TESIS DOCTORAL

Resumen lingüístico de series de datos mediante técnicas de Soft Computing: una aplicación a los cubos OLAP con dimensión tiempo

Rita María Castillo Ortega

Directores:

Dr. Nicolás Marín Ruiz

Dr. Daniel Sánchez Fernández

Programa Oficial de Doctorado en Tecnologías de la Información y la Comunicación

Departamento de Ciencias de la Computación e I.A.

Universidad de Granada

UNIVERSIDAD DE GRANADA

E.T.S. DE INGENIERÍAS INFORMÁTICA Y DE TELECOMUNICACIÓN

Departamento de
Ciencias de la Computación
e Inteligencia Artificial

TESIS DOCTORAL

Resumen lingüístico de series de datos mediante técnicas de Soft Computing: una aplicación a los cubos OLAP con dimensión tiempo

Rita Mª Castillo Ortega

Granada, septiembre de 2012

Editor: Editorial de la Universidad de Granada
Autor: Rita María Castillo Ortega
D.L.: GR 746-2013
ISBN: 978-84-9028-420-9

Resumen lingüístico de series de datos mediante técnicas de Soft Computing: una aplicación a los cubos OLAP con dimensión tiempo

memoria que presenta

Rita Mª Castillo Ortega

para optar al grado de

Doctor en Informática

Septiembre de 2012

DIRECTORES

Dr. Nicolás Marín Ruiz

Dr. Daniel Sánchez Fernández

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
E INTELIGENCIA ARTIFICIAL

E.T.S. de INGENIERÍAS INFORMÁTICA Y DE TELECOMUNICACIÓN
UNIVERSIDAD DE GRANADA

La memoria titulada "Resumen lingüístico de series de datos mediante técnicas de Soft Computing: una aplicación a los cubos OLAP con dimensión tiempo", que presenta Dña. Rita María Castillo Ortega para optar al grado de Doctor en Informática, ha sido realizada en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada bajo la dirección de los Doctores D. Nicolás Marín Ruiz y D. Daniel Sánchez Fernández.

Granada, septiembre de 2012.

La Doctoranda | Los Directores

Fdo. Rita Mª Castillo Ortega | Fdo. Nicolás Marín | Fdo. Daniel Sánchez

Agradecimientos

Cualquiera que me conozca bien se habrá dado cuenta de que no soy muy ducha en estos lances pero, igualmente, si me conocen, saben que les estoy enormemente agradecida simplemente por estar ahí cuando los necesito.

Sí me gustaría mencionar a mis directores, sin los cuales el trabajo de todos estos años nunca hubiera llegado a buen puerto. No sólo he de agradecerles su ayuda y dedicación a mi trabajo durante este periodo, sino el haberme dado la oportunidad de adentrarme en el mundo de la investigación y enseñarme sus entresijos.

Tampoco puedo dejar de nombrar a Carlos Molina por prestarse a colaborar en la integración de mi trabajo en su plataforma web para el manejo de bases de datos multi-dimensionales, y por su paciencia con mis numerosas dudas y problemas. Del mismo modo, quisiera agradecer a Andrea G.B. Tettamanzi por todo el apoyo prestado durante nuestra colaboración para desarrollar la técnica evolutiva implementada en este trabajo. Ya en el ámbito más personal siempre recordaré su inestimable ayuda y amabilidad que hizo más agradable mi estancia en Crema. Otras personas que recuerdo con cariño de mi paso Crema son Olga, Alberto, Antonia y Mauro, siempre se preocuparon por mí y me ayudaron en todo cuanto estuvo en sus manos.

A todos los miembros del departamento de Ciencias de la Computación e Inteligencia Artificial en los que además de compañeros he encontrado amigos y gente siempre dispuesta a echar una mano cuando se necesita.

No me olvido de los compañeros que me han aguantado en este tiempo, primero en *el torreón*, luego en *Orquídeas* y finalmente en el *CITIC*, en especial a Clara, Úrsula, Sergio y mis telecos favoritos que me ayudaron a tomarme las cosas con más calma y que siempre tenían unas palabras de ánimo preparadas para mí. También quiero mencionar aquí a mis amigos por todos los momentos que hemos pasado juntos y que espero seguir compartiendo con ellos. Ha sido una suerte conoceros, me habéis ayudado a ser mejor persona.

Esta sección no estaría completa sin expresar la enorme gratitud que siento hacia mis padres, que confiaron en mí desde el principio y aguantan mis estallidos de mal humor; a mi hermana que siempre logra sacarme una sonrisa al final de un mal día, y a mis tíos y primos que siempre me han prestado apoyo incondicional y me han ayudado a evadirme en momentos difíciles. Por último, gracias a mi mejor amigo con el que espero formar una bonita familia. Gracias por aguantarme y hacerme crecer como persona. Si me has aguantado en estos 4 años, en especial este último, no te costará aguantarme el resto de nuestras vidas.

Siempre me acordaré de tí. Te quiero Abuela.

Índice general

Índice de figuras

Índice de tablas

1 Introducción

1.1. Español

La capacidad para manejar grandes volúmenes de datos se perfila cada vez más necesaria en una sociedad que, sin duda alguna, está basada en el conocimiento. El importante número de grandes empresas, así como de organizaciones u organismos públicos, que generan y consumen ingentes cantidades de datos con el fin de llevar a cabo sus actividades es un buen ejemplo de ello. En este sentido, cabe destacar que la mayoría de estos datos están relacionados con la dimensión temporal de una u otra manera.

Pero no sólo el manejo de los datos en sentido estricto es útil; del mismo modo, el proceso que permite realizar la extracción de la información a partir de grandes conjuntos de datos se está volviendo cada vez más importante para nuestro entorno. La importancia de este proceso se debe al hecho de que permite a los usuarios realizar tareas tan fundamentales como el análisis en la toma de decisiones, pronóstico o predicción [8] de una forma más sencilla y, por tanto, menos costosa.

El proceso de análisis de grandes conjuntos de datos con el fin de encontrar información útil se denomina *Extracción de Conocimiento en Bases de Datos (Knowledge Discovery in Databases, KDD)* o, en su sentido amplio, *Minería de Datos (Data Mining, DM)*. En ésta última disciplina se usan técnicas de Inteligencia Artificial y herramientas estadísticas y de análisis de datos avanzadas para revelar patrones que, de otra forma, no hubieran sido detectados. De este modo, podemos decir que, de forma general, la minería de datos es el proceso de análisis de datos desde diferentes perspectivas para encontrar patrones ocultos que los describan y resumirlos para obtener información útil.

Existe una clasificación ampliamente aceptada por la comunidad investigadora que pretende establecer una tipología de tareas diferenciadas de entre todas las incluidas en el amplio concepto de minería de datos. Dicha tipología está compuesta por seis tareas fundamentales como son la detección de anomalías, el aprendizaje de reglas de asociación, la agrupación o clustering, la clasificación, la regresión y, por último, el resumen. De entre las anteriores, la última será la que se estudiará a lo largo de la presente memoria.

Los *Sistemas de Información (Information Systems, IS)* y, en concreto, las soluciones creadas dentro del área de la *Inteligencia de Negocio (Business Intelligence, BI)* se benefician de las técnicas de minería de datos con el objetivo de mejorar los procesos de toma de decisiones. Las empresas que utilizan a la hora de tomar decisiones información obtenida a través de técnicas de BI disponen de una herramienta que las sitúa en una posición privilegiada a la hora de obtener mejores resultados a

nivel de mercado que aquellas que no lo hacen.

Como se ha mencionado, la información obtenida a través de técnicas de minería de datos juega un papel muy importante en el sector de *los Sistemas de Apoyo a la Toma de Decisiones (Decision Support Systems, DSS)*. Recientemente, varios investigadores han centrado su atención en esta relación tan significativa [104]. Es un hecho que los decisores que cuentan con información extraída de los datos convenientemente procesada ejercen su trabajo con más facilidad que aquellos que cuentan sólo con los datos en cuestión. Además, el proceso se desarrolla de forma mucho más amigable si la información que manejan se encuentra expresada en formato textual.

Los decisores, aunque sean usuarios con conocimiento experto del negocio, no dejan de ser seres humanos a los que resulta conveniente transmitirles la información en una manera lo más amigable posible. A esto se une que, en el ámbito de la minería de datos, es clave que el conocimiento obtenido sea comprensible. En este sentido, y en la mayoría de las situaciones, cuando tratamos con personas, la mejor manera de establecer dicha comunicación es hacerlo a través del lenguaje natural, ya que es la forma *natural* de pensar y comunicarse de los seres humanos. Incluso se puede dar el caso de que la comunicación en lenguaje natural sea la única alternativa posible en determinadas situaciones (por ejemplo, en el caso en el que el emisor no pueda utilizar información visual o el receptor no sea capaz de manejarla).

El área de *Generación de Lenguaje Natural (Natural Language Generation, NGL)* concentra los esfuerzos de investigación destinados a resolver el problema de la creación de textos en lenguaje natural mediante el uso de computadores. El conjunto de técnicas que engloba representa una buena herramienta que, entre otras cosas, nos permite mejorar los sistemas de apoyo a la decisión. En esta memoria, se consideran de especial interés los enfoques que incorporan el uso de herramientas de *Soft Computing* como, por ejemplo, las etiquetas lingüísticas y las sentencias cuantificadas, para llevar a cabo la trasformación de datos numéricos a información textual [83, 183].

Otro aspecto de especial importancia en esta memoria y que tiene relación con los elementos anteriormente descritos, es el modelado multi-dimensional de datos. Con el fin de facilitar al decisor humano el acceso a los datos, se ha introducido en el ámbito empresarial el uso de este modelo. En este sentido, volviendo al campo de la inteligencia de negocio, debemos decir que una parte muy importante de las herramientas disponibles hacen uso de cubos construidos en base a un conjunto de dimensiones que almacenan grandes cantidades de datos que, además, pueden ser consultados mediante *procesamiento analítico en línea (OnLine Analitycal Processing, OLAP)* [116]. Este modelo se encuentra ampliamente extendido.

Debido al papel fundamental que juega el tiempo en nuestra sociedad, la dimensión temporal es una de las que siempre aparecen en los cubos de datos. La mayoría de las operaciones OLAP aplicadas a los cubos de datos con dimensión temporal dan como

resultado series de datos temporales. Dado su interés, un gran número de autores han prestado atención a estudiar este tipo particular de conjunto de datos.

En el contexto presentado anteriormente, y después de un estudio del estado del arte en la materia, se detecta la *necesidad de un modelo general y configurable que permita a usuarios no expertos la obtención automática de resúmenes altamente intuitivos, personalizables y de calidad a partir de series de datos* que, como caso particular, podrían ser extraídas realizando operaciones sobre almacenes de datos multidimensionales con hipercubos organizados alrededor de la dimensión *tiempo*.

Con la idea de dar un paso importante para paliar esta necesidad, y al mismo tiempo hacer un aporte general en el ámbito del NLG, se ha elaborado el trabajo de investigación presentado en esta memoria. Los objetivos planteados van desde el estudio de herramientas para construir dicho modelo, hasta la aplicación de las mismas en el ámbito de las series de datos temporales. También se ha considerado la elaboración de una herramienta software que permita incorporar el modelo propuesto para transferirlo al ámbito de las herramientas de análisis de datos OLAP y el modelo de datos multi-dimensional.

A continuación se presentan los objetivos mencionados de una manera ordenada y detallada:

1. **Estudiar el concepto de resumen y su aplicación al ámbito del resumen de datos y series de datos**. Este estudio debe ser afrontado, en primera instancia, partiendo de una perspectiva general de modo que, luego se pueda concretar para el ámbito de las series de datos temporales y su uso en la toma de decisiones.

2. **Analizar las técnicas y modelos presentes en la literatura y que se centran en la generación automática de resúmenes textuales mediante el uso de computadores**. Para ello será necesario, partiendo del ámbito de la generación de lenguaje natural, establecer las fases que se deben considerar para afrontar la correcta elaboración de los resúmenes, identificando sus características más importantes y señalando el camino que se debe seguir cuando haya que evaluar la calidad del producto obtenido. Se prestará especial atención al uso de técnicas de Soft Computing por su probada utilidad y eficacia en la tarea de convertir datos numéricos en palabras del lenguaje natural, más cercanas éstas al usuario humano.

3. **Proponer un modelo general para la realización automática de resúmenes lingüísticos de series de datos temporales basado en el uso de técnicas de Soft Computing**. El modelo debe ser fácilmente configurable para poder adaptarlo en la medida de lo posible al contexto, así como portable para permitir su aplicación en otros ámbitos y con diferentes tipos de datos.

La elección de la estructura general que debe tener el resumen, así como el uso de una jerarquía de conceptos, debe permitir que el resumen obtenido sea lo más parecido posible a aquel que generaría un resumidor humano. Una vez obtenido el resumen es necesario poder evaluar la calidad del mismo, lo que conduce al siguiente de los objetivos.

4. **Elaborar un modelo general de calidad que permita, por un lado, determinar si un resumen dado es bueno y hasta qué punto lo es y, por otro, diseñar algoritmos que estén destinados a la construcción de buenos resúmenes**. En esta línea, el modelo debe responder a dos restricciones importantes:

   - Debe ser configurable y adaptable, de manera que pueda usarse tanto para el modelo que se aporte como respuesta al objetivo 3, como para otros modelos presentes en la literatura.
   - Debe incorporar criterios cuantificables que permitan su implementación a través de una propuesta algorítmica y también permitan afrontar, en lo posible, la comparación entre sí de los distintos resultados obtenidos.

   Cuando nos planteamos medir la calidad surgen diferentes aspectos que se deben considerar. Sin embargo, una buena parte de estos conceptos es altamente subjetiva y puede variar de una persona a otra o, incluso siendo la misma persona, de una situación a otra. Un segundo problema surge cuando nos damos cuenta de que la mayoría de dichos aspectos no son fácilmente cuantificables. En este sentido es necesario desarrollar un modelo multi-dimensional de medida de la calidad que tenga en cuenta los aspectos más fácilmente ponderables sin dejar de tener en cuenta, en la medida de lo posible, algunos aspectos subjetivos. En cualquier caso, en esta memoria, un resumen de calidad será aquel que represente de forma sucinta, veraz y específica la totalidad del conjunto de datos que se desea resumir.

5. **Presentar propuestas algorítmicas concretas que permitan construir los resúmenes según el modelo lingüístico propuesto, basándose en el modelo de calidad**. Para llevar a cabo dicha tarea con éxito será necesario, en primer lugar, conocer el tamaño o complejidad del problema propuesto así como diferentes opciones algorítmicas disponibles en la literatura que respondan tanto a criterios de velocidad como de diversidad de los resúmenes generados.

   Debido a la subjetividad inherente al proceso, no existe un único resumen para un conjunto de datos. La actividad de identificar el mejor resumen entre todos los posibles se asemeja al proceso de búsqueda en el espacio total de soluciones. Debido a ello, la complejidad del problema se debe determinar en función del tamaño del espacio de búsqueda del mismo, y las técnicas algorítmicas consideradas deben ser técnicas de búsqueda convenientemente adaptadas.

6. **Llevar el modelo a un campo de aplicación real a través del uso de una herramienta amigable y sencilla de utilizar**. En este sentido, dado el marco de motivación que se acaba de describir dentro del ámbito de la inteligencia de negocio, se desarrollará una herramienta de análisis de cubos OLAP de forma que incorpore habilidades de resumen lingüísticas.

   Ya se ha mencionado anteriormente que la minería de datos y, en concreto, el resumen de datos son procesos que en el área de la inteligencia del negocio proporcionan buenos resultados para apoyar la toma de decisiones. Si además los resúmenes de dichos datos se presentaran en lenguaje natural la tarea llevada a cabo por el decisor se facilita. En este contexto ya no es necesario que el decisor se enfrente a grandes cantidades de datos sino que se podría ayudar de herramientas que realicen todo el proceso de extracción de información novedosa, previamente desconocida y potencialmente útil.

Al cumplimiento de los objetivos 1 y 2 dedicaremos el Capítulo 2 de la memoria. En el Capítulo 3 abordaremos la presentación y desarrollo de sendos modelos para la obtención del resumen y para la medida de la calidad del mismo, objetivos 3 y 4. Las propuestas algorítmicas a las que hace referencia el objetivo 5 para que implementen los modelos se presentarán en el Capítulo 4. El Capítulo 5 se dedicará al estudio de la generalización y portabilidad del modelo. Por último el Capítulo 6 se dedicará a abordar la funcionalidad comentada por el objetivo 6.

1.2. English

The ability to manage large volumes of data is increasingly turning into an essential issue in a society based, without any kind of doubt, on knowledge. The significant number of large companies, as well as of organizations or public bodies, which generate and consume huge amounts of data in order to perform their activities is a good example of this fact. In this sense, it is worth highlighting that most of these data are related to the temporal dimension in one way or another.

But not only data management in a strict sense is useful; in the same way, the process which allows the discovery of the knowledge from large datasets is turning more and more important for our environment. The relevance of this process lies in the fact that it allows users to perform such essential tasks as the decision making analysis, prediction, or forecast [8] in a simpler and, as a result, cheaper way.

The analysis process of large datasets so as to find useful information is known as *Knowledge Discovery in Databases (KDD)* or, in a broader sense, *Data Mining (DM)*. In the latter discipline, Artificial Intelligence techniques and advanced statistical and data analysis tools are used to reveal patterns which would not have been detected otherwise. In this way, we can state that, generally speaking, data mining is the data

analysis process from different perspectives to find hidden patterns which describe those data and to summarize them so as to obtain useful information.

There is a widely accepted classification by the research community intending to establish a differentiated task typology from among all tasks which are included in the broad concept of data mining. Such typology consists of six main tasks, namely anomaly detection, association rule learning, clustering, classification, regression, and lastly, summarization. From among these tasks, this document will focus on the last one of them.

The *Information Systems (IS)* and, particularly, the solutions created within the *Business Intelligence (BI)* area, benefit from the data mining techniques in order to improve the decision making processes. The companies which use BI techniques for their decision making processes are in a privileged position for obtaining better results on a market level than those companies which do not use them.

As has been mentioned, the information obtained through data mining techniques plays a very important role in the *Decision Support Systems (DSS)* sector. Recently, several researchers have focused their attention on this significant relationship [104]. The decision makers who have appropriately processed information extracted from the data perform their work in an easier way than those who just have the actual data. Besides, the process takes place in a more user-friendly way if the managed information is expressed in a textual format.

The decision makers, although in the case that they are users with an expert knowledge of the business, are just human beings, due to which it is important to convey the information to them in the most user-friendly way as possible. Together with this, in the scope of data mining, a key issue is that the obtained knowledge is understandable. In this sense, and in most of the situations, when we deal with individuals, the best way to establish such communication is to do it through natural language, as that is the *natural* way to think and communicate for human beings. Even more, natural language communication is the only possible alternative in certain situations (for example, the case in which the emitter cannot provide visual information or the receptor is not able to manage it).

The *Natural Language Generation (NLG)* area concentrates the research efforts directed to solving the problem of the creation of texts in natural language by means of the use of computers. The set of techniques of which it consists represents a good tool which, among other things, allows us to improve the decision-making support systems. In this research work, the approaches which include the use of Soft Computing tools, as for example, linguistic labels and quantified sentences [83, 183] are considered to be of special significance.

Another particularly relevant aspect in this document and which is related with

the previously described elements is the multi-dimensional data modelling. In order to facilitate the access to the data to the human decision maker, the use of this model has been introduced in the entrepreneurial area. In this sense, back to the field of the business intelligence, we must say that a very important part of the available tools use cubes based on a set of dimensions storing large amounts of data which, besides, can be accessed by means of the *OnLine Analytical Processing (OLAP)* [116]. This model is a widely spread one.

Due to the essential role played by time in our society, the temporal dimension is always present on data cubes. Most of the OLAP operations applied to data cubes with a temporal dimension lead to temporal data series. Due to their interest, a large number of authors have paid attention to the study of these particular data sets.

In the previously presented context, and after a study of the state of the art in this issue, we detected the *need for a general and configurable model allowing non expert users to automatically achieve understandable, customizable, and high quality summaries from data series* which could be extracted, in a given case, by implementing operations on multi-dimensional data warehouses with hypercubes organized around the *time* dimension.

The research work presented in this report is intended to address this need and at the same time, to make a general contribution in the area of NLG. The objectives of this work range from the study of tools to build the aforementioned general summarization model, to the particular application within the scope of the temporal data series. The creation of a software platform which allows the addition of the proposed model to transfer it to the area of the OLAP data analysis tools and of the multi-dimensional data model has also been considered.

The mentioned objectives are listed below in a detailed and ordered way:

1. **Studying the concept of summary and its application to the area of data summaries and data series.** This study must be faced, in the first instance, beginning from a general perspective so that it can be later on focused on the temporal data series and on its use for the decision making process.

2. **Analysing the techniques and models which are present in the literature and which focus on the automatic generation of linguistic summaries by means of computers.** For such purpose, and beginning from the natural language generation area, it will be necessary to establish the stages which must be taken into account to face the correct drafting of the summaries, identifying their most important features and indicating the path that have to be followed in order to assess the quality of the obtained product. Special attention will be paid to Soft Computing techniques, due to their proven usefulness

and efficiency for converting numeric data into natural language words, which are closer to the human user.

3. **Proposing a general model for the automatic creation of linguistic summaries of temporal data series based on the use of Soft Computing techniques.** The model must be easily configurable, so as to be able to adapt it as much as possible to the context, and portable to allow its application in other areas and with different types of data.

   The choice of the general structure of the summary, as well as the use of a concept hierarchy, must allow the obtained summary to be as similar as possible to those generated by a human summarizer. Once the summary is obtained, it is necessary to be able to evaluate its quality, which leads to the following objective.

4. **Providing a general quality model allowing, on the one hand, to establish whether a given report is good or not and to which extent it is good, and on the other hand, to design algorithms destined to the creation of good summaries.** In this line, the model must meet to important restrictions:

   - It must be configurable and adaptable, so that it can be used both for the model provided as the answer to objective 3 and for other models present in the literature.
   - It must feature quantifiable criteria allowing its implementation through an algorithmic proposal and enabling as well, as much as possible, the comparison of the obtained results.

   When thinking about measuring quality, different aspects which might be taken into account arise. However, many of those concepts are highly subjective and can vary from person to person. A second problem arises when we realize that most of those aspects are not easily quantifiable. In this sense, it is necessary to develop a multi-dimensional quality measuring model which takes into account those aspects which are more easily assessed and also, to the possible extent, some subjective aspects. In any case, in this memoir, a quality summary will be that one representing in a brief, truthful, and specific way the whole data set to be summarized.

5. **Presenting specific algorithmic proposals allowing us to build the summaries according to the proposed linguistic model, based on the quality model.** To carry out such task successfully, it will be necessary on the first place to know the size or complexity of the proposed problem, as well as different algorithmic options from those available in the literature which meet both speed criteria and diversity criteria of the generated summaries.

Due to the inherent subjectivity of the process, there is not one single summary for a dataset. The activity of identifying the best summary from among all possible summaries is similar to the search process in the total space of solutions. Due to this, the complexity of the problem must be established according to the size of the search space of such problem and the considered algorithms must be appropriately adapted search techniques.

6. **Taking the model to a real application field through the use of a user friendly interface.** In this sense, due to the motivation framework which has just been described within the scope of business intelligence, an OLAP cube analysis tool will be developed, so that it features linguistic summarization abilities.

   It have been already mentioned before that data mining and, particularly, data summarization are processes which provide good results in the area of business intelligence for the support of decision making processes. If the summaries are besides presented in natural language, the task to be performed by the decision maker is made easier. In this context, it is no longer necessary for the decision maker to face large amounts of data, but to use some tools which perform the discovery of new, previously unknown, and potentially useful information.

Chapter 2 will be devoted to accomplish objectives 1 and 2 in the memoir. In Chapter 3 we will present and develop both models for the summarization and quality measure, objectives 3 and 4. The algorithmic proposals referenced by objective 5 to implement the models will be introduced in Chapter 4. Chapter 5 will be dedicated to the study of the generalization and portability of the model. Finally, in Chapter 6 we will present the functionality commented in objective 6.

2 *Estudio preliminar*

"No abras los labios si no estás seguro de que lo que vas a decir es más hermoso que el silencio"
Proverbio árabe

A lo largo de este capítulo se realiza un estudio de la idea o concepto de resumen, comenzando desde las definiciones más generales hasta llegar a áreas más concretas, y que son relevantes para el objetivo de esta tesis, como el resumen lingüístico de series de datos temporales mediante el uso de técnicas procedentes del Soft Computing.

¿Qué es resumir?, ¿qué es un resumen? o ¿cuáles son las características y funciones de un resumen? son algunas de las preguntas que se abordan y tratan de contestar a lo largo de la primera parte del presente capítulo con la intención de, paso a paso, introducir al lector en la problemática que se afronta en este documento.

A continuación, nos adentramos en el campo de estudio relacionado con el resumen de datos y sobre todo en las series de datos temporales. Con esta intención se presentan diferentes herramientas y técnicas existentes para su estudio. Para ello seguimos una línea cronológica en la que se comienza por modelos clásicos y se finaliza con las metodologías más novedosas en este campo.

Como se ha mencionado anteriormente, se introduce el término Soft Computing así como qué representa y cómo puede ser utilizado para mejorar los procesos de resumen lingüístico automatizados. Entre las herramientas más destacadas encontraremos elementos tan importantes como los Conjuntos Difusos y las etiquetas lingüísticas asociadas, o los Algoritmos Evolutivos.

Por último, pero no por ello menos importante, se presenta al lector un completo estudio de diversas propuestas de modelos que hacen uso del Soft Computing para enfrentarse a la tarea de generar resúmenes mediante la utilización de computadores. Los diferentes trabajos se clasifican a través de las herramientas utilizadas durante su desarrollo. Este estudio pretende poner en antecedentes al lector dándole una idea clara de los métodos existentes y sus principales características así como las tareas que presentan especial dificultad o que quedan pendientes de resolver.

2.1. Resumen lingüístico

En la sección actual centraremos nuestra atención en el concepto de resumen y, en particular, el de resumen lingüístico. Veremos qué es, así como las características de un buen resumen, sus funciones y cómo obtener un resumen de calidad. En último lugar enfocaremos nuestra atención en el resumen lingüístico de datos, área en la que centraremos nuestro interés a lo largo del documento.

2.1.1. ¿Qué es resumir?

De acuerdo con el diccionario de la Real Academia Española de la Lengua, *resumen* es una "exposición resumida en un asunto o materia". Se trata también de la "acción y efecto de resumir o resumirse". Y lo que hacemos al *resumir* es "reducir a términos breves y precisos", o, "considerar tan sólo y repetir abreviadamente lo esencial de un asunto o materia".

No sólo es posible realizar resumen lingüístico de información representada textualmente, sino que también puede hacerse de información con otras representaciones, como por ejemplo, de conjuntos de datos de otra índole o tipo. En este sentido se han seguido las ideas de María Pinto en [108] cuando define el documento original objeto del proceso de resumen "como la acumulación permanente y estable de signos". En esta definición no se establece la necesidad de que el documento sea exclusivamente un documento textual, permitiendo otras muchas representaciones del mismo.

En el contexto de este trabajo se siguen dichas ideas y se considera que un documento representa a un conjunto de datos de distinto tipo, ya sea textual, numérico, etcétera. De modo que un documento no se tomará como exclusivamente textual, ni un conjunto de datos como exclusivamente numérico. A partir de este momento, se usarán las expresiones *documento* y *conjunto de datos* indistintamente, considerando que representan lo mismo.

De nuevo, siguiendo las ideas e investigaciones de Pinto [45], diremos que el "resumen es el documento referencial más completo y por consiguiente el que mejor representa la información original, ofreciendo una visión global del contenido del documento".

Volviendo a la definición de resumen dada por la Real Academia Española de la Lengua, debemos puntualizar que el hecho de que un resumen sea *breve* no implica que en él se refleje lo *esencial* acerca de algo, aunque esta situación sería la más deseable. En muchas ocasiones estas dos palabras son usadas como sinónimos cuando en realidad no lo son. Es cierto que existen diversas circunstancias o situaciones en las que no es necesaria tanta precisión al definir términos, ya que existen patrones de conocimiento común respecto a un cierto tema, o que son compartidos por una determinada comunidad y que por tanto no necesitan ser explicitados. Por todo lo

dicho anteriormente, durante el trabajo se han tenido muy en cuenta las diferencias existentes entre los términos *breve* y *esencial*.

La situación ideal a la hora de enfrentarnos a la realización de un resumen sería que fuésemos capaces de ofrecer la información esencial de una forma breve y concisa. Pero en la mayoría de los casos, las situaciones no se caracterizan por ser ideales. Podemos encontrarnos, por ejemplo, con que toda la información esencial no pueda ser reflejada de forma breve; aunque lo que realmente representa un problema son las distintas percepciones que tienen diferentes personas de los conceptos *breve* o *esencial*. Esta concepción vendrá marcada por los diferentes intereses de cada uno o la utilidad que se le vaya a dar al resumen obtenido. Lo que para un individuo o grupo puede resultar interesante o esencial, puede no serlo para otros, y viceversa; incluso se pueden presentar visiones diferenciadas acerca de qué extensión puede ser considerada o no como breve. Los procesos de confección de un resumen e interpretación del mismo son altamente subjetivos, de modo que podemos decir que son procesos sensibles al contexto en el que se desarrollan.

Un ejemplo práctico de lo comentado anteriormente con respecto a la disparidad de puntos de vista, lo podemos encontrar cuando se pide que se realice una descripción de una persona. Cada individuo realizará una descripción diferente en función de las circunstancias, sus gustos o de elementos que hayan captado su atención. Incluso un mismo individuo puede confeccionar diferentes descripciones de la persona ajustándose al receptor de los mismos. Por norma general, las personas realizan este proceso de ajuste de forma natural e inconsciente. Las descripciones que dos amigos hacen de un tercero pueden diferir, pero ser más parecidas entre sí que la descripción que nos dará la madre de ese tercero, y que, a su vez, será muy distinta de la que dé el jefe del mismo a su superior. Los primeros reflejarán posiblemente su personalidad y la manera de actuar con sus amigos, mientras que la madre se centrará en aspectos relacionados con la vida familiar, y el jefe hará hincapié en la formación y las habilidades profesionales que posee.

El proceso de resumir es una actividad inherente al ser humano, y como se ha comentado anteriormente, con altas dosis de subjetividad. Las personas continuamente reciben información del exterior que someten a una serie de procesos (transformación, reducción, almacenaje, recuperación o utilización) según sus capacidades y necesidades, con vistas a una futura aplicación. De este modo, y siguiendo las ideas de Neisser en [109], podemos decir que:

> "La información que asimila la persona se selecciona y condensa condicionada a sus intereses. Las sensaciones que no se acomoden a ellos son olvidadas, desechadas y reemplazadas por otras en un proceso continuo basado en la *elección* y *selección* de aquellas".

Por tanto, podemos considerar que resumir implica una actividad de reducción natural de información en la mente humana. En este proceso se pasa a fijar los conceptos más importantes o significativos de entre todos los datos percibidos. Como dice Pinto, resumir "se trata pues de un proceso de abstracción, que va de lo específico a lo general, eliminando lo que no se considere esencial" [108]. En nuestro caso, el trabajo que realizamos va encaminado a lograr que dicho proceso se pueda realizar de forma automatizada por ordenador incorporando una cierta cantidad de información contextual.

En general el resumen pretende ser lo mismo pero en tamaño más pequeño y no una parte arbitraria de lo que tenemos que resumir. Por ejemplo, si se resume un texto es para tener al alcance, en muy poco tiempo y de una sola ojeada, la información importante de dicho texto. El mismo principio puede ser aplicado al realizar un resumen de datos de otro tipo, mediante técnicas de análisis de los mismos. En definitiva, es la representación abreviada y precisa del contenido de un conjunto de datos o situaciones, entre otros, fruto de la transformación experimentada a través de un doble proceso de análisis y síntesis.

Funciones del resumen

Una de las funciones principales del resumen es servir de anticipo del documento original, y capacitar al lector para decidir sobre la conveniencia o no de consultar dicho documento al completo. En ocasiones puede incluso actuar como sustituto del mismo en caso de que el usuario haya decidido no consultarlos, de modo que se evita la lectura de información que pueda resultar marginal. De esta forma se convierte en una ayuda muy importante en las tareas de búsqueda retrospectiva y recuperación de la información.

De forma más precisa, y basándonos en Reques [133], podemos decir que las funciones del resumen son:

- Servir de anticipo del conjunto de datos, al identificar de forma rápida y precisa el mismo, permitiendo al usuario decidir sobre la conveniencia o no de consultar los datos de forma íntegra.

- Convertirse en sustituto del conjunto de datos, en los casos en que por tener éste un interés marginal, el resumen suministra información suficiente al usuario.

- Contribuir a superar las barreras técnicas, siendo el resumen en ocasiones el único medio de acceso a la información sustancial de un documento almacenado.

- Ayudar en las tareas de búsqueda automatizada de información.

Ejemplos de resumen de datos: textual y numérico

El resumen, y las técnicas para realizarlo de manera apropiada, han sido intensamente trabajados en relación, sobre todo, a documentos compuestos por datos textuales y datos numéricos. Está claro que no se pueden aplicar las mismas técnicas para realizar un resumen de un documento textual que para resumir un conjunto de datos numéricos. Al ser la naturaleza de los datos diferente, se deben aplicar también diferentes técnicas, donde cada una de las cuales ha sido diseñada para intentar obtener los mejores resultados en relación a las particularidades de cada disciplina.

Con motivo de presentar algún ejemplo de **resumen de datos en formato texto** vamos a tomar prestado un párrafo del libro *La sombra del viento* de Carlos Ruiz Zafón.

> *"En una ocasión oí comentar a un cliente habitual en la librería de mi padre que pocas cosas marcan tanto a un lector como el primer libro que realmente se abre camino hasta su corazón. Aquellas primeras imágenes, el eco de esas palabras que creemos haber dejado atrás, nos acompañan toda la vida y esculpen un palacio en nuestra memoria al que, tarde o temprano -no importa cuántos libros leamos, cuántos mundos descubramos, cuánto aprendamos u olvidemos-, vamos a regresar. Para mí, esas páginas embrujadas siempre serán las que encontré entre los pasillos del Cementerio de los Libros Olvidados".*

Una vez finalizada la lectura del párrafo anterior un resumen adecuado podría ser *"un joven oyó que el primer libro importante marca a un lector sin importar cuántos vengan detrás"*. Por supuesto no es el único que se puede realizar, y posiblemente no es el mejor, pero no es una mala elección. Otros resúmenes correctos en igual medida son: *"el primer libro que nos llega al corazón, nos deja marcados de por vida"*, o *"las primeras páginas que marcaron al personaje para siempre, las encontró en el Cementerio de los Libros Olvidados "*.

Los distintos resúmenes son igualmente válidos, es decir, describen de forma resumida lo que aparece en el párrafo, pero cada uno pone el énfasis en una parte o simplemente transmite su idea de forma diferente. Existen tantas maneras de resumir, como personas distintas hacen el resumen y más aún cuando se consideran contextos diferentes. Esto se hace mucho más patente sobre todo en textos más largos: cuanto más largo es el texto, más resúmenes diferentes se pueden hacer de él. ¿Cuál es el mejor? Eso dependerá de la persona que realiza el resumen, de la situación que la rodea, de la utilidad que se le quiera dar al mismo, de la persona que lo va a recibir, y un largo etcétera.

Día de la semana	Correos publicitarios S_1	Correos publicitarios S_2
Lunes	12	11
Martes	9	12
Miércoles	5	11
Jueves	2	12
Viernes	12	12
Sábado	20	12
Domingo	25	15

Tabla 2.1: Cantidad de correos publicitarios en una cuenta de correo en las semanas S_1 y S_2.

Para ejemplificar el **resumen de datos en formato numérico**, supongamos ahora que disponemos de los datos representados en la Tabla 2.1. Estas cantidades nos informan del número de correos de publicidad que han llegado a una cuenta de correo durante las semanas S_1 y S_2. En esta ocasión para realizar el resumen de los datos contamos con la ayuda de las herramientas de análisis de datos.

Un resumen que podríamos hacer de los datos correspondientes a la semana S_1 sería, por ejemplo, la información dada por la media, mediana y moda. En este caso obtendríamos las medidas $media \approx 12{,}14$, $moda = 12$ y $mediana = 12$.

Por supuesto existen medidas estadísticas más complejas y que nos ayudarán a completar la información que compondrá el resumen, pero todas ellas no dejan de ser más datos numéricos, que a los ojos humanos resultan iguales que los anteriores y que por lo general se presentan difíciles de interpretar por usuarios sin conocimiento experto. Además, éstas medidas dan cierta información de los datos pero no dan una idea clara en ciertas ocasiones, vemos para ello el resumen para la semana S_2. En esta ocasión los datos son totalmente diferentes a los obtenidos anteriormente, sin embargo si nos fijamos en las medidas estadísticas obtenemos que $media \approx 12{,}14$, $moda = 12$ y $mediana = 12$. De este modo vemos que dos configuraciones diferentes de datos pueden dar lugar al mismo resumen numérico, ¿cómo hacernos una idea de la verdadera distribución de los datos si únicamente contamos con dicho resumen?.

Existen situaciones en las que, debido al volumen de datos, necesitaremos multitud de medidas de este tipo que transmitan la información pero que puedan resultar difíciles de entender, y por supuesto, el problema crecerá al aumentar el volumen de datos.

Junto con otros autores, nosotros creemos que una buena forma de solucionar esta situación es hacer uso de los resúmenes lingüísticos. En este caso para la semana S_1 obtendríamos resúmenes del tipo *"La recepción de correos es alta al comenzar la semana, reduciéndose al llegar a la mitad e incrementándose de nuevo al acercarnos al*

fin de semana" o *"Los días cercanos a los que componen el fin de semana son aquellos es los que se recibe más publicidad"*. Mientras que para S_2 tendremos *"La cantidad de correos publicitarios se mantiene casi constante durante toda la semana percibiéndose un ligero incremento hacia el final de la misma"*.

En este caso, la ventaja del resumen lingüístico es que puede aportarnos información fácilmente procesable por los receptores humanos. Además, los resúmenes para S_1 y S_2 reflejan las diferencias existentes en la configuración de los datos. Finalmente. la situación no será tan dramática conforme se vaya aumentando el cantidad de datos, ya que el proceso, al ser automático y presentar igualmente una salida lingüística, será transparente para el usuario.

2.1.2. Obtención de un buen resumen

La obtención de un buen resumen es algo muy importante. No basta con obtener un resumen, dicho resumen debe satisfacer las necesidades del usuario y debe hacerlo con una cierta *calidad*. Para determinar la calidad de un resumen deberemos tener en cuenta una serie de cualidades y la medición de las mismas.

Ser capaces de determinar la calidad de un resumen, ya sea de forma individual o como mecanismo que nos permita la comparación entre resúmenes, es un aspecto esencial de esta investigación, de forma que volveremos sobre él en capítulos posteriores. Como adelanto decir que las pautas de calidad son tan importantes para nosotros que las introduciremos en el proceso de elaboración de los resúmenes en sí.

La elección de características relevantes es una acción que determinará qué partes son las que queremos destacar o a cuáles prestar mayor atención.

Cualidades del buen resumen documental

El resumen, como texto independiente y representativo de los datos originales, debe aspirar a una serie de cualidades como son la *objetividad*, la *brevedad*, la *relevancia*, la *homogeneidad*, la *claridad*, la *coherencia*, la *profundidad* y la *consistencia* [108].

La objetividad es muy importante y juega un papel prominente en la calidad del resumen, aunque como se puede intuir, es difícil de conseguir completamente. La brevedad, comentada anteriormente, se consigue suprimiendo la información no relevante o repetitiva. La relevancia hace que el resumen se adecue al mensaje representativo de los datos, sin ningún tipo de omisiones y/o interpretaciones de datos. La estructura del resumen debe ser homogénea. El texto debe ser claro, coherente y profundo, en función de los diferentes niveles de descripción necesarios. El ajuste a las pautas, recomendaciones, consejos y normas, repercutirá en la consecución de un resumen consistente.

Debemos destacar que de entre las cualidades listadas anteriormente hay algunas que son más fáciles de medir que otras. En cierto sentido es sencillo obtener la brevedad de un resumen (en relación al tamaño del conjunto de datos y la longitud del resumen final) o si es consistente con los datos a los que representa, pero, cómo se podría evaluar la claridad o la homogeneidad es algo menos claro. Otro aspecto que debemos tener en cuenta es la relación patente entre la capacidad de medir una cierta cualidad y si ésta es objetiva o subjetiva al usuario o la situación en la que utilizará el resumen. Volveremos sobre ello.

Evaluación de un resumen

Una vez que hemos obtenido un resumen o una serie de ellos debemos ser capaces de medir de alguna forma la calidad del mismo. El concepto de calidad de un resumen nos ayudará a discernir si un determinado resumen es de nuestro interés o se ajusta a nuestras necesidades, permitiendo también realizar comparaciones entre distintos resúmenes o incluso establecer una ordenación o "ranking" con los mismos. Desde un punto de vista general, en este campo encontramos referencias como la norma UNE 50-103-90 [44], también denominada ISO 214:1976, o el trabajo [15] entre otros.

Como pautas generales habrá de valorarse si el resumen cumple los siguientes puntos:

- Contiene los puntos esenciales del original.
- Si son descritos exacta y sucintamente.
- Coherencia y legibilidad del estilo.
- Permite al lector prever si el ítem resumido es relevante para sus intereses.
- Comparación con el resumen ideal.

Lo breve o lo esencial. La calidad y la cantidad no son conceptos encontrados sino complementarios. La cantidad juega un papel en una calidad que para poder medirse se manifiesta en términos cuantitativos. Según S. Richard y Pinto, para medir la calidad de un resumen hay que fijar un conjunto de medidas de calidad derivadas del punto de vista adoptado [134] según las necesidades, y en las que cada atributo desempeñe un protagonismo parcial en consonancia con una escala multi-atributos [108].

Estas pautas para medir la calidad de un resumen son muy intuitivas pero vemos que no son fácilmente cuantificables. En la actualidad no existe un conjunto único de medidas calculables que nos informen sobre la calidad de un resumen y esto se debe a que, como se comentó anteriormente, existe un alto componente de sensibilidad al

contexto o subjetividad que hace que un resumen sea perfecto en una situación pero no sea adecuado en otra.

Contar con un conjunto de medidas que nos indiquen la calidad de un resumen es, ciertamente, muy útil en situaciones naturales, pero lo es todavía más cuando trabajamos con resúmenes lingüísticos generados automáticamente por un ordenador. A lo largo de esta memoria volveremos a tratar la calidad de un resumen y presentaremos una métrica multi-dimensional con la que medir la calidad del resumen generado.

Un punto muy importante a la hora de obtener un buen resumen es la elección de las características sobre las que deseamos informar en él; pero más importante, si cabe, es establecer si existe la necesidad de informar de todas ellas en cualquier situación, o sólo cuando presenten eventos relacionados o valores anormales y por lo tanto interesantes. No sólo la elección de las características es algo moldeable en función de la persona que va a recibir el resumen, sino que también lo son los momentos o situaciones en las que se va a informar de la variación de dichas características. Vemos una vez más la importancia que tiene la subjetividad en el proceso de resumen. La selección de características relevantes, y por tanto el establecimiento de información que es posible descartar, es una tarea que se puede prefijar desde el comienzo del proceso.

En la Figura 2.1 se puede observar la evolución del precio del petróleo en estos últimos años, desde 2001 hasta enero de 2009 aproximadamente. Si estamos interesados en realizar un resumen destinado a consumidores normales sería adecuado informar de *"un periodo de subida del precio hasta 2008, año en el que comenzó a bajar"*, o destacar *"el precio en la actualidad es equiparable con el que existía en 2005"*, incluso que *"el precio más elevado tuvo lugar a mediados del 2008 y fue alrededor de 80 a 90 euros el barril brent"*. En caso de que los receptores del resumen fueran economistas la redacción debería ser diferente, por lo que el concepto de esencial se ampliaría; en este caso, distintas características serían consideradas como relevantes. En esta ocasión podría resultar de interés informar de los diferentes máximos o mínimos locales del precio así como en qué momento tuvieron lugar. Esta información les podría resultar útil a la hora de relacionar estos precios con eventos destacables en la evolución de la Historia, como conflictos, guerras, periodos de crecimiento económico, etcétera.

Al establecer las características sobre las que queremos información es inevitable que se produzca un descarte controlado de información; de modo que la pérdida de información es algo que en mayor o menor medida siempre aparece ligado a la confección de un resumen. La mayoría de las veces la información que se queda en el camino es irrelevante, pero en cualquier caso hay que sacrificarla en pos de la brevedad. Una vez que se tiene asumido esto hay que establecer los límites en el nivel de información que estamos dispuestos a sacrificar. Esto lo haremos, como en otras ocasiones, atendiendo al uso que recibirá el resumen resultante. Si volvemos a los diferentes es-

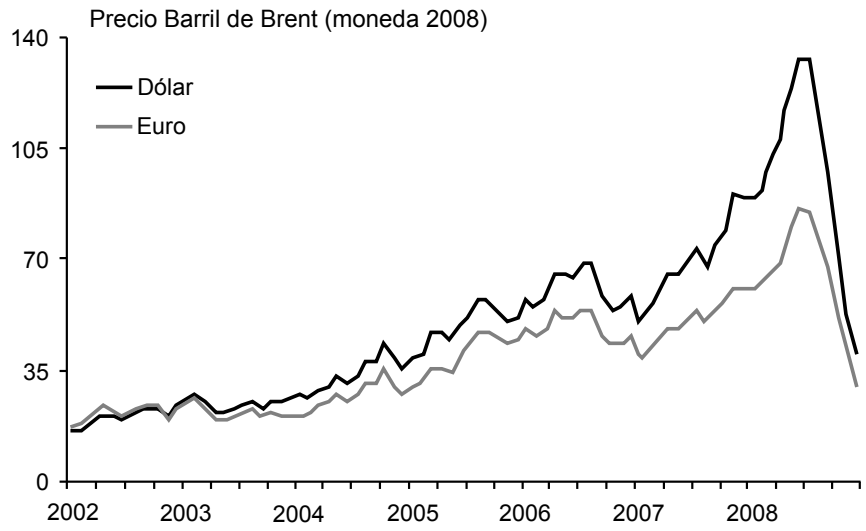


Figura 2.1: Evolución del precio del petróleo en los últimos años.

tilos de resúmenes de la Figura 2.1, podemos observar que en los más generales se pierde la información acerca de las diferentes cimas o valles locales, destacándose sólo la tendencia general o conceptos globales. En esta ocasión se sacrifica información, que en principio creemos que no es relevante o esencial. En cada situación el grado de información que estemos dispuestos a perder será diferente según las diferentes circunstancias que la rodeen.

Podemos afirmar que el resumen debe presentar las ideas esenciales de un determinado tema de forma breve y siendo conciso. Como decíamos, un componente muy importante en la disciplina del resumen es el factor subjetivo (quién lo hace, a quién va dirigido o con qué fin entre otras), que marcará las líneas generales y características del resumen. Teniendo en cuenta los factores anteriores, se deben seleccionar las características de más interés; pero no sólo eso, ya que también se deben establecer las situaciones en las que se considera de interés informar de ellas y qué técnicas son las más adecuadas.

Teniendo los puntos anteriores en mente, vamos a delimitar nuestro campo de trabajo profundizando respecto al estudio de resúmenes lingüísticos de series de datos con representación numérica. En particular el resumen lingüístico presenta bastantes dificultades, pero más si tenemos en cuenta que lo que se pretende es realizarlo de una manera automatizada por parte de una computadora y con unos estándares adecuados de calidad.

Debemos tener en cuenta que si el proceso puede en ocasiones ser complicado para las personas que están acostumbradas a realizarlo (dependiendo del contexto, la familiarización con el mismo o el receptor del resultado), mucho más lo va a ser para las máquinas, que no son capaces de manejar adecuadamente el lenguaje natural. Precisamente son los procesos que los humanos realizan inconscientemente los más complicados de modelar o simular, ya que no se sabe con certeza la secuencia de acciones que se llevan a cabo en el cerebro en estos casos.

Los procesos automáticos o asistidos por ordenador para la obtención de resúmenes pretenden imitar de algún modo al proceso llevado a cabo por las personas. El "resumidor" humano es el modelo a imitar, intentando para ello, "conocer sus métodos espontáneos de percepción, interpretación y producción" [108], así como sus modos de proceder en general.

El interés de la creación de resúmenes lingüísticos de forma automatizada radica en la cercanía del resultado a los usuarios humanos que posteriormente manejarán la información y que ello se realice en un lapso de tiempo adecuado. De nada sirve un complejo y fantástico resumen de un conjunto de datos, si el receptor no va a saber cómo enfrentarse a él, o si tiene que esperar más de la cuenta para obtenerlo. El resumen debe ser lo más intuitivo posible para los receptores, de modo que sea de interés y utilidad; y, para ello, nosotros pensamos que una buena forma de hacerlo es usando el lenguaje natural.

2.1.3. Resumen lingüístico de datos

Poseer gran cantidad de *datos* no es equivalente, de forma directa, a poseer una gran cantidad de *información*. La diferencia entre estos dos conceptos es un proceso muy complejo durante el cual se realiza un tratamiento de los datos de modo que podamos llegar a conseguir la información deseada. Generalmente, las personas tienen dificultades a la hora de enfrentarse a grandes volúmenes de datos, que en ocasiones, y a no ser que posean conocimiento experto, no saben cómo manejar o tratar. Por razones como las comentadas, ya hemos introducido anteriormente la necesidad de realizar procesos de resumen de los datos. Del mismo modo, hemos tratado ya el interés y la problemática de realizar los resúmenes de forma automatizada.

Las actividades de resumen pueden ser aplicadas a casi cualquier cosa, por ejemplo, imágenes, sonidos, sabores, sensaciones, circunstancias, etcétera, pero principalmente se aplican a conjuntos de datos; que en nuestro caso particular, se encontrarán almacenados digitalmente.

Entre los conjuntos de datos con los que más se suele trabajar en nuestro área de investigación encontramos aquellos que contienen información textual o bien información numérica (además de estos, también son muy populares los conjuntos de datos que representan imágenes). En estas situaciones es muy habitual que la salida

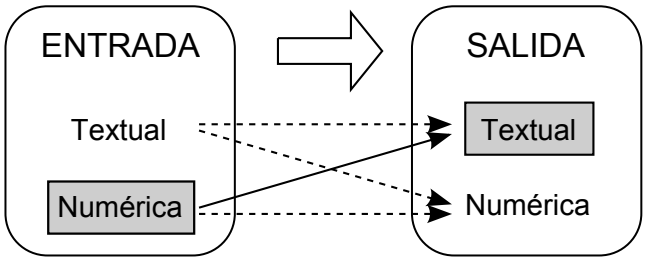


Figura 2.2: Relación entrada-salida al realizar resumen.

misma del proceso de resumen se presente a su vez por texto o más datos numéricos. La razón es que en estas circunstancias se dan las condiciones necesarias para que el proceso se pueda realizar de forma automatizada por un computador.

Ejemplos de lo expuesto anteriormente son la cantidad de programas informáticos que, a partir de un conjunto de datos numéricos, presentan informes numéricos con distintas medidas, por ejemplo, media, mediana, moda, varianza o desviación típica entre muchas otras (entrada y salida numéricas). Del mismo modo, se han realizado amplios estudios sobre el resumen automático de textos, creando un resumen textual del documento original (entrada y salida textual) o mostrando una serie de medidas numéricas, como pueden ser la longitud en palabras del texto, las ocurrencias de dichas palabras o número de palabras no-vacías del texto entre otras (entrada textual y salida numérica).

El estudio de estos supuestos queda fuera del ámbito de estudio de esta tesis que se centrará exclusivamente en resumen lingüístico de series de datos numéricos, área que al igual que numerosos investigadores consideramos de mucha importancia y que comparativamente con las anteriores creemos que ha sido objeto de menos investigación, pero sobre todo, en la que todavía queda mucho por hacer. En la Figura 2.2 quedan reflejadas las situaciones más habituales, encontrándose marcada mediante un sombreado la casuística para la que desarrollaremos nuestro modelo.

La utilidad y potencia del resumen lingüístico de datos numéricos radica en la capacidad de adaptar la información obtenida a patrones de expresión comprensibles y manejables por los seres humanos. De esta manera, la información es más interpretable y por lo tanto más útil.

Entre las series de datos numéricos más extendidas se encuentran las series de datos temporales. Las mencionadas series juegan un papel muy importante en nuestra vida cotidiana debido a que el tiempo juega un papel importante en sí mismo.

Cuando intentamos pensar en series de datos, la mayoría de las que acuden a nuestra mente pueden ser clasificadas como series de datos temporales, por ejemplo, *"la*

variación de precios de un producto determinado a lo largo del tiempo", *"la observación de procesos meteorológicos en un periodo determinado"*, *"la variación del stock de un almacén a lo largo de la campaña navideña"*, y muchas más.

Aunque en la práctica, y como veremos en el Capítulo 5, el modelo para resumen que presentamos puede ser aplicado a otros conjuntos de datos, hemos enmarcado el trabajo en el uso de series de datos temporales debido al interés que su estudio ha suscitado a lo largo del tiempo y al transcendental rol que éstas juegan en la sociedad moderna.

Otro factor decisivo es la posibilidad de usar nuestro modelo en entornos de almacenes de datos (del inglés, Data Warehouse - DW) donde, entre las múltiples dimensiones que estos alojan, la dimensión temporal suele encontrarse siempre presente. Esta situación hace que de forma fácil y sencilla podamos obtener innumerables series de datos temporales con las que trabajar, y sobre las que sería muy interesante aplicar nuestras técnicas para obtener resúmenes lingüísticos.

2.2. Resumen de series de datos temporales

La presente sección nos servirá para presentar de forma general las series de datos, en particular series de datos temporales, y la importancia del análisis que de ellas se realiza dentro de diferentes campos de estudio. Comenzaremos con una visión histórica del uso y estudio de las series de datos temporales. A continuación repasaremos diversos conceptos y conocimientos relacionados con las series de datos así como sus diversas aplicaciones.

2.2.1. Introducción

Desde los tiempos más remotos, el ser humano ha medido el paso del tiempo con diferentes métodos y herramientas, algunos de ellos de gran precisión. A pesar de ello, el estudio de las series de tiempo en sí posee un origen relativamente reciente. Se piensa que fue hace aproximadamente 1000 años cuando se produjo la primera representación gráfica de los eventos dividiendo un eje horizontal en intervalos de igual amplitud para representar iguales periodos de tiempo. Es a partir del siglo XIX cuando, a través del uso de la estadística teórica, el estudio se hace más extensivo debido al interés que suscitaba el análisis de las series temporales generadas en campos como la economía, la demografía o la física, entre otros.

En este trabajo tomaremos las ideas de Kendall cuando afirma que "podemos considerar el tiempo como un flujo que corre a lo largo de un mundo lleno de fenómenos a un paso uniforme. Delimitar puntos en el tiempo resulta fácil y nos permite medir los intervalos entre ellos con gran precisión" [87].

Más información acerca de la evolución histórica en el campo del estudio de series temporales a través de la estadística puede encontrarse en trabajos como [151] o [118].

2.2.2. Series temporales

En este trabajo seguiremos las ideas de Peña cuando afirma que "una serie temporal es el resultado de observar los valores de una variable a lo largo del tiempo en intervalos regulares (cada día, cada mes, cada año, etcétera)" [118]. Las primeras series temporales estudiadas correspondían a datos astronómicos y meteorológicos.

Una vez presentado el concepto de serie temporal pasemos a ver lo que consideraremos como *longitud* de la misma. Cuando nos referimos a la longitud de las series de datos temporales es muy usual el pensar que viene determinada por el periodo de tiempo comprendido entre el inicio y el fin de la serie recogida. Sin duda, este comportamiento sería el apropiado si el fenómeno se ha recogido de forma continua. Sin embargo, el uso común en análisis de series temporales es considerar como longitud el número de medidas tomadas en intervalos regulares, sea cual sea el intervalo de tiempo cubierto. Por ejemplo, una serie de longitud 60 es aquella cuyos datos se han tomado en 60 intervalos regulares de tiempo con independencia de si se han hecho en un minuto, una hora o un año.

Además de la longitud, podemos encontrar que las series de datos vienen caracterizadas por otras propiedades. Algunas de estas propiedades de las series pueden ser la *estacionariedad* y *estacionalidad*. Diremos que una serie es *estacionaria* o constante si los valores oscilan alrededor de un valor constante (si esto ocurre la media y la variabilidad se mantienen constantes a lo largo del tiempo). Si las series no cumplen esta propiedad son denominadas como *no estacionarias* (la media y/o variabilidad cambian a lo largo del tiempo). Las series no estacionarias pueden mostrar cambios en la varianza, o mostrar una tendencia, es decir, que la media crece o decrece a lo largo del tiempo. Por otro lado, cuando la serie es no estacionaria, si se observa que un mismo comportamiento se repite a lo largo del tiempo, diremos que la serie es *estacional*. Un ejemplo muy claro de estacionalidad se puede observar en algunas series temporales en las que los valores o el valor medio de la variable observada depende del mes considerado. Este fenómeno es bastante frecuente en series de variables económicas, sociales o climáticas.

Ejemplos de resúmenes lingüísticos que pongan de manifiesto la estacionalidad de una serie serían, *"Normalmente, en la provincia de Granada, durante el mes de Enero, las temperaturas máximas son bajas"* o *"De forma habitual, la ocupación hotelera durante el verano en la Costa Mediterránea es alta o muy alta"*. Aunque se ha comentado anteriormente que, por lo general, las series no suelen ser estacionarias a lo largo del tiempo, si trabajamos con periodos de tiempo acotados podemos tener series razonablemente estacionarias que nos permitan hacer resúmenes como *"Durante*

el año pasado, el precio de la patata se mantuvo entre 1 y 1.5 euros el kilo" o *"Las precipitaciones en Enero de este año han sido escasas".*

En cierto modo el modelo que presentaremos saca partido de estas características de las series temporales para, marcando un contexto lingüístico adecuado, realizar los resúmenes de la información que contienen.

Representación de series de datos temporales

Las series de datos temporales pueden ser representadas mediante una sucesión de las medidas tomadas. Si estas medidas no se han tomado en intervalos regulares, o necesitamos obtener más información acerca del momento de tiempo al que corresponden, dichas medidas pueden ser acompañadas por el instante de tiempo concreto con mayor o menor nivel de detalle dependiendo de nuestras necesidades. Sin embargo, estas formas de representación, bien sea en texto plano o mediante tablas, no suelen ser muy intuitivas.

En algunas ocasiones puede que el usuario que recibe la información no posea conocimiento experto en el tema específico. Otras veces, puede que la cantidad de datos sea tan elevada o la diferencia entre sus valores tan notable, que hagan complicado el proceso de análisis de los datos. Sea como fuere, incluso con las series de datos más sencillas, en muchos de los casos la representación gráfica aporta una buena herramienta de representación de las series de datos temporales. Por desgracia, la representación gráfica de las series no siempre es fácil de interpretar, ya que en ciertas circunstancias las series son muy complicadas o incluso tenemos varias series relacionadas entre sí presentadas en el mismo gráfico.

Un problema notable del que adolece la representación gráfica de series es que necesita de un dispositivo gráfico, bien pantalla, papel o similares, para poder mostrarse, y no en todas las situaciones es posible contar con el equipamiento técnico adecuado. Pero, incluso si contamos con el equipamiento adecuado, siguen existiendo inconvenientes en este tipo de representación, por ejemplo, si no cuentan con la resolución adecuada. Este sería el caso si necesitáramos mostrar un gráfico complejo en una pantalla de un dispositivo móvil de pequeño tamaño.

Puede ocurrir asimismo que la persona a la que está destinado el resumen no pueda ver adecuadamente la pantalla. Incluso cuando contamos con los medios necesarios que, además, poseen la resolución adecuada, en determinadas situaciones se requiere algo más. Véase como ejemplo el caso de personas con reducida capacidad visual. En esta situación, como en las anteriores se hace necesario otra herramienta que nos ayude a mostrar la información.

En todos los casos, con independencia de la complejidad de la serie o series, el resumen lingüístico de series de datos temporales es un herramienta potente que permite

presentar, a usuarios no expertos y/o con ciertas necesidades especiales, información acerca de la serie en un formato comprensible y fácil de interpretar.

Veamos como ejemplo los datos referentes a la ocupación hotelera en España en el año 2008. Más específicamente nos centraremos en los puestos de trabajo que esta ocupación acarrea (datos obtenidos de la página del Instituto Nacional de Estadística, INE, [30]).

Los datos numéricos obtenidos de la página del INE son los presentados a continuación:

154.713, 164.123, 181.012, 188.959, 218.808, 231.528
244.738, 249.026, 236.581, 207.809, 160.979, 153.603

Como se puede apreciar, los datos por sí solos puede que sean útiles para realizar análisis complejos por personas u ordenadores, pero no resultan muy intuitivos a la hora de ser entendidos por los usuarios sin formación. Esta situación mejora en la Tabla 2.2, donde los datos aparecen tabulados y se han insertado los meses correspondientes a cada medición. En este caso, las personas poseen más información y un formato mucho más agradable y amigable para ellos. Aún así cuando la cantidad de datos crezca o su variabilidad sea elevada puede que se necesite alguna ayuda extra.

Mes del año	Personal ocupado
Enero	154.713
Febrero	164.123
Marzo	181.012
Abril	188.959
Mayo	218.808
Junio	231.528
Julio	244.738
Agosto	249.026
Septiembre	236.581
Octubre	207.809
Noviembre	160.979
Diciembre	153.603

Tabla 2.2: Personal ocupado durante el año 2008 en el sector hotelero.

En la Figura 2.3 se presentan los datos anteriores. En el eje X se ha representado el tiempo dividido según los diferentes meses del año; y en el eje Y aparece el número de personas ocupadas. La representación mediante gráficos nos ayuda a obtener más información en menos tiempo. Con una sola ojeada podemos entender qué se representa, la tendencia general, los meses con valores máximos y mínimos, etc.

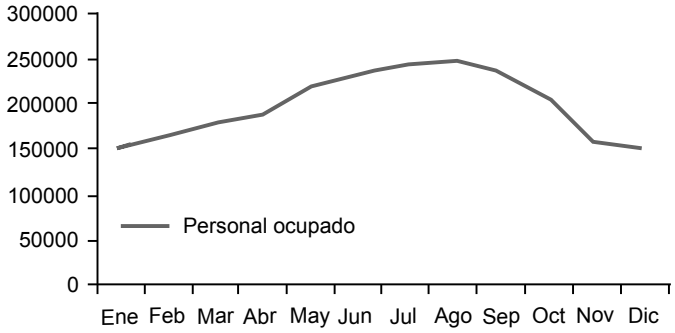


Figura 2.3: Personal ocupado durante el año 2008 en el sector hotelero.

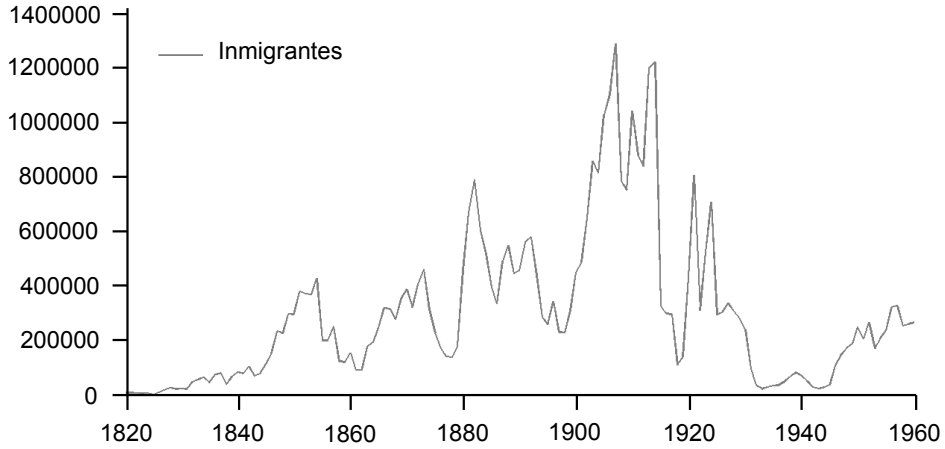


Figura 2.4: Inmigración en USA desde 1820 a 1962.

¿Pero qué ocurre cuando incluso la representación gráfica es difícil de entender porque no contamos con los medios adecuados? Algunos ejemplos de situaciones en los que la representación gráfica no aporta la suficiente información o no lo hace con claridad e interpretabilidad se presentan en las Figuras 2.4 y 2.5.

La Figura 2.4 muestra la inmigración hacia los Estados Unidos desde el año 1820 hasta 1962. Los datos han sido tomados del libro de Kendall [87], y aunque no incluyen los valores de los últimos años, nos sirven para ejemplificar una serie de datos cuya representación gráfica no es tan sencilla. La serie abarca un periodo de tiempo más amplio que en el ejemplo anterior. Además, la variabilidad es muy alta, y si no contamos con la resolución suficiente podríamos no apreciar bien ciertos matices. Del mismo modo, podría ocasionar problemas si el número de datos es muy elevado ya que, a simple vista, podemos ver tendencias que, al usar un mayor nivel de detalle, desaparecen (y viceversa).

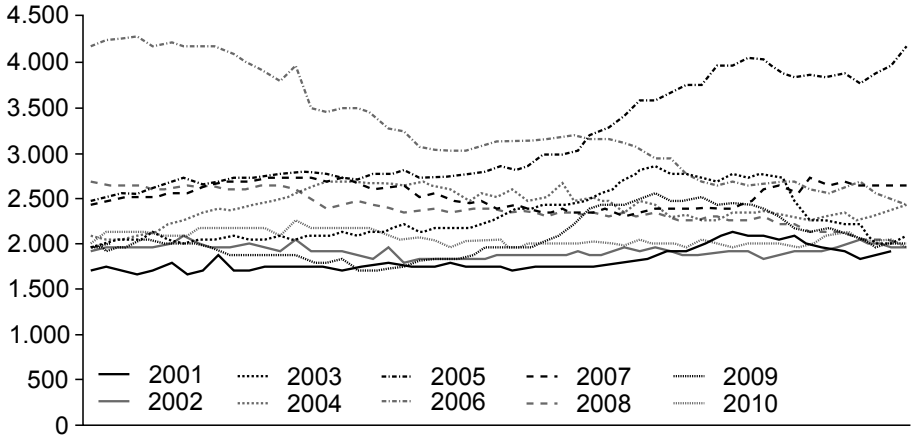


Figura 2.5: Precios del aceite de oliva en España desde 2001 a 2010.

La Figura 2.5 muestra el precio medio (euros por tonelada) del aceite de oliva virgen extra en España en el periodo desde 2001 a 2010, ambos inclusive. Los datos han sido tomados de la página web POOLRed - Sistema de Información de precios en origen de mercado de contado del aceite de oliva [31]. En esta ocasión, los problemas comentados anteriormente se ven agravados debido a la cantidad de información representada en el mismo espacio.

En estos casos, entendemos que la utilización de resúmenes lingüísticos es una alternativa muy interesante y que merece la pena ser tenida en cuenta.

2.2.3. Análisis de series temporales

El propósito del estudio o análisis de las series de datos temporales puede ser dividido en dos grandes áreas. La primera de ellas es "entender o modelar el mecanismo estocástico de una serie observada" y la otra "predecir o pronosticar los valores futuros de series basadas en la historia de esas series y, posiblemente, otras series o factores relacionados" [27]. De modo que podemos decir que el análisis de series temporales comprende métodos que ayudan a interpretar los datos, extrayendo para ello información representativa, tanto referente a los orígenes o relaciones subyacentes como a la posibilidad de extrapolar y predecir su comportamiento futuro. De hecho uno de los usos más habituales de las series de datos temporales es su análisis para predicción.

En nuestro caso no vamos a centrar nuestra atención en áreas como la predicción, sino en la posibilidad de la descripción de las series. La tarea de obtención de la información subyacente en las series de datos temporales es importante y permite conocer tendencias, eventos destacados o patrones, que permitirán una mejor toma de decisiones.

Estudio descriptivo de series de datos: Modelo clásico

El estudio descriptivo de series temporales se basa en la idea de descomponer la variación de una serie en varias componentes básicas. Este enfoque no siempre resulta ser el más adecuado, pero es interesante cuando en la serie se observa cierta tendencia o cierta periodicidad. Hay que resaltar que esta descomposición no es en general única. Este enfoque descriptivo consiste en encontrar componentes que correspondan a una tendencia a largo plazo, un comportamiento estacional y una parte aleatoria. Podemos ver la descomposición en la Ecuación (2.1).

$$X_t = T_t + S_t + I_t \tag{2.1}$$

donde T_t es la *tendencia* secular o regular que refleja el comportamiento o evolución de la serie a largo plazo, es decir, el cambio a largo plazo de la media de la serie. S_t es la *variación estacional* que representa el movimiento periódico de corto periodo y se debe a la influencia de ciertos fenómenos que se repiten de forma periódica en un año (las estaciones), una semana (los fines de semana), un día (horario laboral), o cualquier otro corto periodo establecido. Por último, I_t recoge *variaciones aleatorias* que afectan a las componentes anteriores. Estas variaciones son debidas a fenómenos de carácter ocasional, accidental o errático, como pueden ser tormentas, terremotos, inundaciones, huelgas, guerras, atentados, etcétera.

En ocasiones se puede considerar una cuarta componente denominada *variación cíclica*, C_t. Esta componente es la equivalente a la variación estacional pero considerando periodos temporales de duración superior a un año. Refleja movimientos irregulares alrededor de la tendencia cuyo período o amplitud pueden ser variables, pudiendo clasificarse como cíclicos, cuasi-cíclicos o recurrentes. La nueva ecuación obtenida es presentada a continuación por la Ecuación (2.2).

$$X_t = T_t + S_t + C_t + I_t \tag{2.2}$$

El esquema presentado se denomina esquema aditivo, pero no es el único que existe para la consecución de la serie temporal como tal. En la Ecuación (2.3) podemos ver representado el esquema multiplicativo y en la Ecuación (2.4) el esquema mixto.

$$X_t = T_t * S_t * C_t * I_t \tag{2.3}$$

$$X_t = T_t * S_t * C_t + I_t \tag{2.4}$$

Un esquema aditivo, es adecuado, por ejemplo, cuando S_t no depende de otras componentes, como T_t. Si por el contrario la estacionalidad varía con la tendencia, el modelo más adecuado es un esquema multiplicativo. El esquema multiplicativo puede ser transformado en aditivo, si aplicamos logaritmos. El problema que se presenta, es modelar adecuadamente las componentes de la serie.

Al enfrentarnos a la realización del análisis de series temporales, nos centraremos en primer lugar en el estudio de la tendencia. La aplicación de *filtros* de datos a la serie nos permite detectar la tendencia y eliminarla de la serie. Estos filtros son funciones matemáticas que aplicamos a los valores de la serie y que producen nuevas series con unas características determinadas. Entre esos filtros encontramos las *medias móviles*.

Existen otros procedimientos para extraer la tendencia, como *ajuste de polinomios* o *alisado mediante funciones exponenciales*. Para profundizar más sobre el tema se pueden consultar referencias como [13], [27], [87] o [54] entre otros.

En el proceso del estudio de la estacionalidad de una serie temporal juega un papel muy importante la *función de autocorrelación*. La función de autocorrelación mide la correlación entre los valores de la serie distanciados un lapso de tiempo k, y viene determinada por la Ecuación (2.5) dados N pares de observaciones $(y,\, x)$.

$$r = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(x_i - \bar{x})^2}} \tag{2.5}$$

Podemos considerar a la función de autocorrelación como el conjunto de coeficientes de autocorrelación r_k desde 1 hasta un máximo que no puede exceder la mitad de los valores observados. La importancia de la función de autocorrelación con respecto al estudio de la estacionalidad radica en que si ésta existe, los valores separados entre sí por intervalos iguales al periodo estacional deben estar correlacionados de alguna forma. Es decir, que el coeficiente de auto-correlación para un retardo igual al periodo estacional debe ser significativamente diferente de 0.

En este estudio de la estacionalidad, y relacionada con la función de auto-correlación, encontramos la función de auto-correlación parcial. Al igual que el anterior, en el coeficiente parcial de orden k se calcula la correlación entre parejas de valores separados por una distancia estacional k, pero, esta vez, suprimiendo el efecto debido a la correlación producida por retardos anteriores a k.

Modelos autorregresivos

Los métodos clásicos son en ocasiones insuficientes si pretendemos encontrar explicación para las muchas y variadas dinámicas de las series de datos. Los modelos autorregresivos (**AR** - autoregressive model) son creados con el objetivo de poder

explicar un valor presente de la serie de datos como una función de los p valores anteriores, $x_{t-1}, x_{t-2}, ..., x_{t-p}$, donde p representa el número de observaciones pasadas necesarias para pronosticar el valor actual. En el caso de procesos estacionarios con distribución normal, la teoría estadística de los procesos estocásticos dice que, bajo determinadas condiciones previas, toda x_t puede expresarse como una combinación lineal de sus valores pasados más un término de error.

Los denominados modelos de medias móviles (**MA** - moving average model) son aquellos que explican el valor de una determinada variable en un período t en función de un término independiente y una sucesión de errores correspondientes a períodos precedentes, ponderados convenientemente. Un modelo de medias móviles puede obtenerse a partir de un modelo autorregresivo simplemente realizando sucesivas sustituciones.

Los modelos **ARMA** (autoregressive moving average model) integran a los modelos AR y a los modelos MA en una única expresión y son usados para el análisis de series estacionarias. Por tanto, la variable x queda explicada en función de los valores tomados por la variable en períodos anteriores, y los errores cometidos en la estimación.

Los modelos ARMA son el punto de partida para la construcción de los modelos **ARIMA** (autoregressive integrated moving average model) para el análisis de las series no estacionarias. Este tipo de modelos son también conocidos como modelos Box-Jenkins debido a los estadísticos George Box y Gwilym Jenkins, los cuales aplicaron modelos ARMA y ARIMA para realizar pronóstico en su trabajo con series de datos temporales.

El modelo ARIMA se basa en la aplicación del modelo ARMA sobre las series de tiempo después de pasar por un proceso mediante el cual se usan las diferencias entre valores consecutivos en lugar de usar los valores en sí.

Más información sobre el proceso de análisis de series temporales mediante los modelos ARMA o ARIMA pueden encontrarse en [118], [144], [56], [55], [53], [14], [161], [69], [165], [88], [20] o [150] entre otros.

Todos estos métodos son de probaba exactitud y corrección, pero, ¿son fácilmente utilizables por personas sin conocimientos específicos? Por desgracia, la respuesta es no. ¿Son los resultados amigables para las personas sin conocimientos específicos? La respuesta es que tampoco. Para evitar esta situación de "desamparo" de los usuarios no expertos, si no en el uso de las herramientas, al menos si en la interpretabilidad de los resultados, surge una nueva forma de análisis de datos, el resumen de datos con técnicas relacionadas con la minería de datos y áreas relacionadas.

Minería de datos

Existen diferentes enfoques cuando se realiza abstracción o resumen de series de datos, en relación a su aplicación en campos como el descubrimiento de conocimiento (del inglés, Knowledge Discovery in Databases - KDD) o la minería de datos (del inglés, Data Mining - DM).

Siguiendo a Höppner diremos que "la abstracción o resumen corresponde a la segmentación de las series de datos y a la caracterización de los datos dentro del segmento. Por segmento de la señal entendemos una secuencia de medidas soportadas por el intervalo contiguo más amplio en el que cierta propiedad se cumple" [64].

El proceso de segmentación de la serie puede realizarse de modo supervisado o no supervisado. En el primero de los casos se definen los atributos de interés y el conjunto de etiquetas que los describen a priori. En cambio, en el segundo de los casos, esta información no se conoce y debe ser aprendida a partir del conjunto de datos.

Podemos considerar cuatro grandes enfoques cuando hablamos de los modelos inductivos o no supervisados: Clustering de secuencias embebidas (más en [141], [117], [84], [29]), clustering de modelos embebidos, clustering mediante warping cost [139] y clustering usando modelos de Markov (más en [145], [143]).

En los modelos deductivos o supervisados el proceso de segmentación se realiza buscando los puntos en los que las primeras derivadas toman el valor cero. En muchas ocasiones estos valores son introducidos por el ruido, de modo que debe ser eliminado mediante funciones de suavizado o aproximación.

En general, los métodos inductivos son bastante costosos en términos computacionales. En parte, ello es debido a que el número de conjuntos o "clusters" obtenidos es bastante grande. Por otro lado, los métodos deductivos asumen que el nivel de ruido a lo largo del tiempo es constante, lo cual no es siempre cierto. Para solventar estas limitaciones surgen los métodos multi-escala.

Los métodos multi-escala basan su funcionamiento en obtener varias abstracciones a diferentes escalas en lugar de partir de unos parámetros iniciales y abstraer las series de tiempo a una sola escala. Una característica llamativa en este tipo de modelos es que el proceso de suavizado debe ser altamente intuitivo, de modo que se eliminen máximos y mínimos en lugar de crear nuevos.

Una buena manera de representar las diferentes abstracciones teniendo en cuenta diferentes escalas es usando el árbol intervalar de escalas (del inglés Interval Tree of Scales) donde la coordenada x representa el tiempo mientras que la coordenada y representa la escala que se tiene en cuenta para realizar el proceso de abstracción en diferentes intervalos. Más sobre esta herramienta puede ser encontrado en los trabajos de Höppner [62–64].

Este mismo investigador en sus trabajos [65] y [61] trabaja en el descubrimiento de patrones temporales y reglas informativas. Para llevar a cabo esta tarea utiliza la lógica intervalar de Allen, de modo que las descripciones obtenidas podrían ser del modo *A antes de B* o *A solapa a B* entre otros.

Aunque los métodos expuestos anteriormente siguen usándose mucho, en el caso específico del resumen lingüístico se necesita algo más. Existe una necesidad acuciante de incorporar conocimiento adicional referente al contexto a la hora de realizar la segmentación. Aunque se puede usar una estrategia de segmentación de los tipos comentados anteriormente, corremos el riesgo de que los segmentos encontrados sean muy precisos y correctos pero que no nos satisfagan por no ser intuitivos o no ser lo que esperábamos. En otras palabras, es una buena práctica tener en cuenta el contexto a la hora de realizar el proceso de segmentación para asegurarnos de que ofrecemos al usuario información que le sea útil y relevante.

La incorporación de información contextual al proceso puede realizarse de diversas formas. En esta tesis estamos interesados en técnicas para definir el marco lingüístico del problema basadas en Soft Computing.

2.3. Uso del Soft Computing en resumen de datos

Como se ha indicado en secciones anteriores, en este documento se considera que el uso de resúmenes lingüísticos favorece la presentación de información al usuario, y que por tanto facilita el manejo y entendimiento de dicha información por parte del mismo. Asimismo, se ha presentado el problema tan interesante que surge en la actualidad al intentar que estos resúmenes lingüísticos se generen de forma automatizada por ordenadores tratando de imitar el comportamiento humano. A este respecto, el uso de técnicas Soft Computing son de gran ayuda para salvar las distancias en la comunicación eficiente entre máquina-humano, que tan difícil se revela.

Fue Zadeh quien en 1994 propuso una primera definición de Soft Computing (SC) estableciéndola en los siguientes términos [181]:

> *"Básicamente, Soft Computing no es un cuerpo homogéneo de conceptos y técnicas. Mas bien es una mezcla de distintos métodos que de una forma u otra cooperan desde sus fundamentos. En este sentido, el principal objetivo de la Soft Computing es aprovechar la tolerancia que conllevan la imprecisión y la incertidumbre, para conseguir manejabilidad, robustez y soluciones de bajo costo. Los principales ingredientes de la Soft Computing son la Lógica Difusa, la Neurocomputación y el Razonamiento Probabilístico, incluyendo este último a los Algoritmos Genéticos, las Redes de Creencia, los Sistemas Caóticos y algunas partes de la Teoría de Aprendizaje. En*

esa asociación de Lógica Difusa, Neurocomputación y Razonamiento Probabilístico, la Lógica Difusa se ocupa principalmente de la imprecisión y el Razonamiento Aproximado; la Neurocomputación del aprendizaje, y el Razonamiento Probabilístico de la incertidumbre y la propagación de las creencias".

Como queda reflejado en la definición, se puede considerar a la Soft Computing como un conglomerado de técnicas individuales en las que metodologías difusas (del inglés, fuzzy) se usan de una u otra manera. De este modo se intentan superar las dificultades que surgen al tratar de solucionar problemas reales en los que, de forma natural, existe un alto componente de imprecisión e incertidumbre.

Una de las herramientas incluidas en el Soft Computing que nos servirán para acercar, en términos comunicativos, a máquinas y seres humanos, son los conjuntos difusos (del inglés, Fuzzy Sets - FS). En [34] podemos encontrar una disertación acerca de la necesidad de usar lenguaje con expresiones vagas o difusas con el objetivo de hacer que los textos generados sean más amigables para los seres humanos, que serán los receptores últimos de los resultados. Es en estas situaciones en las que los conjuntos difusos y las variables lingüísticas juegan un papel clave.

2.3.1. Conjuntos difusos y variables lingüísticas

Como Zadeh presentó en su trabajo [178], un *conjunto difuso* es un conjunto sin un límite definido, es decir, la transición entre "pertenecer a un conjunto" y "no pertenecer a un conjunto" es gradual. Dicha transición suave es caracterizada por una función de pertenencia que toma valores en el intervalo [0,1] en lugar de hacerlo en el conjunto {0,1}. De este modo, se puede considerar el concepto de conjunto difuso como una generalización del concepto básico de conjunto.

Los conjuntos definidos de forma imprecisa desempeñan un papel importante en el pensamiento humano, particularmente en los dominios del reconocimiento de patrones, de la comunicación de la información o de la abstracción entre otros. En nuestro caso, los conjuntos difusos nos aportan una inestimable ayuda a la hora de obtener resultados lo más amigables posibles para los receptores de los resúmenes, es decir, los seres humanos.

Muy relacionadas con los conjuntos difusos encontramos a las llamadas *variables lingüísticas*. Podemos considerar a las variables lingüísticas como variables cuyos valores se representan mediante términos lingüísticos, de modo que el significado de estos términos se determina mediante conjuntos difusos.

Las variables lingüísticas nos brindan la oportunidad de modelar de una forma más cercana a la realidad conceptos del mundo real. En la mayor parte de las oca-

siones existen muchos problemas al intentar modelar conceptos del mundo real con variables clásicas o "crisp". En este sentido, las variables lingüísticas nos ofrecen mayor versatilidad para modelar conceptos. Por esta razón, son una herramienta muy ampliamente utilizada en diversos métodos de resumen cuando en el proceso aparece implicado el lenguaje natural.

Mediante el uso de variables lingüísticas podemos convertir la sentencia *"Los chicos de mi despacho tienen alturas de 1.80, 1.85, 1.78 y 1.92"* que para un usuario, aunque lo comprenda, es difícilmente manejable o fácil de recordar; por un resumen del estilo de *"Los chicos de mi despacho son altos"* que resulta más amigable.

Los conjuntos difusos y las variables lingüísticas nos ayudan en el proceso de convertir datos numéricos, que poco o nada dicen a los seres humanos, a sentencias compuestas por términos lingüísticos. Recordemos que parece adecuado pensar que la mejor forma de presentar resultados a los seres humanos es mediante lenguaje natural, ya que es su manera habitual de comunicarse e interactuar entre ellos, de pensar, de crear, de soñar, ...

Para facilitar el mencionado proceso de comunicación, además de los términos lingüísticos, es necesario establecer una estructura de presentación adecuada. Existen diferentes plantillas o esquemas, siendo las más ampliamente utilizadas herramientas como las reglas de asociación o las sentencias cuantificadas. Como se desarrollará posteriormente, el modelo que presentamos en esta tesis hará uso de estas segundas, en su forma concreta *"Q de D son A"* (Capítulo 3).

2.3.2. Un problema de optimización

Debido a la complejidad inherente al contexto lingüístico, el número de posibles resúmenes de un mismo conjunto de datos se hace inmanejable. Como hemos comentado anteriormente, no existe un resumen que sea el mejor, sino un conjunto de ellos, que satisfarán en mayor o menor medida al usuario dependiendo del contexto. Esto no quita que de alguna manera no podamos establecer una ordenación con los que nos satisfacen más. Para ello se utilizarán las medidas de calidad que se aplican a los resúmenes.

Como consecuencia de la amplitud del espacio de búsqueda, podemos considerar que la búsqueda de un resumen adecuado, o el más adecuado, se asemeja a un problema de optimización global.

La Figura 2.6 nos muestra una clasificación en tres grandes categorías de los enfoques de optimización globales, así como algunos de los ejemplos más significativos en algunas de ellas. Dicha categorización ha sido tomada de [25][1].

[1]En la Figura 2.6, los nombres de las diferentes técnicas se han conservado en inglés debido a que muchos de ellos son términos muy extendidos que no se suelen traducir.

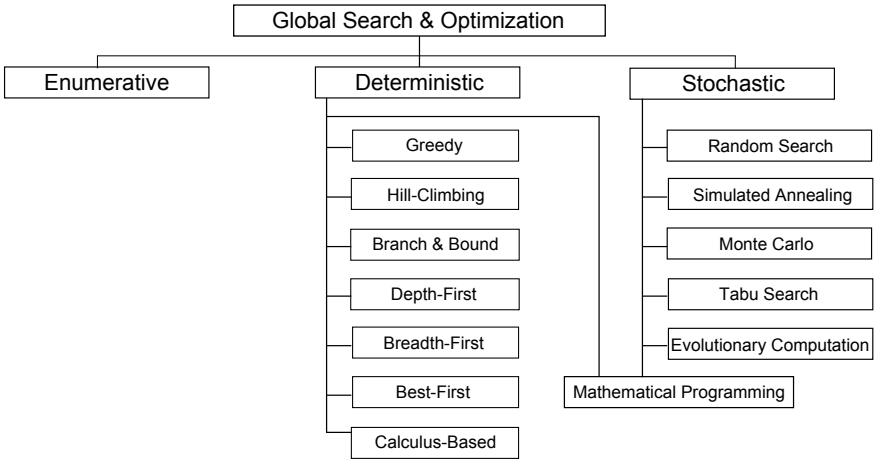


Figura 2.6: Enfoques de optimización global.

El *esquema enumerativo* engloba a las que se pueden considerar las estrategias de búsqueda más sencillas. Este esquema consiste en realizar un barrido exhaustivo de las soluciones para encontrar la mejor de ellas, también denominada como el máximo global. Como se puede suponer, este método consume mucha memoria, pero sobre todo, consume una cantidad ingente de tiempo; esta situación hace que el problema sea poco escalable cuando crece el espacio de búsqueda. Desgraciadamente, los problemas de la vida real se caracterizan por contar con un espacio de búsqueda bastante grande.

Los *algoritmos determinísticos* intentan solucionar el problema de la escalabilidad incorporando al problema conocimiento del dominio, o como le hemos llamado hasta ahora, información del contexto o contextual; de esta forma se acota el espacio de búsqueda que ahora ya no se exploraría en su completitud. Las soluciones que se obtendrían usando estos métodos no tienen porque ser los máximos globales, pero seguro que son máximos locales, es decir, soluciones aceptables o lo suficientemente buenas.

Los *enfoques estocásticos* surgen como una alternativa para resolver problemas irregulares, como ha quedado dicho que suelen ser todos los de la vida real. Éstos enfoques insertan una componente aleatoria que marca la forma de explorar el espacio de soluciones.

En esta memoria se abordará el problema de resumen lingüístico mediante diversas técnicas de optimización global (Capítulo 4). En primer lugar se presentará una estrategia de resumen determinística basada en un enfoque Greedy. A continuación hacemos uso de una técnica estocástica, como es la computación evolutiva y más

concretamente los algoritmos genéticos *multi-objetivo*. Estos últimos, si recordamos la definición que dábamos al principio de la sección, son también parte de la Soft Computing.

Cuando hablábamos sobre crear una ordenación o "ranking" de resúmenes, comentamos que para hacerlo tendríamos en cuenta los diferentes ítems que usamos para determinar si un resumen es de calidad o no, o de si su calidad es mayor o menor que la de un segundo resumen. Como podemos intuir, de entre las medidas de calidad, también conocidas en este ámbito como objetivos, habrá algunas que sean complementarias o contrarias a otras. De modo que mejorar en un aspecto nos puede llevar a empeorar otro.

Esta situación se denomina "problema multi-objetivo" y para enfrentarse a ella deberemos llevar a cabo un proceso de *optimización multi-objetivo* (también conocido como optimización multi-criterio o multi-atributo). En general, en esta tipología de problemas no se pretende optimizar un sólo objetivo, sino dos o más de ellos que habitualmente son contradictorios y además suelen estar sujetos a ciertas restricciones.

Los problemas multi-objetivo no son una excepción en nuestro entorno, sino más bien una constante. No es extraño encontrar este tipo de problemas en numerosos campos como el diseño de productos y procesos, las finanzas, el diseño de automóviles, aviones, etcétera. Sin ir más lejos, nosotros mismos podríamos querer invertir nuestro dinero, y para ello querremos obtener los mayores beneficios posibles asumiendo el menor riesgo posible. O más sencillo aún, planear un fin de semana con los amigos donde lo que se busca el maximizar la diversión y las actividades a realizar pero haciendo una inversión de dinero que sea asumible para nuestros bolsillos.

En este tipo de problemas, para que una solución se pueda considerar mejor que la otra debe ser mejor en todos y cada uno de los objetivos del problema. Del mismo modo, para que sea peor, debe ser peor en todos y cada uno de los objetivos. Pero, ¿qué pasa cuándo una solución A satisface mejor un objetivo y una solución B satisface mejor otro? En este caso diremos que las soluciones no se dominan entre ellas. Habitualmente cuando se trata de estos problemas, intentar mejorar un objetivo puede repercutir en el empeoramiento del otro. Normalmente las soluciones no dominadas se suelen incluir en conjuntos denominados *frentes de Pareto*.

Existe una rama dentro de la computación evolutiva que se encarga de dar solución a este tipo de problemas de optimización multi-objetivo, estos son los llamados algoritmos evolutivos multi-objetivo. Los MOEAs (del inglés, Multi Objective Evolutionary Algorithms) son muy populares optimizando problemas multi-objetivo y la mayoría de ellos están basados en esquemas de establecimiento de rankings de Pareto.

2.4. Enfoques en la realización de resumen lingüístico

En esta sección se presentan algunos de los diversos enfoques presentes en la literatura, con el fin de realizar resumen lingüístico sobre las series de datos temporales o que nos pueden ser de utilidad para hacerlo.

2.4.1. El resumen lingüístico y las técnicas Soft Computing

En trabajos como los de Mitra y otros [103] y Chen y otros [22] podemos encontrar completos estados del arte sobre técnicas de obtención de información a partir de grandes cantidades de datos mediante el uso de técnicas Soft Computing. Aunque dichos trabajos no se centran únicamente en la obtención de resúmenes lingüísticos, si que aparecen en alguna de las secciones. En general son trabajos bastante interesantes en los que se nos ofrece una composición de situación y un buen punto de partida para el estudio de otros trabajos más concretos y actuales. En [48] de Fu, se presenta un completo estado del arte acerca de la minería en series de datos. Es un trabajo amplio y útil que nos da una visión completa y que dedica una sección al resumen lingüístico de datos.

En [7] los investigadores I. Z. Batyrshin y L. Sheremetov nos presentan un extenso estudio sobre diferentes técnicas de minería de datos basadas en percepciones, es decir, haciendo un uso intensivo de diversos conceptos relacionados con el Soft Computing, Computing with Words (computación con palabras), etcétera. Entre las mencionadas técnicas de minería de datos se hace referencia al resumen en lenguaje natural.

Existen distintos enfoques a la hora de afrontar la tarea de producir un resumen lingüístico de datos. Como denominador común cabe destacar que la gran mayoría de las técnicas hasta ahora desarrolladas hacen uso de etiquetas lingüísticas. Como ya se presentó con anterioridad, las etiquetas lingüísticas son esenciales ya que introducen la posibilidad de tratar con la imprecisión y vaguedad necesarias para trabajar con el lenguaje natural.

2.4.2. Las propuestas de Yager

El autor Ronald R. Yager aborda la confección de resúmenes lingüísticos apoyándose en el uso de sentencias enriquecidas por conceptos difusos [167] (introducidos éstos años antes por Lofti A. Zadeh). Dichas sentencias representan el grado de satisfacción de una cierta cualidad por un grupo de objetos. Para ello se cuenta con un resumidor S, una cantidad de acuerdo Q y una medida de validez o verdad.

De esta forma si contamos con un conjunto de datos D de la forma $D = \{25, 13, 19, 37, 25, 56, 45, 73\}$ representando un conjunto de edades, podremos tratar de obtener la cantidad de individuos del grupo que cuentan aproximadamente con 15 años usando $S =$ *"aproximadamente 15"* y como $Q =$ *"algunos"*; o saber si la mayoría de la gente

es adulta usando S = *"adulta"* y Q = *"la mayoría"*. En estos casos, T representa el valor de verdad de la afirmación. El cálculo del valor de T se realiza usando el cardinal de Zadeh [170, 172, 175]. Posteriormente, y para la resolución de casos más complejos, Yager desarrolló los denominados operadores OWA (Ordered Weighted Averaging) [169] y los aplicó en la evaluación de sentencias para la obtención de resúmenes lingüísticos en [173, 174].

En [168] Yager trabaja con el concepto de sentencias cuantificadas (extensión de las sentencias enriquecidas anteriores) para el apoyo a la toma de decisiones multi-objetivo de forma lingüística. En este caso se usan sentencias cuantificadas con la siguiente estructura: *"Q Es son A"* y *"Q BEs son A"* donde Q es un cuantificador lingüístico (como el Q anterior) y A y B son subconjuntos difusos que cumplen ciertas propiedades.

En otro trabajo posterior del mismo autor, esta vez con la colaboración de Petry [176], se trabaja con la introducción de ontologías de conceptos para construir un enfoque multi-criterio del resumen.

Desde que Zadeh y Yager sentaran las bases del uso de conjuntos difusos para de alguna forma agregar o resumir datos y presentarlos de forma lingüística, muchos son los investigadores que han seguido sus pasos. A continuación veremos algunos de los más interesantes divididos en una serie de apartados que nos ayudarán a formar una composición de lugar.

2.4.3. Obtención de los mensajes

En la presente subsección nos centraremos en la clasificación de los distintos métodos o modelos propuestos por diversos investigadores según el tipo de sentencias que usan para la composición de los mensajes del resumen o la forma en la que se han construido las mismas.

Sentencias enriquecidas o cuantificadas

J. Kacprzyk y otros usan en sus trabajos resúmenes representados por sentencias cuantificadas [70, 76–78, 80–82]. En ellos se explota el uso de distintas protoformas con las que construir diferentes tipos de resúmenes. Cada uno de los tipos de resumen será usado por el usuario dependiendo de sus necesidades concretas en el momento determinado en el que toma lugar el proceso de resumen. En este caso se presenta a los usuarios una colección de sentencias ordenadas por el grado de cumplimiento, de modo que se tenga acceso rápido a las sentencias que mejor describen los datos.

En la misma línea encontramos *Quantirius*, una herramienta desarrollada por Daniel Pilarski en [123]. El modelo realiza resumen lingüístico de bases de datos, y al igual que en el caso anterior lo realiza mediante el uso de sentencias cuantificadas.

En este caso, una vez obtenidas las sentencias de resumen se llevan a cabo dos fases de reducción de las mismas con el fin de obtener un producto final manejable para el usuario. En la primera fase de reducción se buscan términos lingüísticos usados que puedan estar incluidos en otros del resumen y se eliminan. A continuación, en la segunda fase se lleva a cabo una reducción por superposición de términos lingüísticos unimodales.

Debemos decir que el tipo de resúmenes obtenidos en los anteriores trabajos son de diferente naturaleza que los que nosotros perseguimos, además de intentar cumplir diferentes objetivos. Destacaremos que en los anteriores un resumen es una sola sentencias mientras que para nosotros un resumen será un conjunto homogéneo y completo de sentencias.

A. Niewiadomski es otro de los investigadores que usan las sentencias cuantificadas para expresar en lenguaje natural el resumen obtenido de un conjunto de datos dado. En el trabajo [112] se hace uso de los conjuntos difusos de tipo 2, como generalización de aquellos de tipo 1, y se presentan ejemplos donde se muestra la utilidad de los mismos. En [113] se aborda de nuevo la misma línea de acción pero esta vez haciendo un estudio más en profundidad sobre ciertos aspectos. En el primero de los trabajos se asumía de manera implícita que los universos de discurso de las expresiones lingüísticas eras discretos, sin embargo en este trabajo se extiende la idea para poder tratar con conjuntos de dominio continuo, tanto si son finitos como si son infinitos.

Otro de los autores que últimamente se ha interesado por la potencia que ofrecen los resúmenes lingüísticos es James Keller. En sus trabajos se usan protoformas de la forma *"X_c is S_i in P_k for T_j"*, obteniendo resúmenes de la forma *Derek está de pie en el laboratorio durante 11 segundos* que en una segunda fase transforman en *Derek está de pie en el laboratorio durante un periodo moderado de tiempo*. Dichas protoformas son modificaciones o adaptaciones de las sentencias cuantificadas nombradas anteriormente. Este tipo de información se utiliza en [2, 3] para realizar el resumen lingüístico del comportamiento de cierto individuo a través de imágenes tomadas en un determinado lugar con ambiente controlado por videocámaras. En concreto, es interesante el estudio que realizan acerca de si la persona que se esta estudiando ha sufrido una caída y cómo la lógica difusa puede ser aplicada en dicha tarea. Además de lógica difusa y visión en estéreo, se han usado algoritmos genéticos para realizar el seguimiento del sujeto de estudio.

Otra línea relacionada con la anterior en la que Keller ha colaborado con M. Ros y otros [136] es realizar resúmenes lingüísticos de patrones de comportamiento con el fin de detectar cambios en la conducta habitual de las personas estudiadas. En este caso lo que se hace es obtener resúmenes de curvas de inferencia difusas que representan los estados del objeto tridimensional que representa a la persona. En [4] de Anderson y otros el interés se centra en presentar un informe lingüístico al

estimar la edad de una persona con fines forenses. Dicho trabajo es una ampliación de uno anterior en el que no se empleaban conjuntos difusos para realizar la tarea. Por último presentamos [166] de Wilbik y otros, donde a partir de mediciones obtenidas a través de sensores colocados en un entorno controlado, se puede obtener un resumen lingüístico con la conducta de un sujeto en dicho entorno.

Como se puede apreciar, el conjunto de las técnicas presentadas son aplicadas a situaciones de la vida cotidiana de modo que son trabajos altamente implementables una vez desarrollados. Sin embargo, en ocasiones puede resultar que los resúmenes sean en cierta medida abrumadores por su detalle y cantidad de información.

En [185] de Zhang los resúmenes lingüísticos se realizan usando Degree Theory y FCA (Formal Concept Analysis). El autor diferencia entre el proceso que se debe seguir para obtener lo que él denomina resúmenes simples (sentencias cuantificadas) y resúmenes complejos (agregación de sentencias cuantificadas). En este caso, los resúmenes simples sirven de punto de partida para ir obteniendo resúmenes complejos mediante el uso de conjunciones lógicas. Es posible elegir entre ellos dependiendo de la situación o el uso que se les vaya a dar a los mismos.

Otro trabajo interesante es el de Carrasco y Villar [16]. En él se presenta un modelo para realizar resumen lingüístico de fuentes de datos heterogéneas como pueden ser diferentes páginas web relacionados con el turismo. A pesar de que la fuente principal no son sólo datos, el trabajo es mencionado aquí por la cantidad de herramientas relacionadas con el Soft Computing que son usadas y por la profusión de referencias a otros trabajos que, a su vez, pueden ser de interés.

En trabajos como [39, 40, 126], Bugarín y otros presentan su visión a la hora de construir resúmenes lingüísticos. Bugarín y Díaz-Hermida, también se han enfrentado al problema del resumen mediante el uso de sentencias cuantificadas. En [41] Díaz-Hermida y otros exponen los problemas de los que adolece la tarea de resumen lingüístico de datos y presentan una descripción de las fases más importantes en este proceso en las cuales se encuentran involucradas las sentencias cuantificadas.

Finalmente, G. Triviño y otros también han abordado la generación de resúmenes utilizando lenguaje natural mediante un modelo denominado *Modelo Lingüístico Granular de un Fenómeno*, donde se utilizan variables lingüísticas y reglas difusas para representar percepciones relacionadas con el fenómeno que se pretende describir, y la relación entre el cumplimiento de las mismas. Estas técnicas se han aplicado a una gran cantidad de problemas del mundo real: la descripción de las posturas que toman los humanos así como su actividad [156, 158], la descripción de consumo de energía [162], la descripción del tráfico rodado en una rotonda y de la evolución del tráfico en carreteras [157] y [138], y generación de resúmenes textuales en el campo del análisis financiero [115], entre otros.

Reglas

Otro tipo de protoforma de uso bastante extendido, sobre todo en minería de datos y aprendizaje automático, son las reglas de asociación y las reglas difusas. Éstas son utilizadas para descubrir asociaciones o correlaciones entre un conjunto de elementos, objetos o, como es nuestro caso, datos.

La estructura tipo de las reglas es *"Si A Entonces B"* pudiendo ser A y B un comportamiento o una agregación de comportamientos, por ejemplo, $A = \{A_1 \ o \ A_2 \ y \ A_3\}$ o $B = \{B_1 \ y \ B_2\}$. De esta forma podemos reflejar o describir situaciones como *"Si llueve mucho entonces el nivel de humedad es alto"* o *"Si hace sol y la humedad es alta entonces la temperatura es elevada"*.

En [9] Batyrshin y Wagenknecht presentan un modelo de descripción lingüística de datos basada en reglas junto con la definición de términos lingüísticos para describir dependencias en series de datos. La salida es textual y por lo tanto fácilmente interpretable por usuarios no expertos. De nuevo en [6] Batyrshin, esta vez en solitario, procede a la representación de dependencias cuantitativas mediante el uso de reglas. Como cara negativa del trabajo se destaca que en ocasiones el conjunto final de reglas a obtener podría ser, aunque interpretable, difícilmente manejable para el usuario receptor del resultado.

Otro enfoque donde se usan reglas de asociación es el presentado por Chen y otros en [21]. El modelo propuesto usa en primer lugar una ventana deslizante para crear subsecuencias continuas de la serie de datos temporal, para a continuación extraer los itemset frecuentes de dichas subsecuencias. Por último, una etapa de post-proceso debe ser llevada a cabo con el fin de eliminar patrones redundantes. Aún con esta etapa de post-proceso, y al igual que en [9], podría ocurrir que el número de reglas que se presentan como resultado final fueran inmanejables para el usuario.

Otro enfoque relacionado en cierto modo con los anteriores es el presentado en [42] por Wu y Mendel. En éste el objetivo es realizar resúmenes lingüísticos compuestos por reglas y conjuntos difusos intervalares de tipo 2. Con el fin de que el usuario sea capaz de manejar esta herramienta, se ha implementado una interfaz de usuario.

Liu y otros presentan en [96] una técnica de extracción de relaciones difusas de un modelo de series de datos a través de aproximación de conceptos. El objetivo de dicho trabajo es permitir realizar predicciones junto con la ayuda de los resultados obtenidos por el modelo. En este caso el uso de lógica difusa ayuda a obtener el resultado final, pero como se puede observar en el trabajo no hace que éste sea más fácilmente interpretable por los usuarios encargados de recibirlos.

El mayor inconveniente del que adolecen los modelos presentados deriva del elevado número de reglas de asociación que pueden llegar a extraerse de un determinado

conjunto de datos. La presentación de dicha cantidad de información al usuario no experto puede dar lugar a una comunicación poco exitosa de los resultados.

En [50] Moreno presenta una aplicación de su modelo de resumen con lógica difusa al campo de la electricidad. En [49] el autor establece unos mecanismos que le permiten la obtención de resúmenes amigables de modelos dinámicos. En este caso se usan reglas para obtener el resumen final aunque estas no sean claramente visibles después de un post-proceso que permite obtener un texto más fluido y amigable que describe las tendencias.

En [93] se hace una reflexión acerca de la proximidad conceptual entre las reglas de asociación y los resúmenes lingüísticos especialmente cuando nos encontramos en el ámbito de las bases de datos multi-dimensionales.

Otras plantillas

Otro autor que muestra su interés en los resúmenes lingüísticos es I. Kobayashi. Por un lado, y junto con N. Okamura se centra en el resumen de series de datos temporales [90, 91]. Posteriormente, junto con Mami Noumi y Atsuko Hiyama, se propone hacer lo mismo pero aplicándolo para el resumen de la conducta de una persona en una habitación [89]. En este último caso, la fuente de información son imágenes tomadas por cámaras. Hacen uso de etiquetas lingüísticas para la anotación de las imágenes.

En estos casos no podemos decir que los resúmenes sigan un patrón determinado que se repite, sino que se cuenta con una batería de expresiones que se usan en las situaciones adecuadas. Se pueden obtener dos tipos de resumen, aquel que hace referencia a la información de los datos, y aquel que además añade información acerca de la representación gráfica de éstos.

Un ejemplo del primer tipo es: *El precio de cierre de la media de stock del Nikkei en el mercado de stock de Tokio en el 15 de Agosto de 2005 se recuperó. Subió 160.78 yenes, o el 1.38 %, hasta alcanzar los 12452.51 yenes y se recuperó el nivel de 12400 yenes. Éste ascendió hasta un alto nivel. El precio del mercado subió. El precio de stock presentaba valores elevados a lo largo del día.*[2].

Mientras que un ejemplo del segundo tipo sería: *Durante la sesión de mañana, las órdenes de venta avanzaron. Después, el comercio permaneció firme, a continuación, el volumen de incremento fue pequeño. Durante la sesión de tarde, el comercio fue*

[2]La traducción se ha hecho a un nivel muy literal para conservar las estructuras utilizadas originalmente: *The closing price of Nikkei stock average at Tokyo stock market on August 15, 2005 rebounded. It added 160.78 yen, or 1.38 percent, to reach 12452.51 yen and has recovered at the level of 12400 yen. It rose to high level. The market price rose. The stock price was moving at a high level troughout the day.*

creciendo continuamente. A continuación, el volumen de incremento se expandió. Durante el cierre de sesión, los precios se decrementaron[3].

Como se puede observar en el primer tipo se describe el comportamiento de los datos, en tanto que en el segundo tipo se incluye información temporal y de tendencias, que en este caso viene extraída de la gráfica. En nuestra opinión, un resumen más idóneo sería una mezcla de los dos anteriores. Desde nuestro punto de vista, la información mostrada en el segundo tipo de resumen se puede extraer de los datos en sí junto con algo de información contextual sin la necesidad de implementar reconocimiento de patrones y formas en la representación gráfica.

Ehud Reiter y Robert Dale han expresado su inquietud por construir sistemas que generaran resultados en lenguaje natural [130,131]. En su trabajo dentro del contexto de la Generación de Lenguaje Natural (NLG, Natural Language Generation) [132] Reiter y otros se centran en la labor de elegir "palabras" para componer una predicción meteorológica. El producto final de este trabajo es el sitema SUMTIME-MOUSAM. Este sistema está conectado a un corpus de predicciones meteorológicas reales con el fin de adecuar el lenguaje natural generado por la máquina al que podría producir un experto en la materia. En [177] Reiter y otros cambian de contexto y nos explican cómo seleccionar el contenido en resúmenes textuales de series de datos de gran tamaño introduciendo el uso de ontologías de términos lingüísticos.

En [99] Mahamood y Reiter introducen el proyecto Baby Talk [51] (Albert Gatt, François Portet, Ehud Reiter, Jim Hunter, Saad Mahamood, Wendy Moncur y Somayajulu Sripada) cuyo objetivo es desarrollar software que permita generar resúmenes de los datos médicos en el entorno de la Unidad de Cuidados Intensivos para bebés. Estos resúmenes están pensados para que puedan facilitar la tarea de traspaso de conocimiento y toma de decisiones por el personal sanitario. En este caso, se presenta un nuevo modelo, que se podría considerar parte de la familia de modelos Baby Talk, llamado BT-Family, pensado sobre todo para generar informes para los padres de los bebés ingresados y no para el personal sanitario. En este caso el lenguaje usado no es tan técnico lo que facilita que los padres puedan comprender el informe presentado.

En [135] Rieser y Lemon presentan y evalúan su modelo de generación de lenguaje natural para sistemas de diálogo oral. Este modelo esta basado en planificación estadística. El sistema sigue tres estrategias, el resumen, la comparación y la recomendación, y podemos ver cómo las aplican al corpus MATCH que contiene diversa información sobre restaurantes.

En [66] Umano y otros se enfrentan a la descripción de series a través de las ten-

[3]La traducción se ha hecho a un nivel muy literal para conservar las estructuras utilizadas originalmente:*At the morning session, sell order was ahead. Afterwards, trading was steady, therefore, the width of rising was small. At the afternonn session, trading was continuosly rising. Therefore, the width of rising was expanded. At the closing session, the prices were decline.*

dencias globales y características locales de las mismas. Para llevar a cabo el proceso los autores hacen uso de intervalos difusos, reglas y el corpus de términos MuST (Multimodal Summarization for Trends Information).

Otras herramientas para la construcción de mensajes

Conjuntos difusos de tipo 2

El concepto de conjunto difuso de tipo 2 fue introducido por Zadeh [179] como generalización de los conjuntos difusos (de tipo 1). Éstos nos permiten incorporar la incertidumbre en la definición de la función de pertenencia.

Dado un conjunto X, un conjunto difuso de tipo 2 A se define de la siguiente forma:

$$A = \{(a, x, \mu_a(x)|a \in A, x \in [0, 1])\}$$

donde la función $\mu_a(x)$ representa el grado en el que el valor x es el valor de pertenencia de a a A.

Recordemos que tanto en [112,113] de Niewiadomski, como en [42] de Wu y Mendel se presenta la utilidad del uso de conjuntos de tipo 2 al realizar resúmenes lingüísticos en lugar de los extendidos conjuntos de tipo 1.

En ambos trabajos se introduce el uso de este tipo de conjuntos como una herramienta más que nos permite imprimir una mayor flexibilidad al afrontar el proceso de conversión de datos numéricos a una salida o resumen en lenguaje natural.

Cuantificadores y cardinalidades difusas

P. Bosc presenta un modelo de resumen difuso con componentes lingüísticos pero con un producto final que necesita cierta interpretación. En el trabajo [11] se realiza resumen difuso de datos usando cardinalidades difusas. Junto con Allel HadjAli, Hélène Jaudoin y Olivier Pivert presentan sus técnicas para consulta flexible para bases de datos en contextos distribuídos [12].

En [39] y [40] se realiza el proceso de resumen usando cuantificadores difusos y semi-difusos respectivamente. En [126] se presenta un ejemplo práctico de aplicación de este tipo de técnicas para la obtención de informes climatológicos de temperatura usando como punto de partida series de datos meteorológicos.

Uso de OWA

Además de Yager en [173] y [174] también otros científicos han utilizado los operadores OWA para el resumen. En la ya mencionada herramienta *Quantirius* [123]

además de la implementación de validación mediante el método propuesto por Zadeh, se ha ampliado la funcionalidad y se ha introducido la posibilidad de trabajar con los mencionados operadores OWA.

Por su parte, los investigadores Pei y otros presentan en [119] un método para extraer resúmenes complejos usando agregaciones lingüísticas en el entorno de una base de datos, en concreto una base de datos de personal. Como herramienta para el resumen se usa el operador LOWA, extensión de los OWA de Yager, mencionados anteriormente. Una vez obtenido el resumen, se realiza una optimización del número de términos lingüísticos así como de sus grados de pertenencia a través del uso de algoritmos genéticos. De este modo el resultado final será un resumen con una grado de verdad más elevado después de este post-proceso mediante técnicas evolutivas. Como en trabajos anteriores se echa de menos poder ver un resultado compuesto por sentencias en lenguaje natural que ejemplifique el modelo presentado.

En [16] de Carrasco y Villar también hacen uso del operador OWA, así como de su extensión LOWA, para obtener resúmenes automatizados de datos heterogéneos almacenados de forma digital en una base de datos con opiniones de usuarios de hoteles y servicios.

2.4.4. Uso de jerarquías y ontologías

El uso de jerarquías es algo que ha sido explorado por diversos investigadores como herramienta que dota de mayor versatilidad al resumen. La gran ventaja que se nos presenta al usar jerarquías es la capacidad de obtener resúmenes con diferentes niveles de granularidad o abstracción en la descripción de conceptos. Del mismo modo, el uso de ontologías nos permite obtener diferentes niveles de descripción en los términos lingüísticos usados en el resumen.

La investigadora Anne Laurent presenta su conjunto de técnicas en [93]. En su trabajo se centra en la obtención de resúmenes lingüísticos a partir de bases de datos multi-dimensionales difusas y mediante la aplicación de operaciones OLAP sobre ellas. El uso de jerarquías toma un papel muy importante en [93] ya que en este caso encontramos una jerarquía en la dimensión temporal que viene heredada de la ya existente en el cubo de datos de donde se extrae la serie de datos. De esta manera se obtiene una colección de resúmenes con diferente nivel de detalle en la descripción de la dimensión temporal.

En [127] los investigadores G. Raschia y N. Mouaddib presentan su modelo *SaintEtiq* para resumir bases de datos. En este modelo se trabaja con conjuntos de conceptos lingüísticos organizados jerárquicamente. La consulta de la base de datos dará lugar a una jerarquía de resúmenes con diferentes niveles de abstracción.

Posteriormente, R. Hayek y otros aplican *SaintEtiq* para crear *PeerSum*, que ofrece

resúmenes de servicio para aplicaciones P2P [58]. Siguiendo en el ámbito de las redes P2P, los mismso autores presentan [59] donde se ofrecen resúmenes de actuación en redes P2P no estructuradas. Aunque, como veremos más adelanta en el trabajo, en nuestra propuesta las jerarquías de conceptos jugarán un papel muy importante, el uso que se hace de ellas no es el mismo. En nuestro caso, en un resumen final podrán aparecer conceptos de diferentes niveles en la jerarquía, en lugar de tener una jerarquía de diferentes resúmenes que no se pueden combinar entre sí. Del mismo modo, podemos afirmar que aunque el modelo tiene un alto componente lingüístico, los resúmenes finales no se presentan completamente en lenguaje natural, por lo que es necesario una etapa de interpretación que no aparece en nuestro modelo.

En [160] L. Ughetto y otros, se vuelve a trabajar con *SaintEtiq* pero esta vez haciendo más hincapié en las grandes posibilidades que ofrece el que los usuarios, y por tanto receptores de los resúmenes, sean parte activa en la definición del vocabulario que se va a usar.

El uso de jerarquías de conceptos también ha sido explotado por los investigadores Petry y Zhao en [121] como instrumento para ganar poder descriptivo de nuestros datos al pasar la representación de los mismos al lenguaje natural.

En otro sentido, pero muy relacionado con los anteriores, Lee y Kim [94], obtienen una jerarquía de conceptos pero basándose en el uso de relaciones difusas ISA (o ES UN si usamos el término castellano).

Yager y Petry hacen uso de las ontologías de conceptos en el proceso de resumen lingüístico en su trabajo [176]. Del mismo modo, en [177] Reiter y otros se centran en cómo seleccionar el contenido en resúmenes textuales de series de datos de gran tamaño introduciendo el uso de ontologías de términos lingüísticos.

2.4.5. Medidas de calidad

No existe un consenso acerca de cuáles son los aspectos en los que nos debemos fijar para poder conocer la calidad de un resumen en términos cuantitativos. Diversos autores han propuesto diferentes medidas, que sin ser iguales, en algunos casos presentan cierto grado de similitud.

La medida más básica para obtener la calidad de un resumen compuesto de sentencias cuantificadas es conocer la validez o verdad de dichas sentencias. A partir de ahí, Yager [167] introduce el término de *informativeness*, es decir, el grado en el que es informativa la sentencia. De este modo podemos modelar el hecho de que una sentencia con grado de cumplimiento o validez igual a cero, nos pueda resultar muy informativa. La información que obtenemos se considera relevante ya que nos permite conocer un comportamiento que no está presente en el conjunto de datos.

En [176] Yager y Petry introducen un conjunto de seis medidas de calidad relacionadas con conceptos tales como la cobertura del resumen, la relevancia del mismo, la concisión o la utilidad.

Sin embargo, en [111] Niewiadomski presenta un conjunto de seis medidas diferentes de medida de la calidad. Podemos encontrar entre ellas ideas como la imprecisión o la cardinalidad de la cuantificación o la longitud entre otros.

Más tarde, el mismo Niewiadomski en [112] propone una ampliación de estas medidas para el caso específico del uso de conjuntos difusos de tipo 2. El conjunto se compone de un total de diez medidas que se agregan en una onceava usando pesos para ponderarlas entre ellas (*degree of truth, degree of imprecision, degree of covering, degree of appropriatenes, length, type-2 quantification imprecision, type-2 quantification cardinality, type-2 summarizer cardinality, imprecision of the type-2 query, cardinality of the type-2 query*).

Algunas de las medidas anteriores más otras diferentes se pueden encontrar en los diversos trabajos de Kacprzyk y otros (*a truth value, degree of imprecision (fuzziness), degree of covering, degree of appropriateness y lenght of a summary*). Reiter y otros también han expresado su interés y preocupación a la hora de realizar una buena evaluación de los resúmenes obtenidas con sus métodos [128, 129, 137].

Además de en la generación de los resúmenes, Triviño y otros han trabajado en el desarrollo de un procedimiento que les permita medir la calidad de dichas descripciones de datos [120].

Debido a la importancia que ostenta la capacidad de determinar la calidad de un resumen, un estudio más detallado se presenta en la Sección 3.5. Además, como fruto de nuestra investigación, presentamos también un modelo multi-dimensional para la medida de calidad de los resúmenes generados. Del mismo modo, se introduce una instanciación ejemplo de dicho modelo.

2.4.6. Post-proceso del resultado

El post-proceso también es importante para algunos autores que buscan mejorar la legibilidad del resultado eliminado repeticiones o agregando resultados entre otras tareas. Gracias a un buen post-proceso podemos hacer que el resultado sea menos extenso y por lo tanto más manejable, y que presente un aspecto lo más parecido posible a un resumen lingüístico producido por una persona.

En *Quantirius* de Pilarski se ha implementado una fase final en la que se aborda la reducción del resultado. Al presentar al usuario final un resumen sin repeticiones, además de hacerlo más corto, le facilitaremos la comprensión del mismo.

Con la misma filosofía Chen et al [21] implementan un mecanismo que reduce el

número de reglas producidas a un conjunto menor, y por lo tanto, más manejable por el usuario.

Por su parte en [119] dicho proceso de reducción y adaptación del resultado se lleva a cabo mediante el uso de algoritmos genéticos que optimizan el número de términos lingüísticos y sus grados de pertenencia.

2.4.7. Interfaces de usuario

En esta sección se presentarán aquellos modelos que incorporan una interfaz de usuario.

En este apartado debemos mencionar que Kacprzyk y otros han realizado una implementación de su método basado en consultas en la herramienta *FQUERY*, un paquete para Access de Microsoft. De este modo se capacita al usuario de Access para realizar consultas que den lugar a resúmenes lingüísticos además de los clásicos informes convencionales.

Quantirius de Pilarski [123] también incorpora un entorno, en este caso no basado en Access. Como se ha comentado anteriormente en [42] junto con el método de extracción de reglas para el resumen se presenta una interfaz gráfica. El elevado número de reglas de asociación que pueden llegar a extraerse unido a la representación gráfica de las mismas (que no parece muy intuitiva) hacen que los resultados, aunque correctos para usuarios expertos, sean poco amigables para usuarios no expertos

En el trabajo [164] se le dota de una interfaz de usuario a la herramienta anteriormente nombrada *SaintEtiq*. De esta manera se pretende que su capacidad pueda ser explotada al máximo por los usuarios de la manera lo más simple y transparente posible.

En el trabajo de A. Laurent [93], los métodos propuestos han sido implementados en el sistema Oracle Express Server usando Oracle Express Objects, Java y C++; lamentablemente el artículo no ofrece ninguna captura de pantalla que nos dé una idea de la usabilidad de los mismos por parte del usuario final.

En [24] Chiang y otros también expresan su interés en el resumen lingüístico de series de datos temporales. En este caso presentan un modelo que pretende realizar minería de datos en el contexto de series de datos por medio de un sistema de resumen lingüístico. En este caso, lo que ellos presentan como resumen lingüístico no deja de ser un resumen numérico en el que se han utilizado para su obtención etiquetas lingüísticas. Aunque se ha implementado una interfaz para usuario, en ella podemos ver que los resultados son numéricos y por tanto difíciles de interpretar.

En [90] de Kobayashi y otros se ha implementado una interfaz gráfica que presenta al usuario tres áreas bien diferenciadas. Por un lado se muestra el conjunto de

datos tanto en forma tabular como gráficamente, y por otra parte se muestra el texto generado para resumir dichos datos.

2.4.8. Objetivo de la propuesta

En muchas de las propuestas analizadas el modelo se aplica para la situación específica para la que fue concebido. En los casos anteriores, para la descripción de las series de datos.

Kacprzyk y otros, además de por el simple resumen lingüístico de series de datos temporales, se han interesado también por la realización de resúmenes de la comparación de dichas series, véase [73, 74].

Como veremos en el Capítulo 5 el modelo en el que hemos trabajado, además de para la descripción y resumen de series, también se puede aplicar a la comparación de series desde diversos puntos de vista. Además se presenta una propuesta en la que el modelo se utiliza para la descripción lingüística de imágenes almacenadas digitalmente.

Algunas de las propuestas, además de centrarse en el proceso mismo de resumen, prestan especial atención al proceso seguido para la segmentación en la línea temporal. La estructura de periodos obtenidos en dicha dimensión juega un papel crucial durante el proceso de resumen. En algunas ocasiones dicha estructura deberá ser aportada por el usuario, como es nuestro caso, mientras que en otras se obtiene de forma automática o semiautomática.

2.4.9. Discusión

La Tabla 2.3 recoge las principales características de los modelos comentados en esta sección. Las columnas están separadas en tres grandes bloques. Los dos primeros corresponden a herramientas utilizadas para la construcción de los mensajes. El último bloque es más interesante ya que refleja herramientas para construir el contexto, medir la calidad y utilidades para facilitar al usuario la usabilidad de los modelos.

Como podemos observar estamos ante una tabla muy dispersa. Esta situación no es preocupante en los dos primeros bloques, pero merece un poco más de atención cuando se refiere al último de los bloques.

En éste se refleja si se han usado jerarquías u ontologías a la hora de de flexibilizar el contexto lingüístico (columna 7), si se han presentado medidas de calidad con las que evaluar los resultados obtenidos (columna 8) o si se han implementado medidas de post-proceso o interfaces de usuario para facilitar al receptor la comprensión del resultado (columnas 9 y 10 respectivamente).

	Obtención de mensajes						Jerarquías ontologías	Medidas de calidad	Post-pro. de resultado	Interfaz usuario	Objetivo propuestas
	Protoformas			Otras herramientas							
	Sent. cuantif.	Reglas asoc.	Otras	CD de tipo 2	Cuantif. dif.	OWA					
Yager [167, 173]	X					X		X			
Kacprzyk y otros [80, 81]	X							X		X	X
Pilarski [123]	X					X			X		
Niewiadomski [112, 113]	X			X				X		X	
Keller [2]	X										
Zhang [185]	X										
Carrasco y Villar [16]	X					X					
Bugarín y otros [39, 126]	X				X						
Batyrshin [9]		X									
Chen y otros [21]		X							X		
Wu y Mendel [42]		X								X	
Liu y otros [96]		X		X							
Moreno [49]		X									
Triviño y otros [120, 157]	X		X		X			X			
Laurent [93]		X					X			X	
Kobayashi y otros [90, 91]			X							X	
Reiter y otros [128, 129, 137]			X					X			
Rieser y Lemon [135]			X								
Umano [66]			X								
Bosc y otros [11]					X						
Pei y otros [119]						X			X		
Raschia y otros [127, 160]							X			X	
Petry y Zhao [121]							X				
Lee y Kim [94]							X				
Yager y Petry [176]							X	X			
Chiang y otros [24]										X	

Tabla 2.3: Comparativa de modelos de resumen.

Por último, se ha considerado también si el modelo se ha aplicado a campos diferentes (con características diferentes) de para el que fue diseñado (columna 11).

Como decíamos, la tabla en este bloque está muy dispersa, de modo que podemos decir que los modelos propuestos no cubren todos los aspectos que por unanimidad se han considerado importantes al enfrentarse al desarrollo del modelo. En los casos analizados los autores se han centrado en aspectos específicos sin llegar a dar una visión de conjunto y sin dar este conjunto de características que permitirán al usuario el fácil manejo del modelo así como la comprensión del mismo.

En este sentido, nosotros presentaremos un modelo para el resumen lingüístico de datos que haga uso de jerarquías durante la definición del contexto, más concretamente en la dimensión temporal. Asimismo se presentarán una serie de medidas de calidad con las que medir la bondad de los resúmenes generados y nos darán oportunidad de compararlos entre ellos. Además realizamos una etapa de post-proceso que ayudará a transformar el resultado tal y como se obtiene a la salida del proceso a un párrafo lo más parecido posible a cómo lo generaría un resumidor humano y también se ha implementado una interfaz de usuario que facilite la interacción del humano con el modelo. Por último, y debido a la generalidad del modelo, éste es portable a otros tipos de datos, como veremos.

2.5. Conclusiones

A largo de este capítulo se han presentado los conceptos más importantes y necesarios para afrontar sin grandes problemas el resto de capítulos de la memoria.

Comenzando por la idea básica de resumen, pero sobre todo de resumen lingüístico, y continuando por la aplicación del mismo a series de datos. Aunque el modelo desarrollado puede ser aplicado a distintos tipos de series, nos hemos centrado en las series temporales debido al papel que estas juegan en nuestra vida cotidiana.

Otro aspecto fundamental es la introducción de la Soft Computing en la tarea y la presentación de los conjuntos difusos y las variables lingüísticas que nos serán de gran utilidad al presentar nuestro modelo, jugando éstas un papel decisivo en la tarea de conversión de datos a resumen textual. Asimismo, se han presentado distintas estrategias que nos ayudan a buscar de forma automática el mejor resumen, adelantando las técnicas que se usarán en nuestro modelo. Por un lado, un algoritmo Greedy y por otro un algoritmo evolutivo multi-objetivo.

Por último se ha realizado un estudio de las técnicas que aplican Soft Computing a problemas de resumen lingüístico. En este repaso se han presentado las obras más representativas que además nos ayudarán luego a comprender mejor nuestro planteamiento en el siguiente capítulo. Los distintos trabajos han sido analizados desde diversas perspectivas o puntos de interés. Por último se ha confeccionado una tabla

comparativa, en la que hemos visto que existe gran dispersión. En capítulos posteriores presentaremos el modelo que se ha desarrollado y veremos cómo incorpora todas las características fundamentales.

3

Un modelo para el resumen lingüístico de series de datos

"El más pequeño gato es una obra maestra"
Leonardo da Vinci

Este capítulo tiene como objetivo introducir un nuevo modelo para el resumen lingüístico de series de datos. En particular, y por motivos que han sido expuestos con anterioridad, nuestro interés va a estar especialmente centrado en las series de datos temporales, aunque los resultados, como veremos, son aplicables mas allá de éstas.

El mencionado modelo pretende establecer una base formal bien definida con la que la computadora pueda trabajar pero manteniendo la cercanía con el usuario final, tanto en su forma de manejo como en la claridad de los resultados que ofrezca.

El texto en lenguaje natural que compone los resúmenes que se mostrarán al usuario final no es un texto desestructurado, sino que sigue unos patrones de construcción concretos. Dicho texto, como conjunto, se encuentra compuesto por una serie de sentencias con una forma prefijada. Existen diversas construcciones tipo que se consideran factibles a la hora de construir un resumen. En nuestro caso, para definir esta disposición en los elementos que conformen el resumen, nos hemos decantado por el uso de sentencias cuantificadas. Este tipo de sentencias posee un esqueleto lo bastante cerrado como para facilitar el proceso de creación de lenguaje natural por parte de la computadora, pero que, al mismo tiempo, se considera muy cercano a aquellos que se usan diariamente por los seres humanos en sus procesos de comunicación, y concretamente, en los procesos de resumen.

El proceso de resumen lingüístico de datos se considera como un caso particular de la disciplina denominada Generación de Lenguaje Natural (del inglés, Natural Language Generation, NLG). En todo proceso de creación de un texto en lenguaje natural por parte de una computadora existen una serie de pasos o fases que se deben seguir. Dichas fases, así como su relación con la creación de resúmenes en nuestro modelo, serán presentadas a lo largo de las siguientes secciones.

Pero no sólo es importante el proceso de creación del resumen, sino que también lo es contar con un conjunto de medidas que nos permitan evaluar la calidad del

texto generado. Por este motivo, acompañamos la formalización de nuestro modelo con una propuesta que permite evaluar la calidad de los resúmenes generados. Para ello, mostraremos distintos enfoques existentes a la hora de evaluar la calidad de un resumen final, para, a continuación, presentar un modelo de evaluación de la calidad de un resumen desde un punto de vista general que luego particularizaremos para poderlo usar en nuestro enfoque de resumen lingüístico de series de datos.

3.1. Resumen lingüístico en el ámbito de la Generación de Lenguaje Natural

Como puede parecer intuitivo, existe una relación entre la creación de resúmenes lingüísticos de datos de forma automatizada por medio de computadores y la disciplina conocida como Generación de Lenguaje Natural, y es que la primera es un caso particular de la segunda. En este sentido, en ambas disciplinas se cuenta con un conjunto de datos almacenados digitalmente a la entrada y se ofrece un texto en lenguaje natural como salida. En concreto, el texto de salida del proceso de resumen deberá cumplir las características que lo acrediten como tal y que abordaremos con detalle más adelante en este capítulo cuando nos centremos en la calidad del resumen.

Una vez hecha esta aclaración, parece lógico considerar que el proceso discurrirá de forma idéntica en ambos casos. Existe un consenso general acerca de cuáles deberían ser las fases que se sigan cuando se pretende generar lenguaje natural automáticamente. Dichas fases fueron recogidas por Reiter y Dale en su trabajo "Building Natural Language Generation Systems" [131] (una versión más resumida y concisa del libro se puede encontrar en el artículo [130]). A continuación, pasaremos a repasarlas sucintamente con objeto de poder relacionarlas con las etapas seguidas en la creación de nuestro modelo.

Las tareas que se deben tener en cuenta para crear un buen texto en lenguaje natural son:

1. Definición del contenido (Content determination).
2. Estructura del documento (Document structuring o Discourse planning).
3. Agregación de componentes (Aggregation).
4. Elección del léxico (Lexical choice).
5. Generación de expresiones referenciales (Referring expression generation).
6. Realización lingüística (Linguistic realisation).

Que se encuentre un consenso sobre la existencia de dichas tareas, no implica que siempre deban aparecer todas o que deban aparecer en ese orden. Algunas de ellas están muy relacionadas entre sí, pudiendo incluso llegar a fusionarse. Otras pueden dejar de ser necesarias en determinados contextos. Veámoslo con más detalle.

Durante la *definición del contenido* del texto final se llevarán a cabo las decisiones pertinentes acerca de lo que será mencionado en el texto final y lo que no. Es decir, la información que se considera relevante y la que se puede omitir. Esta etapa es

muy dependiente del uso que le vayamos a dar al texto final, por lo que es una fase altamente sensible al contexto. Al definir la *estructura del documento*, lo que se hace es llegar a un acuerdo sobre la forma de presentación de la información obtenida. Estas dos fases se encuentran muy relacionadas, formando parte de lo que se conoce como la planificación del texto final (del inglés, text planning).

La *agregación de componentes* es una fase muy importante a la hora de dotar de legibilidad al texto final. En esta fase se fusionan entre sí sentencias simples, con el fin de ganar naturalidad en el resultado. Otra fase crucial es aquella en la que se realiza la *elección del léxico* para representar los conceptos con las palabras adecuadas. Para ello nos podemos ayudar de diccionarios, ontologías y conocimiento experto, entre otras herramientas. Estas dos fases se encuentran estrechamente relacionadas, siendo práctica habitual unirlas en una sola fase. Durante la *generación de expresiones referenciales* se decidirá qué términos serán los adecuados para denominar objetos que aparecen en el resumen. Junto con las anteriores, compone la denominada planificación de sentencias (del inglés, sentence planning).

Por último ya sólo nos quedaría construir el texto propiamente dicho, en la etapa de *realización*. Esta etapa cosiste en la aplicación de las reglas gramaticales para la construcción de un texto que sea sintáctica, morfológica y ortográficamente correcto.

En el caso de nuestro modelo la tarea 1, es decir, la definición del contenido del resumen, se abordará durante la definición del contexto y el diseño y ejecución de las técnicas algorítmicas. Al definir el contexto estamos modelando el entorno del problema. En la fase de diseño se toman decisiones que influyen de manera fundamental acerca de lo que finalmente aparecerá en el resumen al ejecutar el o los algoritmos implementados para automatizar su generación. Al finalizar esta fase sabremos qué mensajes y conceptos son los que componen el resumen final. La fase de diseño de las técnicas algorítmicas que solucionen el modelo debe ser afrontada concienzudamente de forma que el resumen final que obtengamos cumpla nuestras expectativas.

La tarea 2, o estructura del documento, queda definida de forma completa al seleccionar el tipo de estructuras que compondrán el documento y de qué forma lo harán. En nuestro caso, el resumen es una colección de sentencias cuantificadas. Una vez superada la tarea 2 se está en disposición de acometer la tarea 3, agregación de componentes. En nuestro caso, a partir del conjunto final de sentencias cuantificadas se obtendrá un párrafo completo en lenguaje natural. Para ello se eliminarán las repeticiones de elementos fusionando sentencias, de este modo el resultado final gana en legibilidad y es más cercano a la manera de expresarse usada por los seres humanos.

Las tareas de la 4 a la 6 están altamente relacionadas en nuestro modelo, y muy influenciadas por la información del contexto. Nuestro modelo fusiona las tareas 4 y 5, elección del léxico y generación de expresiones referenciales, cuando se asignan términos lingüísticos a los conceptos modelados durante la tarea 1. Por ejemplo si en

la tarea 1 hemos definido un conjunto difuso representado por el trapecio $[3, 4, 5, 6]$, durante esta tarea se debe asociar un término lingüístico a dicho conjunto, por ejemplo, *"aproximadamente entre 4 y 5"*. Con respecto a la tarea 6, si la estructura de la sentencia cuantificada es correcta (sintaxis y morfología) y los términos que la instancian también lo son (ortografía), la realización lingüística no requiere ninguna acción adicional para poder asegurarla.

En los sucesivos apartados posteriores se tratarán con más detalle los distintos pasos que se han seguido y las acciones que han sido tomadas para abordar cada una de las fases anteriores en nuestro modelo.

3.2. Mensajes del resumen

El texto en lenguaje natural que se presenta al usuario como resultado del proceso de resumen lingüístico puede ser considerado como un conjunto de mensajes. Cada uno de estos mensajes individuales se representa por medio de una sentencia, pero no por una cualquiera, sino por una con una estructura formal bien definida.

En numerosos trabajos (véase por ejemplo [81, 82, 182]) se usa el concepto de *protoforma* como esqueleto tipo o plantilla en la que se basa la construcción del mensaje final. El término protoforma es una abreviación de "forma prototípica"[1] y fue introducido como tal en el ámbito de la computación por Zadeh en [184]. Este término también es usado en otros campos para denominar la primera forma[2]. Por ejemplo, en lingüística se usa para las primeras construcciones que usan los bebés al comenzar a hablar, o las primeras construcciones verbalizadas de las antiguas culturas.

Existen diferentes estructuras susceptibles de ser usadas como protoformas. A este respecto, consideramos que las protoformas más extendidas al realizar resumen lingüístico son las reglas de asociación [9, 21, 42, 96] y las sentencias cuantificadas [41, 77, 123, 157, 167]. Incluso podemos llegar a considerar que existe una estrecha relación entre estas dos construcciones [79].

A pesar de que las reglas de asociación son bastante intuitivas y cercanas al lenguaje natural, en nuestro modelo nos hemos decantado por el uso de sentencias cuantificadas. Estas sentencias se encuentran bastante extendidas y han sido ampliamente usadas en la literatura. Además, en cierto modo, se puede considerar que la cuantificación y el resumen son tareas altamente relacionadas entre sí, de modo que nos parece muy adecuado su uso para solucionar el problema que se aborda aquí.

De alguna manera lo que hacemos en esta disciplina es agrupar una serie de hechos en su totalidad y resumirlos en una sola expresión. Aunque el punto de partida se

[1] del inglés, protoform = "prototypical form"

[2] del latín, proto- forma combinatoria de prōtos-, primero, se usa para denotar la condición de primero en algún orden, especialmente temporal, o la condición de incipiente y primitivo.

tiene en la cuantificación clásica en la que los cuantificadores son sólo dos, el cuantificador existencial $\exists$ y el universal $\forall$, para flexibilizar este proceso y adecuarlo al modo en el que los humanos cuantificamos, surgen en la literatura familias enteras de cuantificadores que nos permiten describir la realidad que nos rodea de una manera más adecuada.

"Hay alumnos que han aprobado" y *"Todos los alumnos han aprobado"* son ejemplos de sentencias cuantificadas clásicas. Pero, si no son todos o ninguno, ¿cómo expresamos cuántos de los alumnos han aprobado?. Con una familia de cuantificadores más amplia podremos describir la situación como *"La mitad de los alumnos han aprobado"*, *"Aproximadamente el 80 % de los alumnos han aprobado"* o *"Sólo dos de los alumnos han aprobado"*, entre muchas otras sentencias. Como podemos observar este tipo de expresiones son más cercanas a las que usaría un ser humano para describir la situación.

Dos corrientes se han seguido a la hora de llevar a cabo esta tarea. La primera de ellas fue presentada en [180] y da como resultado una gran cantidad de cuantificadores difusos que pueden ser clasificados en dos subfamilias: cuantificadores absolutos (alrededor de 3, aproximadamente 5) y relativos (al menos la mitad, aproximadamente el 60 %). Esta corriente se encuentra muy extendida y es muy usada en procesos de cuantificación.

La segunda corriente se basa en la aplicación de la llamada Teoría de Cuantificadores Generalizados (del inglés, Theory of Generalized Quantifiers), en la que se reconocen más de 30 tipos de cuantificadores. En este caso, los cuantificadores son también llamados determinantes (del inglés, determiners). Dicha corriente también ha sido usada ampliamente en este campo y algunos ejemplos de uso se muestran en los trabajos [39, 41, 52]. Más sobre el modelo de de cuantificadores generalizados y su relación con el lenguaje natural puede ser encontrado en [5].

En nuestro modelo hacemos uso de los cuantificadores difusos de Zadeh inmersos en sentencias cuantificadas del tipo,

$$\text{"}Q \text{ de los } D \text{ son } A\text{"}$$

donde Q es un cuantificador lingüístico, y A y D son propiedades difusas definidas sobre los elementos de un conjunto X. Este tipo de sentencias recibe el nombre de sentencias cuantificadas de tipo II. Como caso particular, si D es crisp las sentencias pasan a denominarse de tipo I.

Algunos ejemplos de este tipo de sentencias son:

"La mayoría de motocicletas pesadas poseen gran cilindrada"

"Aproximadamente el 60 % de los días de Abril las precipitaciones fueron abundantes"

De este modo tenemos un conjunto sujeto a una propiedad o restricción ("las motocicletas pesadas" o "los días de Abril") sobre el que se informa en relación con una determinada propiedad ("poseer gran cilindrada" o "tener precipitaciones abundantes") mediante el uso de un cuantificador ("La mayoría" o"Aproximadamente el 60 %"). Podemos ver que este tipo de construcciones no son ajenas a la forma de hablar que utilizan los seres humanos, resultando bastante amigable su uso para crear el resumen final.

En la siguiente sección definiremos con mayor detalle los elementos que compondrán nuestra sentencia, esto es, qué son en nuestro caso los componentes Q, D y A.

3.3. Marco lingüístico del resumen

La presente sección se encuentra dedicada a la exposición y definición del marco lingüístico que se propone en la memoria como parte de nuestro modelo para el resumen de series de datos temporales. Como veremos, nuestro marco lingüístico se compone de una partición difusa en el dominio de la variable bajo estudio y una jerarquía de particiones difusas en el dominio del tiempo. Estas particiones, junto con los cuantificadores lingüísticos utilizados, serán las encargadas de permitir la "traducción" de datos numéricos temporales a texto. A continuación, veamos en profundidad cómo se define en nuestro trabajo una serie temporal y las diversas particiones que se necesitan para verbalizarla.

3.3.1. Términos lingüísticos para el dominio de la variable y del tiempo

En primer lugar, presentamos una definición de serie de datos temporales, que nos servirá de referencia a la hora de formalizar nuestro marco lingüístico para resumirla.

Definición 3.1 (Serie de datos temporales) *Sea T la dimensión temporal y D_T su dominio descrito en su nivel de granularidad más fino como $D_T = \{t_1, ..., t_m\}$ donde cada t_i son los instantes de tiempo. Sea V una variable bajo estudio y D_V su dominio básico. Una serie de datos temporales $TS_{D_T}^{D_V}$ sobre V definida en T se representa como:*

$$TS_{D_T}^{D_V} = \{< t_1, v_1 >, ..., < t_m, v_m >\}$$

donde cada v_i es un valor en el dominio básico D_V.

En los capítulos siguientes, cuando no haya lugar a confusión, nos referiremos a la serie $TS_{D_T}^{D_V}$ como simplemente, TS.

Es decir, una serie de datos temporales sobre una variable está compuesta por un conjunto de duplas <*instante de tiempo, valor de la variable en dicho instante*>.

Con la idea de poder describir los datos de la serie mediante términos lingüísticos, tal y como hemos indicado anteriormente en esta memoria, contaremos con la ayuda de la teoría de conjuntos difusos y la definición de variables lingüísticas.

Ya se ha mencionado anteriormente que los conjuntos difusos y, particularmente, las etiquetas lingüísticas, nos brindan una inestimable ayuda a la hora de expresar información de una forma semejante a la que usamos los seres humanos. Con el fin de poder usar la lógica difusa para describir la información almacenada en forma de series de datos debemos de tener en cuenta dos puntos fundamentales.

En primer lugar, será necesario que el dominio básico de la variable V se encuentre *p*articionado, usando para ello un conjunto de etiquetas lingüísticas. Antes de formalizarlo, vamos a fijar lo que vamos a entender por partición en esta memoria.

Definición 3.2 (Partición difusa) *Sea X un conjunto de referencia y sea $\{X_1, ..., X_n\}$ una serie de conjuntos difusos definidos sobre X; diremos que $\{X_1, ..., X_n\}$ es una partición difusa de X si*

1. $\forall x \in X, \exists X_i, i \in \{1..n\} | \mu_{X_i}(x) > 0$, *donde $\mu_{X_i}(x)$ es el grado de pertenencia del elemento x al conjunto X.*
2. $\forall i, j \in \{1..n\}$, *con* $i \neq j, core(X_i) \cap core(X_j) = \emptyset$, *donde $core(C)$ es el núcleo del conjunto difuso C.*

Como vemos, en esta memoria, una serie de conjuntos difusos constituye una partición del dominio de referencia si todos los puntos del dominio pertenecen a alguno de los conjuntos difusos de la partición con grado mayor que cero y, al mismo tiempo, los núcleos de los distintos conjuntos de la partición no se solapan.

Puesto que nuestro objetivo es describir mediante el uso de lenguaje, nos interesan las particiones difusas de carácter lingüístico.

Definición 3.3 (Partición difusa lingüística) *Sea X un conjunto de referencia, sea $\{Etiqueta_1, ..., Etiqueta_n\}$ un conjunto de etiquetas lingüísticas y sea X_i definido*

sobre X el conjunto difuso que representa la semántica de cada $Etiqueta_i$. Diremos que $\{Etiqueta_1, ..., Etiqueta_n\}$ es una partición difusa lingüística de X si $\{X_1, ..., X_n\}$ es una partición difusa de X.

Para simplificar, en esta memoria, usaremos $Etiqueta_i$ para referirnos indistintamente a la etiqueta o al conjunto difuso X_i asociado.

En nuestro modelo no se imponen restricciones respecto a la forma de las funciones de pertenencia de una etiqueta dada, aparte del hecho de que ésta debe estar normalizada. Por la sencillez de representación y uso, y la versatilidad que ofrecen al representar conceptos, en este modelo hemos trabajado con conjuntos difusos trapezoidales.

Visto lo anterior, podemos presentar la primera componente de nuestro marco lingüístico, que servirá para describir lingüísticamente los valores de la variable V bajo estudio.

Definición 3.4 (Partición del dominio de la variable) *Sea V la variable bajo estudio y D_V su dominio básico. En nuestro modelo, la partición del dominio de la variable es una partición difusa lingüísica $E = \{E_1, ..., E_s\}$ definida sobre D_V.*

En segundo lugar, también será necesario particionar lingüísticamente la dimensión temporal. En esta ocasión la dimensión se encontrará organizada jerárquicamente en niveles, donde cada nivel de la jerarquía contendrá una partición del dominio temporal con distintos niveles de granularidad.

Definición 3.5 (Jerarquía en el dominio del tiempo) *Sea T la dimensión temporal y D_T su dominio básico. La jerarquía que representa la dimensión temporal se define como un conjunto de niveles*

$$L = \{L_1, ..., L_n\}$$

donde cada $L_i = \{D_{i,1}, ..., D_{i,p_i}\}$ es una partición difusa lingüística definida sobre D_T, verificando las siguientes condiciones:

1. $\forall i, j \in \{1..n\}, (i < j) \rightarrow (p_i < p_j)$
2. $\forall i \in \{2..n\}, \forall j \in \{1..p_i\}, \forall k \in \{1..p_{i-1}\} | (D_{i-1,j} \subseteq D_{i,k}) \rightarrow (D_{i-1,j} = D_{i,k})$.

De este modo establecemos que todos y cada uno de los puntos que componen la dimensión temporal debe estar cubierto al menos por una etiqueta lingüística y que

los núcleos de las etiquetas dentro de un mismo nivel no pueden solaparse entre sí. Además, en los niveles con una granularidad más fina debe haber más etiquetas que en los niveles con una granularidad más gruesa, de modo que al bajar en la jerarquía (del nivel 1 al n) se aumente el número de etiquetas (*1*). Además, una etiqueta de un nivel L_i nunca podrá ser generalizada por otra de un nivel inferior L_{i+m} (granularidad más fina), de modo que cuanto más bajemos en la jerarquía (del nivel 1 al n), por lo general, las etiquetas serán más *pequeñas* (*2*).

El uso de jerarquías en la dimensión temporal nos aporta gran versatilidad a la hora de construir un resumen. Entre otros motivos, éstas nos ofrecen la oportunidad de contar con diferentes grados de granularidad en la descripción de los datos en la dimensión tiempo, potenciando la brevedad en el resumen obtenido.

Como se indicó en el capítulo anterior, varios han sido los investigadores que se han decidido a introducir este tipo de estructuras en sus procesos de obtención de resúmenes lingüísticos. Destacaremos, sin embargo, que las jerarquías tal y como son definidas y usadas en este trabajo no aparecen en trabajos anteriores, aunque podríamos decir que hay una relación en la necesidad común que existe de tener varios niveles de descripción con el fin de dotar a las descripciones con una mayor capacidad de resumen.

En la Figura 3.1 se puede ver la representación gráfica de un ejemplo de contexto lingüístico construido siguiendo las pautas establecidas anteriormente.

3.3.2. Cuantificadores y cuantificación

Los cuantificadores lingüísticos, también conocidos como cuantificadores difusos, son etiquetas lingüísticas que nos permiten expresar de manera flexible la cantidad de elementos que satisfacen cierta condición.

Está claro que saber si se cumple una propiedad al menos una vez o en todas las ocasiones es muy útil, pero lo es todavía más ser capaces de expresar si se ha cumplido en el 90 % de los casos o sólo en el 30 % (entre otros ejemplos). Los cuantificadores difusos nos brindan la capacidad de adaptar los resúmenes lingüísticos a la forma natural en que son construídos por los seres humanos.

En este sentido, la cuantificación difusa extiende a la cuantificación convencional. Como comentamos anteriormente (Sección 3.2), esto se consigue generalizando los cuantificadores clásicos $\forall$ y $\exists$ de la lógica de primer orden y llevándolos al ámbito difuso [180]. Un gran número de aplicaciones de esta idea flexible de cuantificación aparecen en la literatura en áreas como la agregación guiada por cuantificadores [169, 171], el resumen lingüístico, la computación con palabras [184] y la cuantificación en lógica descriptiva difusa [147] entre otras.

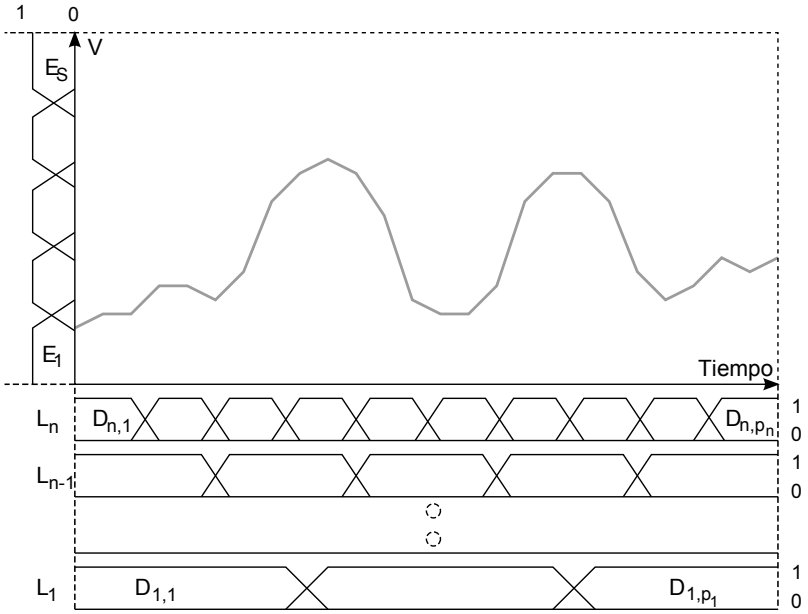
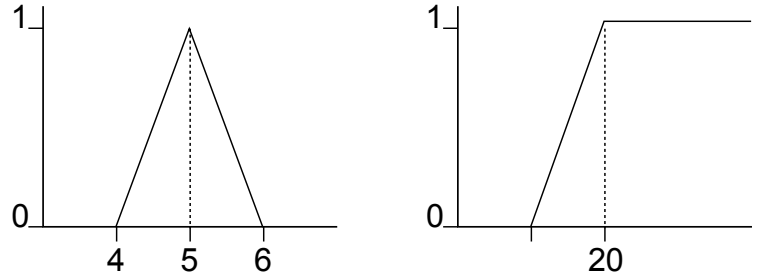


Figura 3.1: Forma general del contexto lingüístico para el resumen de series de datos.



a) Aproximadamente 5 b) Mayor que 20

Figura 3.2: Cuantificadores absolutos.

Habitualmente, los *cuantificadores lingüísticos* son subconjuntos normales y convexos de $\mathbb{Z}$ (cuantificadores absolutos) o de [0,1] (cuantificadores relativos):

- Los *cuantificadores absolutos* expresan cantidades aproximadas sobre el número total de elementos de un determinado conjunto. Figura 3.2.

- Por el contrario, los *cuantificadores relativos* expresan mediciones sobre el número total de elementos que cumplen cierta característica dependiendo del total de elementos posibles. Figura 3.3.

Algunos ejemplos de cuantificadores absolutos son *aproximadamente 5* (Figura 3.2.a) o *mayor que 20* (Figura 3.2.b). En estos casos el grado de verdad del cuantificador dependerá sólo de una única cantidad correspondiente al número total de elementos del conjunto.

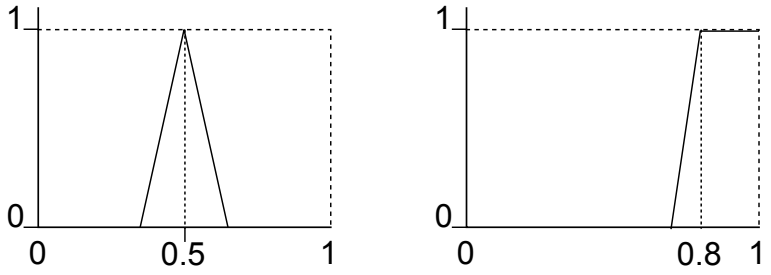
Como ejemplos de cuantificadores relativos encontramos *la mayoría* (Figura 3.3.b), *la minoría* o *aproximadamente la mitad* (Figura 3.3.a). En estos casos el grado de verdad del cuantificador se verá afectado por dos cantidades, el número total de elementos y cuántos de esos elementos cumplen la característica.

En el presente trabajo, los cuantificadores lingüísticos con los que vamos a trabajar serán cuantificadores relativos. La representación que haremos de dichos cuantificadores será mediante funciones trapezoidales. Este tipo de funciones se encuentran ampliamente extendidas debido a su facilidad de representación y uso.

Sin pérdida de generalidad, consideraremos que nuestro marco lingüístico para Q está formado por un subconjunto de cuantificadores lingüísticos relativos organizados según una relación de orden parcial. Para ello, tomamos la definición de *familia coherente de cuantificadores* propuesta en [163]:

Definición 3.6 (Familia coherente de cuantificadores) *Sea* $\mathcal{Q}=\{Q_1, ..., Q_l\}$ *un conjunto de cuantificadores lingüísticos relativos. Se dice que* $\mathcal{Q}$ *es una familia coherente de cuantificadores si verifica las siguientes condiciones:*

(I) *Las funciones de pertenencia de los elementos de* $\mathcal{Q}$ *son no decrecientes.*
(II) *Hay definida en* $\mathcal{Q}$ *una relación de orden parcial* $\succeq$*, que tiene como elemento maximal a* $Q_1 = \exists$ *y como elemento minimal a* $Q_l = \forall$*. Además,* $\forall Q_i, Q_j \in \mathcal{Q}, Q_i \subseteq Q_j \Rightarrow Q_j \succeq Q_i$*.*
(III) *La función de pertenencia del cuantificador* $\exists$ *viene expresada por* $\mu_{Q_1}(x) = 1$ *si* $x \neq 0$ *y* $\mu_{Q_1}(0) = 0$*, y la función de pertenencia del cuantificador* $\forall$ *viene*



a) Aproximadamente la mitad b) La mayoría

Figura 3.3: Cuantificadores relativos.

dada a su vez por $\mu_{Q_l}(x) = 0$ *si* $x \neq 1$ *y* $\mu_{Q_1}(0) = 1$

3.4. Estructura final del resumen

A la vista de las definiciones anteriores que nos permiten referirnos tanto a los valores de la variable como a los valores del tiempo y al cuantificador mediante etiquetas lingüísticas, podemos ahora formalizar el concepto de resumen construido como un conjunto de sentencias cuantificadas.

Nuestro método nos dará como salida un resumen lingüístico en lenguaje natural compuesto por una colección de sentencias cuantificadas construidas tomando elementos del anterior marco lingüístico. Formalmente,

Definición 3.7 (Resumen lingüístico) *Sea* $TS_{D_T}^{D_V}$ *una serie de datos temporales sobre* V *definida en* T. *Sean también* $E = \{E_1, ..., E_s\}$ *una partición lingüística difusa de* V, $L = \{L_1, ..., L_n\}$ *una jerarquía en el dominio del tiempo* T *y* $\mathcal{Q}$ *una familia coherente de cuantificadores. Un resumen lingüístico de la serie* $TS_{D_T}^{D_V}$ *se define de la siguiente manera:*

$$LS_{TS} = \{QS_1, ..., QS_h\}$$

donde QS_i *es una sentencia de la forma "*Q de $D_{i,j}^{TS}$ son A^{TS}*" donde:*

- $D_{i,j}$ *es una etiqueta* j *miembro de un determinado nivel* L_i *de la jerarquía* L *asociada a la dimensión temporal y*

$$D_{i,j}^{TS}(< t, v >) = D_{i,j}(t). \tag{3.1}$$

- A *es una etiqueta o la unión de un subconjunto de etiquetas de la partición* E *de la variable* V *estudiada y*

$$A^{TS}(< t, v >) = A(v). \tag{3.2}$$

- *y* $Q \in \mathcal{Q}$.

Podría plantearse que cualquier conjunto de sentencias que cumpla la anterior definición formal no es un resumen de la serie según la acepción del término resumen en el diccionario y que usamos como punto de partida en el anterior capítulo para motivar nuestro trabajo. Por ejemplo, tendría poco sentido construir un resumen con sentencias falsas o duplicadas. Sin embargo, la definición anterior debe entenderse como un marco formal que define la estructura del resumen en nuestro modelo. La

idoneidad de un resumen estructuralmente válido nos mete de lleno en otro apartado de crucial importancia y que complementa el modelo de resumen que acabamos de presentar: la evaluación de la calidad de un resumen.

3.5. Calidad del resumen

Una vez que tenemos el resumen construido desde el punto de vista estructural, debemos ser capaces de evaluar o medir de alguna forma la calidad del mismo. No nos podemos conformar con cualquier resumen, debemos tener unas medidas de calidad que nos aseguren que la información que aporta el resumen cumple nuestras expectativas.

Como vimos en la Sección 2.1.2 existen unos criterios que nos marcan de alguna forma cuál es la calidad de un resumen. Pero dichos criterios, aunque cualitativos, no nos ayudan a la hora de cuantificar la calidad. Ser capaces de cuantificar la calidad es muy útil para poder fijar la bondad de un resumen, pero sobre todo para ser capaces de comparar varios resúmenes entre sí.

Somos conscientes, y en eso los investigadores están de acuerdo, de que cuando nos enfrentamos a un resumen lingüístico existen una gran parte de factores que no son cuantificables y que se encuentran altamente relacionados con los gustos y preferencias del usuario. En este trabajo nos centramos en aquellos aspectos de la calidad de un resumen que, de alguna forma, sí pueden ser medidos. Para poder cuantificar estos aspectos de la calidad se deberán definir una serie de medidas de calidad u objetivos que se quieren cumplir. Lamentablemente, estas medidas no son tan sencillas de obtener.

Tratar de medir la calidad de un resumen, así como ser capaces de encontrar el mejor resumen, no es una tarea trivial, sino que encierra mucha dificultad.

- La calidad consta de muy diferentes facetas, de modo que para medir la calidad debemos ser capaces de identificar esos aspectos y, a su vez, ser capaces de medirlos, si se puede. Sin embargo, y aunque se pueden apreciar similitudes en las características que sería interesante tener en cuenta, no existe un consenso sobre el modo en el que se deben evaluar.

- Existe una relación muy fuerte entre los diferentes aspectos que dan calidad a un resumen, con la dificultad añadida de que, en la mayoría de las ocasiones, dichos objetivos pueden llegar a ser contradictorios. Esto significa que, en general, no existe el resumen óptimo o el mejor resumen de un determinado conjunto de datos. El problema de medir la calidad se convierte así en un problema de optimización multi-objetivo en el cual aparecen varios objetivos en conflicto.

- La subjetividad es una parte muy importante en la medición de la calidad, así como la sensibilidad al contexto para el que se crean las medidas. De esta manera, lo que tratamos de medir es la idoneidad de un cierto resumen, para una cierta persona, en una cierta situación.

- El espacio de búsqueda que comprende todos los posibles resúmenes de un determinado conjunto de datos (para un contexto y usuario determinados) es normalmente enorme. De modo que encontrar cuál es el mejor de todos ellos requiere un proceso computacional con una alta complejidad.

Existen numerosos métodos con los que medir la calidad de un resumen, casi tantos cómo formas de resumir. En esta sección nos centraremos en algunos de ellos para, a continuación, establecer el método de evaluación que se ha decidido usar en este nuevo enfoque para el resumen lingüístico de series de datos temporales.

3.5.1. ¿Cómo evaluar la calidad?

Cuando tenemos en nuestras manos un resumen generado por computador, la primera duda que tenemos es ¿cómo sabemos que se ha confeccionado bien? Al fin y al cabo, como se ha mencionado anteriormente, las tareas de resumir o generar un mensaje textual son inherentes al ser humano pero no tan fáciles de realizar por parte de las máquinas. El alto componente subjetivo y ligado al contexto, así como el desconocimiento exacto de los procesos que se llevan a cabo en nuestro cerebro, hacen que sea sumamente complicado simular el proceso usando un ordenador.

Un buen método con el que evaluar la calidad del resumen textual es presentarlo a *expertos* en la materia en la que se realice el resumen y preguntarles si el resumen les ha ayudado, y cómo de bien les ha ayudado, en el desempeño de su trabajo. Ésta es una buena manera de evaluación, ya que los veredictos los ofrecen las personas que van a usar los resúmenes y su opinión es la más importante. Un problema de esta técnica es que debemos de contar con la ayuda de profesionales que nos cedan su tiempo para ayudarnos. De este modo se deberá contar con un grupo variado de expertos con el fin de tener la evaluación lo menos sesgada posible. En consecuencia, este tipo de análisis puede resultar complicado de hacer debido primero a la disponibilidad de los expertos así como a los costes económicos asociados o el tiempo que se deberá dedicar a dicha tarea. En Reiter y otros [132] podemos ver un ejemplo en el cual se cuenta con ayuda de expertos que después de usar los resúmenes los valoran para así destacar tanto los pros como los contras de los mismos. La utilización de un conjunto de expertos para evaluar la calidad de un resumen tiene un inconveniente adicional: resulta difícil de materializar en el diseño de los algoritmos de generación automática por cuanto es un método de evaluación a posteriori altamente subjetivo.

Otra estrategia de evaluación de resúmenes es presentar los resúmenes a *un grupo de usuarios* para que los puntúen. En este caso, no se evaluará cómo ayuda en el desempeño del trabajo sino características específicas del resumen a través de un cuestionario. Al igual que anteriormente, deberíamos contar con una población de usuarios lo bastante amplia como para que las opiniones no contengan sesgo alguno. También son importantes los costes económicos asociados o el tiempo que se deberá dedicar a dicha tarea. Un ejemplo de método de evaluación humana sobre la calidad de resúmenes generados automáticamente lo podemos encontrar en [120] de Triviño y otros. Este método también adolece del problema de implementación en los algoritmos de generación.

La última alternativa que presentamos, a priori la menos costosa pero no por ello más fácil de implementar, es la *validación automática* de los resultados. Esta será la alternativa que usemos para nuestra propuesta de modelo de calidad porque, en este punto de nuestra investigación, damos especial importancia a la facilidad de incorporar dicho modelo en propuestas algorítmicas de generación automática de resúmenes lingüísticos y, por tanto, nos interesa hacer una aproximación *cuantitativa* al concepto de calidad. Sin embargo, en el futuro, está propuesta debe ser complementada de forma que se aborden aspectos cualitativos y subjetivos igualmente importantes.

La validación automática de resultados está muy ligada a la persona que diseña las estrategias de medida. La calidad puede ser entendida de diversas formas según la persona a la que se le pregunte. En este caso se podrían usar métricas que nos permitan comparar resúmenes generados automáticamente con otros provenientes de un corpus de resúmenes salidos de la mano de seres humanos. Para ello deberíamos tener tantos corpus como temas en los que queramos resumir y además dichos corpus deberán ser abundantes en resúmenes. De modo que se deberá incurrir en tareas de búsqueda de corpus existentes, y si no los hubiera, en tareas de confección de los mismos.

Pero no sólo imprime dificultad la búsqueda de los corpus adecuados, sino la creación de una métrica que nos permita medir la calidad cuantitativamente. Existen diferentes métricas sugeridas por distintos investigadores. En general, aún no se ha llegado a un consenso sobre qué cualidades debemos de medir en un resumen para considerarlo de calidad o no, pero sí que se observa una gran coincidencia entre algunos de las características de un resumen que debemos evaluar. Como se ha comentado anteriormente, esto se debe en gran medida a la subjetividad y al alto índice de dependencia del contexto de los resúmenes creados.

En [176] Yager y Petry, además de la exactitud propia de la cuantificación de las sentencias, introducen conceptos tales como mínima cobertura, mínima relevancia, ser sucinto y la utilidad[3] para referirse a la calidad de un resumen lingüístico generado por

[3]del inglés, minimum coverage, minimum relevance, succinctness, and usefulness.

computador. En primer lugar, los autores establecen un grado de mínima cobertura para especificar el grado de cumplimiento de una sentencia a partir del cuál el usuario estaría completamente satisfecho con la misma. En este modelo, el grado en el que un resumen es sucinto tiene mucho que ver con la longitud del mismo. Estrechamente ligado con el anterior encontramos el criterio de mínima relevancia que establece el grado hasta el cual estamos dispuestos a perder información con tal de asegurar la brevedad. Por último, la utilidad es un criterio implícito que representa el deseo de que el resumen sea útil para el usuario, y está relacionado con la precisión o imprecisión de los términos usados durante la construcción del resumen.

En [111] Niewiadomski presenta una serie de medidas, en concreto seis, con las que medir la calidad de un resumen. Dichas medidas son: imprecisión de la cuantificación, cardinalidad de la cuantificación, cardinalidad del "resumidor", e imprecisión, cardinalidad y longitud de la consulta o resumen[4]. Vemos que de nuevo aparecen términos como la imprecisión (que antes se relacionaba con la utilidad), la longitud del resumen o la cuantificación. Posteriormente en [112] se proponen un nuevo conjunto de medidas esta vez para evaluar resúmenes compuestos por conjuntos difusos de tipo 2 [5].

Varios son los trabajos en los que Kacprzyk y otros presentan un conjunto de medidas de calidad de los resúmenes generados. Como ejemplo, podríamos citar [78] de Kacprzyk y Yager. En este caso las medidas son cinco: el grado de verdad, el grado de imprecisión, el grado de cobertura, la medida en la que es apropiado un resumen y, por último, la longitud del mismo[6]. En [71] de Kacprzyk y Wilbik, se introduce el concepto de grado de especificidad[7]. Asimismo, en [72], de los mismos autores, se introduce también el concepto de grado de enfoque[8].

Reiter y otros también han expresado su preocupación a la hora de realizar una buena evaluación de los resultados obtenidos al realizar resumen de datos. Prueba de ello son una gran cantidad de trabajos sobre el tema, entre los que podemos destacar algunos como [128, 129, 137].

Como vemos muchos son los autores que han profundizado en el tema de la medida de la calidad. Cada grupo de investigación ha presentado sus propios conjuntos de medidas para evaluar la calidad. Vemos que aunque los conjuntos son muy diferentes, en esencia hay muchos conceptos o ideas que se repiten en todos los trabajos. La

[4] del inglés, quantification imprecision, quantification cardinality, summarizer cardinality, and imprecision, cardinality, and length of the query.

[5] degree of truth, degree of imprecision, degree of covering, degree of appropriatenes, length, type-2 quantification imprecision, type-2 quantification cardinality, type-2 summarizer cardinality, imprecision of the type-2 query, cardinality of the type-2 query

[6] del inglés, a truth value, degree of imprecision (fuzziness), degree of covering, degree of appropriateness y lenght of a summary.

[7] del inglés, degree of specificity.

[8] del inglés, degree of focus.

idea de que el proceso depende en gran medida del contexto y el usuario también se encuentra presente en los trabajos.

Somos conscientes de la dificultad de establecer un conjunto único de medidas que sea aceptado por toda la comunidad ya que la subjetividad siempre estará presente, y lo que es adecuado para unos puede no serlo para otros en las mismas circunstancias, incluso para la misma persona esto puede variar en función del entorno. Muchos investigadores han centrado su área de investigación en la inserción de las preferencias de usuarios en procesos informatizados.

La subjetividad no sólo se encuentra presente en el paso de interpretación de la información obtenida, sino que hace su acto de presencia mucho antes. Durante la fase de definición del contexto lingüístico, la familia de cuantificadores, establecimiento de los límites o de un umbral para la exactitud, se introducen variantes muy subjetivas en el proceso. La definición de conjuntos difusos depende del usuario pero también lo hacen las etiquetas lingüísticas que asociamos a dichos conjuntos, y lo mismo sucede para los cuantificadores.

Teniendo en cuenta los trabajos anteriores y nuestras propias ideas, hemos desarrollado un modelo de medición de la calidad de un resumen que nos permita conocer la bondad de los resúmenes generados por el ordenador y nos ofrezca la posibilidad de la comparación entre resúmenes.

3.5.2. La calidad como medio de comparar resúmenes

El objetivo principal que se busca al llevar a cabo un estudio de la calidad es ser capaz de comparar, en un momento dado, dos resúmenes desde el punto de vista de la calidad de cada uno de ellos. Esto nos ha llevado a realizar un estudio acerca de la calidad desde la perspectiva de una relación binaria de orden en el espacio de posibles resúmenes de un cierto conjunto de datos, es decir, la relación binaria deberá satisfacer al menos las propiedades reflexiva y transitiva.

Supongamos que $d = \{e_1, \ldots, e_L\}$ es un conjunto de datos y S_d el conjunto de posibles resúmenes para d; consideremos entonces la relación de calidad $\leq_C$ definida en S_d que verifica:

- $\forall s \in S_d, s \leq_C s$ (reflexividad)
- $\forall s_1, s_2, s_3 \in S_d, s_1 \leq_C s_2 \wedge s_2 \leq_C s_3 \rightarrow s_1 \leq_C s_3$ (transitividad)

Este tipo de relación recibe el nombre de *preorden*. La posibilidad de que $\leq_C$ sea un orden más estricto depende del cumplimiento de propiedades adicionales tales como la antisimetría y la comparabilidad. Éstas, a su vez, dependen del modelo de calidad empleado para determinar $\leq_C$.

A este respecto, una forma de afrontar la construcción de la relación $\leq_C$ es determinar un criterio asociado a la calidad y definir cómo se medirá el mismo. Bajo este enfoque, si la medida se define sobre el conjunto S_d al completo, la propiedad de comparabilidad está garantizada. Si, además, la medida está definida mediante una función inyectiva, la propiedad antisimétrica también estará garantizada, y en consecuencia obtendremos un orden total.

Desafortunadamente, como ya hemos comentado anteriormente, la calidad tiene diferentes facetas, y además, más de un criterio debe ser considerado. Incluso cuando es posible identificar todos los criterios y definir medidas para cada uno de ellos, el hecho de que dichos criterios deban de combinarse de algún modo (agregación de medidas, definición de algún orden entre criterios, etcétera) de una manera significativa, hace difícil la tarea de mantener las propiedades que nos llevan a un orden completo, y que dependerán del modelo de combinación seleccionado.

Los criterios asociados a la calidad son muchos y muy variados, y en la mayoría de las ocasiones, pueden llegar a ser contradictorios entre sí. Pero lo que causa más dificultades a los investigadores es la subjetividad de los mismos, que estará ligada al usuario final como receptor de resumen. En esta memoria, nos hemos centrado en ese tipo de criterios, que aún pudiendo estar afectados por la subjetividad del usuario, pueden ser calculados de forma automática a partir del conjunto de datos y el resumen obtenido.

3.5.3. Un modelo multi-dimensional de medida

En esta sección presentaremos un modelo multi-dimensional para la medida de calidad. El modelo es general y abierto y, por lo tanto, puede ser fácilmente configurable o adaptable para tener en cuenta las opiniones tanto del diseñador de los procesos de resumen, como del usuario al que van dirigidos los resultados.

Siguiendo las ideas anteriores acerca de la calidad, a continuación presentaremos un modelo multi-dimensional de calidad compuesto por cuatro criterios que cualquier buen resumen debería cumplir. En nuestro caso, diremos que un buen resumen deberá *cubrir los datos de forma sucinta y cierta* de acuerdo a los intereses del usuario. Esta definición corta nos sugiere algunos aspectos importantes que se deben tener en cuenta a la hora de medir la calidad de un resumen.

Las medidas

En primer lugar, centremos nuestra atención en la *cobertura* de los datos.

Definición 3.8 (Cobertura de un resumen - $c_d(s)$) *Consideremos un conjunto de datos d que debe ser resumido lingüísticamente y s un resumen lingüístico de d.*

Jugador	Altura (cm)
J1	183
J2	194
J3	181
J4	190
J5	204
J6	211
J7	190
J8	189
J9	203
J10	204

Tabla 3.1: Datos de ejemplo para calidad: altura de jugadores.

La cobertura de d por s, $c_d(s)$ puede ser definida como la medida en la que todos *los elementos $e_i \in d$ son considerados en s.*

Independientemente de la forma en la que calculemos la cobertura, y sin perder generalidad, podemos asumir que $c_d(s)$ es una medida difusa normalizada en d. Esto es, si $d_s = \{e_i \in d | e_i$ es considerado en $s\}$, entonces:

- Si $d_s = \emptyset$, entonces $c_d(s) = 0$.
- Si $d_s = d$, entonces $c_d(s) = 1$.
- $\forall s_1, s_2, d_{s_1} \subseteq d_{s_2} \rightarrow c_d(s_1) \leq c_d(s_2)$.

De modo que, si ninguno de los elementos de d es considerado en el resumen s, entonces $c_d(s) = 0$. De manera opuesta, cuando todos y cada uno de los datos en d son considerados en s, entonces $c_d(s) = 1$. Conforme crece el número de datos considerados en s, $c_d(s)$ crece también.

Con motivo de ir ejemplificando cada una de las medidas presentadas, iremos ilustrándolas mediante un pequeño ejemplo. En la Tabla 3.1 se muestra un conjunto d que contiene la altura en centímetros de un grupo de jugadores de baloncesto.

Consideremos ahora los siguientes resúmenes de los datos presentados:

- s_1: Hay cuatro jugadores que superan los 200 cms, otros tres en el rango [190, 200], y los otros tres no superan los 190.
- s_2: Hay cuatro jugadores que superan los 200 cms.

El resumen s_1 considera a la totalidad de jugadores, mientras que sólo cuatro de ellos son considerados por el resumen s_2. Podemos ver que el resumen s_1 cubre más datos que s_2, de modo que $c_d(s_1)$ será mayor (o igual) que $c_d(s_2)$.

Después de la definición de la cobertura, y de acuerdo con la definición inicial, nos centraremos ahora en cómo describir los datos de forma *sucinta*.

Definición 3.9 (Brevedad de un resumen - $b(s)$) *Consideremos un conjunto de datos d que debe ser resumido lingüísticamente y s un resumen lingüístico de d.*

La brevedad de s, b_s, representa la medida en la que el resumen es corto.

Sin pérdida de generalidad, y como estamos trabajando con computadoras con limitaciones físicas que imponen limitaciones a la longitud de un resumen, podemos asumir que $b(s) \in [0, 1]$.

Si obtenemos un valor $b(s) = 1$ tendremos que el resumen es lo más corto posible. Por ejemplo, si consideramos la brevedad en términos de número de sentencias, el resumen más corto será aquel con una sola sentencia. Por supuesto, cuanto mayor es la brevedad mejor es el resumen bajo este criterio.

Siguiendo con el ejemplo de los jugadores, es claro que $b(s_1)$ es menor (o igual) que $b(s_2)$.

Finalmente, con respecto a la certeza con la que el resumen describe los datos, podemos considerar dos aspectos principales: la *especificidad* y la *exactitud*.

Definición 3.10 (Especificidad de un resumen - $p(s)$) *Consideremos un conjunto de datos d que debe ser resumido lingüísticamente y s un resumen lingüístico de d.*

La especificidad de s, $p(s)$, es la medida en la que los conceptos en el resumen definen o identifican de manera clara los datos involucrados.

Como estamos considerando conjuntos de datos almacenados en computadores, podemos asumir sin perder generalidad, que $p(s) \in [0, 1]$.

El valor 1 de $p(s)$ nos informa de que los conceptos usados en el resumen describen de forma clara los datos implicados, esto es, a partir del resumen podríamos conocer los datos descritos sin sombra de duda. Cuanto mayor en la especificidad, mejor será el resumen bajo este criterio.

Volvamos de nuevo al ejemplo representado en 3.1,

- s_1: Hay cuatro jugadores que superan los 200 cms, otros tres en el rango [190, 200], y los otros tres no superan los 190.

- s_3: Un jugador excede ligeramente los 210 cms, tres jugadores se encuentran en el rango [200, 205], otros tres entre [190, 195], y los otros tres entre [180, 189].

En este caso, el resumen s_3 usa conceptos que nos ofrecen información más específica acerca de los datos implicados que los que aparecen en el resumen s_1. De modo que $p(s_3)$ será mayor que (o igual a) $p(s_1)$.

Definición 3.11 (Exactitud de un resumen - $a_d(s)$) *Consideremos un conjunto de datos d que debe ser resumido lingüísticamente y s un resumen lingüístico de d.*

La exactitud de s, $a_d(s)$, es la medida en la que lo que transmite en el resumen responde fielmente a la realidad de los datos cubiertos d_s.

Al igual que en la cobertura, independientemente de cómo se calcule la exactitud, y sin perder generalidad, podemos asumir que a_d es una medida difusa normalizada en d_s. Esto es, si $t_s = \{e_i \in d_s |$ lo que dice s acerca de e_i es verdadero $\}$, entonces:

- Si $t_s = \emptyset$, entonces $a_d(s) = 0$.

- Si $t_s = d_s$, entonces $a_d(s) = 1$.

- $\forall s_1, s_2, t_{s_1} \subseteq t_{s_2} \rightarrow a_d(s_1) \leq a_d(s_2)$.

Un valor 1 para la exactitud significa que lo que muestra el resumen para cada punto cubierto responde fielmente a la realidad. Cuanto mayor es la exactitud, mejor es el resumen en términos de este criterio.

Si regresamos al ejemplo 3.1, y consideramos:

- s_1: Hay cuatro jugadores que superan los 200 cms, otros tres en el rango [190, 200], y los otros tres no superan los 190.

- s_4: La mitad de los jugadores excede los 200 cms mientras que la otra mitad no mide menos de 180.

En este ejemplo, la exactitud del primer resumen s_1 es mayor que la del resumen s_4. Destacamos que, en este ejemplo, $c_d(s_1) = c_d(s_4) = 1$.

Estos cuatro criterios constituyen el modelo multi-dimensional de medida de la calidad que definen a una amplia familia de relaciones de calidad. La instanciación de este modelo en una ordenación específica debe tener en consideración la intuición e interés del usuario en dos aspectos:

- La definición de medidas adecuadas para cada uno de los criterios teniendo en cuenta la semántica de cada criterio y cómo éste es entendido por el usuario.
- La combinación de los cuatro criterios para determinar la relación de orden final debe satisfacer $\forall s_1, s_2 \in S, (c(s_1) \leq c(s_2) \wedge b(s_1) \leq b(s_2) \wedge s(s_1) \leq s(s_2) \wedge a(s_1) \leq a(s_2)) \rightarrow s_1 \leq_{C_d} s_2$.

La relación de orden

Definición 3.12 (Relación Básica de Ordenación de Calidad - $\leq^B_{C_d}$) *Consideremos un conjunto de datos d que debe ser resumido lingüísticamente y S_d el universo de resúmenes que puede ser construido para describir d de acuerdo con un determinado mecanismo formal. La Relación Básica de Ordenación de Calidad se define como una relación binaria $\leq^B_{C_d}$ sobre S_d:*

$\forall s_1, s_2 \in S_d, s_1 \leq^B_{C_d} s_2 \Leftrightarrow c_d(s_1) \leq c_d(s_2) \wedge b(s_1) \leq b(s_2) \wedge s(s_1) \leq s(s_2) \wedge a(s_1) \leq a(s_2)$.

Se puede ver fácilmente que se satisface el cumplimiento de las siguientes propiedades:

- $\forall s \in S_d, s \leq^B_{C_d} s$ (Reflexividad)
- $\forall s_1, s_2, s_3 \in S_d, (s_1 \leq^B_{C_d} s_2 \wedge s_2 \leq^B_{C_d} s_3) \rightarrow s_1 \leq^B_{C_d} s_3$ (Transitividad)

De este modo, la Relación Básica de Ordenación de Calidad $\leq^B_{C_d}$ define una relación de preorden en S_d.

El uso de esta relación básica tiene importantes consecuencias cuando tratamos de desarrollar técnicas algorítmicas que resuman lingüísticamente un conjunto de datos d:

- Por un lado, nos permitirá descartar unas soluciones en favor de otras cuando sean comparables.
- Por el otro, nos pone ante la realidad de que hay resúmenes que no son comparables desde el punto de vista de la calidad y deben, por tanto, considerarse como *igualmente* buenos.

3.5.4. Una instanciación del modelo de calidad para nuestro modelo de resumen

Debemos remarcar la generalidad de nuestro modelo de calidad, sobre todo, como característica positiva. El tener un modelo general nos permite concretar las medidas particulares para cada situación. De este modo, se pueden obtener diversos conjuntos de medidas en función del uso o situación para los que se creen.

En esta sección se presenta una alternativa concreta para posibilitar la evaluación de resúmenes lingüísticos generados por ordenador.

Consideraremos un resumen $LS_{TS} = \{QS_1, ..., QS_h\}$ concreto definido sobre una serie de datos $TS_{D_T}^{D_V}$ de tamaño m.

Brevedad

El conjunto de sentencias cuantificadas debe ser lo más pequeño posible. De modo que la *brevedad* se presenta como una cualidad muy importante. El resumen no debe ser excesivamente largo y para ello cada sentencia final debe cubrir un número significativo de puntos en periodos de tiempo amplios.

Una forma de medir la brevedad de un resumen, $b(s)$, es contar las sentencias cuantificadas que conforman el resumen. Si se desea llevar a cabo un proceso de normalización para mantener el valor en el rango $(0, 1]$ (siendo 1 un resumen de los más cortos posibles) se puede hacer de la siguiente forma:

$$b(LS_{TS}) = 1/h \tag{3.3}$$

Especificidad

La *especificidad* debe ser lo más alta posible. Con este término lo que se pretende es medir hasta qué punto los conceptos usados en la confección del resumen definen o identifican claramente a los datos.

Una fórmula que puede usarse para medir la especificidad para cada sentencia QS_i del resumen, $\overline{p}(QS_i)$, es

$$\overline{p}(QS_i) = \frac{area(Q_{QS_i}) + area(A_{QS_i})}{2}$$

donde $area(Q_{QS_i})$ y $area(A_{QS_i})$ son, respectivamente, el área del cuantificador y la etiqueta de valor usadas en QS_i y se suponen normalizadas en $[0, 1]$, ya que d es finito.

Cuanto mayor es el área bajo la representación de los términos lingüísticos para Q y A, menos específicos son éstos. Bajo el cuantificador *La mayoría*, en principio, siempre hay menos área que bajo *Al menos la mitad*.

Para el resumen completo, LS_{TS},

$$p(LS_{TS}) = 1 - \left(\frac{\sum_{i=1}^{\mathrm{h}} p(QS_i)}{\mathrm{h}} \right). \tag{3.4}$$

Exactitud

Cuando hablamos de la evaluación de la exactitud de una sentencia cuantificada nos referimos al proceso mediante el cual se calcula el grado de cumplimiento de dicha sentencia. Es decir, el proceso de cuantificación establece la validez o verdad de la información que aporta el resumen.

Existen diferentes aproximaciones a la hora de enfrentarse a la evaluación de sentencias cuantificadas. Los investigadores A. Niewiadomski y O. Korczak presentan en [114] un repaso de algunos de los métodos más utilizados y extendidos a la hora de realizar evaluación de sentencias cuantificadas. Para un estudio más exhaustivo, que el que aquí se presentará, acerca de las sentencias cuantificadas y sus métodos de evaluación el lector puede consultar [37].

Ninguno de los métodos de cuantificación existentes ha podido resolver el cumplimiento de todas las propiedades concernientes a los cuantificadores relativos. Sin embargo, el método *GD* cumple una serie de ellas que son bastantes interesantes. En nuestro enfoque usaremos el método *GD* de Delgado, Sánchez y Vila [38] debido a su eficiencia y a su carácter no estricto.

Siguiendo el mencionado método la evaluación de una sentencia del tipo "Q de D son A" se realizaría mediante la siguiente ecuación:

$$GD_Q(A/D) = \sum_{\alpha_i \in \triangle(A/B)} (\alpha_i - \alpha_{i+1}) Q \left(\frac{|(A \cap D)_{\alpha_i}|}{|D_{\alpha_i}|} \right) \tag{3.5}$$

donde $(A \cap D)(x) = \mathrm{Min}(A(x), D(x))$, $\triangle(A/D) = \Lambda(A \cap D) \cup \Lambda(D)$, $\Lambda(D)$ es el conjunto de niveles de D, y $\triangle(A/D) = \{\alpha_1, ..., \alpha_p\}$ con $\alpha_i > \alpha_{i+1}$ para $i \in \{1, ..., p\}$, $\alpha_1 = 1$ y $\alpha_{p+1} = 0$ (a pesar de que α_{p+1} no pertenece al conjunto de niveles, se ha considerado en la fórmula). Se asume que el conjunto D está normalizado. Si no es así, D se normalizaría y el mismo factor de normalización se aplicaría a $A \cap D$ [9].

[9] Como el conjunto de datos es finito, también se consideran finitos el conjunto D y el número relevante de α-cortes.

De este modo, podemos calcular la exactitud de una sentencia QS_i, $a_d(QS_i)$, como,

$$a_d(QS_i) = GD_{Q_{QS_i}}(A_{QS_i}/D_{QS_i})$$

donde Q_{QS_i}, A_{QS_i} y D_{QS_i} son las componentes Q, A y D de la sentencia QS_i.

Para el resumen completo, $a_d(LS_{TS})$, tenemos,

$$a_d(LS_{TS}) = \frac{\sum_{i=1}^{\mathrm{h}} \mathrm{a_d(QS_i)})}{\mathrm{h}} \tag{3.6}$$

La anterior medida de exactitud puede acompañarse fácilmente de una restricción que lleve a 0 la exactitud de un resumen si alguna de las sentencias que lo componen no supera un determinado umbral τ propuesto por el usuario.

Cobertura

En nuestro modelo, podemos calcular la cobertura como la razón de puntos de la línea temporal que se encuentran cubiertos por alguna sentencia del resumen. Para cada punto, si éste no está cubierto por alguna etiqueta de tiempo utilizada en el resumen se utiliza un 0 para el cómputo; si, en cambio, sí está cubierto se utiliza un 1.

$$c_d(LS_{TS}) = \frac{\sum_{i=1}^{n} Cov_{LS_{TS}}(t_i)}{n} \tag{3.7}$$

donde $Cov_{LS_{TS}}(t_i) = 1$ si $\exists QS_j \in LS_{TS} | t_i \in supp(D_{QS_j})$ y 0 en otro caso (con *supp* como la función que representa el soporte del conjunto que tiene como argumento).

En nuestro modelo, la *cobertura* juega un papel esencial. Hasta el punto de que puede convertirse en una restricción asociada a un umbral que lleve su valor a 0 en caso de que no se supere.

3.6. Conclusiones

El presente capítulo se ha dedicado a presentar el nuevo modelo para el resumen lingüístico de series de datos temporales.

En primer lugar nos hemos adentrado en las fases que, por consenso, se tratan de seguir cuando se quiere generar texto en lenguaje natural a través de un proceso en el ordenador. A continuación se ha relacionado cada una de ellas con las fases que hemos seguido al crear el modelo.

Del mismo modo se han definido el marco lingüístico del resumen y su estructura como conjunto de sentencias cuantificadas usando términos de ese marco lingüístico.

Finalmente, se ha presentado una discusión sobre la calidad de un resumen lingüístico generado por computador, tratando temas como ¿Qué es la calidad? ¿Cómo se define y se mide?, para, acto seguido, presentar el modelo multi-dimensional propuesto para la medición de la calidad en el modelo.

Una vez presentado el modelo de medida de la calidad se ha ofrecido al lector una posible elección de medidas concretas. Dichas medidas no son las únicas existentes, estando éstas altamente influencias por el uso que vaya a hacerse de las mismas, el tipo de problema o el contexto entre otros.

Debemos reiterar que el modelo es un modelo abierto y no ligado a cierta representación de etiquetas o cuantificadores, ni a un método de evaluación de sentencias o evaluación de la calidad determinados. Se han descrito las elecciones realizadas y se ha tratado de justificar el porqué de las mismas, pero podemos decir que el modelo seguirá funcionando si se optara por otras elecciones.

4 *Aproximaciones algorítmicas al problema*

"La amenaza de la derrota es más terrible que la derrota misma"
Gran Maestro Anatoli Karpov

Una vez presentadas nuestras propuestas, tanto el modelo de resumen lingüístico de series de datos como el modelo de calidad, pasamos a estudiar la implementación de algoritmos para la generación automática de los resúmenes lingüísticos. En primer lugar, estudiaremos la complejidad del problema en términos de la complejidad del espacio de búsqueda asociado. A continuación estudiaremos distintas aproximaciones algorítmicas para la resolución del problema.

Con el fin de sacar provecho de la estructura jerárquica en la dimensión temporal, la primera de las aproximaciones que se presentan hace uso de una filosofía Greedy en la exploración del espacio de soluciones. Los algoritmos Greedy son muy populares entre las diferentes estrategias deterministas para la optimización global de problemas. En este tipo de problemas, y con el fin de no llevar a cabo una búsqueda exhaustiva de la mejor solución entre todas y cada una de las posibles, el algoritmo Greedy optará siempre por la mejor alternativa posible en cada paso durante la construcción de la solución final. De manera que la solución final será, si no una solución óptima, al menos sí una optimal. En nuestro caso, estas decisiones se encuentran basadas en una priorización particular de nuestro modelo de calidad.

Teniendo en cuenta la amplitud del espacio de búsqueda de soluciones, y con el fin de conocer la bondad de la soluciones obtenidas mediante el enfoque Greedy, nos hemos decidido por el uso adicional de algoritmos evolutivos. Las técnicas basadas en este tipo de algoritmos realizan una exploración más amplia, sin llegar a ser exhaustiva, del espacio de soluciones del problema. Como se comentó anteriormente, los objetivos de calidad aplicados al resumen son diversos y, en general, entran en conflicto unos con otros. En este sentido, hemos considerado que el uso de algoritmos evolutivos multi-objetivo nos brinda una buena herramienta con la que enfrentarnos a la optimización de los objetivos de calidad de nuestro modelo.

4.1. El espacio de búsqueda del problema

Dado un conjunto de datos y un marco lingüístico, el espacio de búsqueda del problema está compuesto por todos los posibles resúmenes lingüísticos que se pueden construir mediante conjuntos de sentencias cuantificadas elaboradas en base al mencionado marco. Al considerar un marco finito, el espacio de soluciones es también finito.

El proceso de generación automática de resúmenes lingüísticos puede verse como un proceso de búsqueda del resumen o resúmenes deseados dentro del espacio de resúmenes posibles. El tamaño de este espacio de búsqueda, por tanto, resulta de especial interés.

Recordando que las sentencias tipo que componen los resúmenes poseen la forma

"Q de los D son A"

para determinar el tamaño del espacio de búsqueda debemos tener en cuenta los siguientes elementos:

- En relación a Q, la familia coherente de cuantificadores.

  El tamaño del subconjunto de cuantificadores determinado en el marco lingüístico es un factor a tener en cuenta al determinar el espacio de búsqueda. Normalmente, el número de los mismos no suele ser muy elevado. No es común trabajar con una familia muy extensa de cuantificadores.

- En relación a D, la jerarquía de particiones para el dominio temporal.

  La forma en la que se define la jerarquía temporal dentro de nuestro marco lingüístico se revela como el factor más determinante en el tamaño del espacio de búsqueda del problema. En este punto, de esta jerarquía debemos prestar atención, por un lado, al número de niveles, y por otro, a la partición de etiquetas de cada nivel. Más adelante, como veremos, también habrá que tener en cuenta la relación entre las etiquetas en los diferentes niveles.

- Finalmente, en relación con A, los términos lingüísticos para la variable bajo estudio.

  En general, para construir el conjunto de etiquetas lingüísticas que describan una variable, se suelen utilizar particiones con un cardinal impar en el entorno de cinco. Es decir, un número pequeño de etiquetas suele ser suficiente para describir el comportamiento de la variable (p.e. muy frío, frío, medio, caliente, muy caliente). Se pueden usar menos etiquetas para un menor nivel de precisión (p.e. frío, medio, caliente) o incrementarlas ligeramente en los casos con mayores

requerimientos de precisión (p.e. muy frío, frío, poco frío, medio, poco caliente, caliente, muy caliente). En cualquier caso, se trabaja siempre con un número moderado de etiquetas.

Si condieramos un resumen según la definición 3.7, el tamaño del espacio de búsqueda viene determinado por el número de sentencias posibles y los conjuntos que se pueden construir con ellas. En cada nivel L_i de la jerarquía, el número de sentencias posibles se puede calcular usando la Ecuación 4.1.

$$nSentencias_i = l * p_i * s \tag{4.1}$$

donde l es el número de cuantificadores utilizados, p_i es el número de etiquetas de la partición temporal para L_i y s es el número de etiquetas de la partición del dominio de la variable, tanto de partición realizada con el fin de describir la variable bajo estudio como combinaciones de las mismas.

Teniendo en cuenta lo anterior, podemos obtener el número total de sentencias a través de la Ecuación 4.2.

$$nSentencias = \sum_{i=1}^{n} nSentencias_i \tag{4.2}$$

donde n es el número de niveles en la jerarquía definida sobre la dimensión temporal.

Finalmente, como un resumen es un conjunto de sentencias no vacío, el número de combinaciones que se pueden considerar para construirlo viene determinado por la Ecuación 4.3.

$$Tam = 2^{nSentencias} - 1 \tag{4.3}$$

Recordemos que, en base a nuestro modelo de calidad, puede que no exista un único mejor resumen, sino una colección de buenos resúmenes que no se pueden ordenar entre sí. En cualquier caso, para localizar ese conjunto de resúmenes, la estrategia más directa sería explorar de manera exhaustiva el espacio de búsqueda que acabamos de describir. Por desgracia, en la mayoría de los problemas, el consumo de recursos que esta exploración requiere no es asumible. Por este motivo, presentamos a continuación dos aproximaciones heurísticas a este proceso de búsqueda que responden a los siguientes criterios:

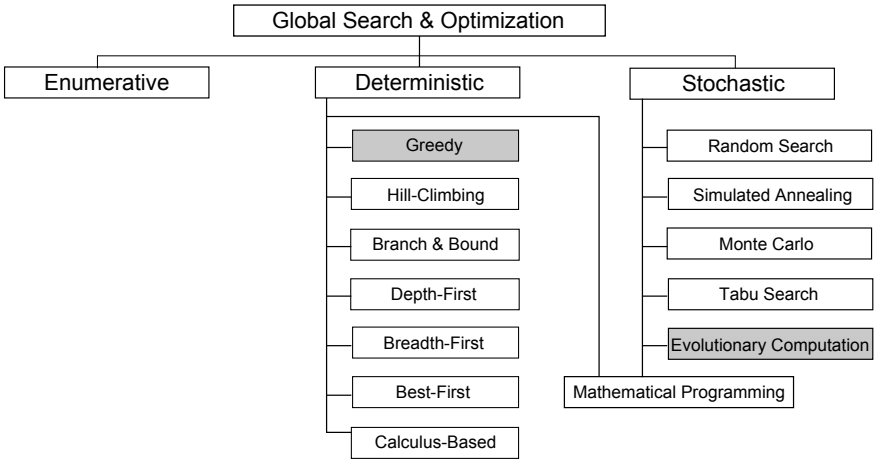


Figura 4.1: Enfoques de optimización global explorados.

- Una aproximación Greedy en la que se utiliza una particularización del modelo de calidad que prioriza algunas de las dimensiones del modelo frente a otras. Esta particularización se utiliza en el propio diseño del algoritmo. Como veremos, esta primera aproximación entrega un único resumen con calidad razonable en un tiempo que permite su uso en herramientas de consulta interactiva.

- Una aproximación evolutiva que realiza una exploración más amplia del espacio de búsqueda dando mayor protagonismo al enfoque multi-dimensional de nuestro modelo de calidad. Esta segunda aproximación, al contrario de lo que ocurre en la anterior, produce un conjunto de resúmenes alternativos para la descripción del mismo problema, aunque el consumo de recursos es muy superior.

4.2. Aproximación Greedy

Los algoritmos Greedy se caracterizan por ir construyendo paso a paso la solución final, tomando la decisión que maximiza la calidad de la solución parcial en cada paso. La principal ventaja que proporcionan estos algoritmos es la rapidez, y por ello nos hemos decantado por esta aproximación de cara a disponer de un algoritmo que pueda utilizarse en aplicaciones interactivas.

En el caso de la construcción de resúmenes lingüísticos, la estrategia Greedy consiste en ir construyendo paso a paso el resumen, añadiendo en cada paso aquella sentencia cuantificada, de entre las no incluidas aún en el resumen, que proporciona un resumen parcial de mayor calidad. Sin embargo, dado que nuestro modelo de

calidad es multi-dimensional, podemos encontrarnos con que en un paso concreto existan diversas sentencias que, una vez añadidas al resumen parcial disponible, nos proporcionen distintos posibles resúmenes parciales que no puedan ser ordenadas en términos de calidad. Por ello, para aplicar la estrategia Greedy, debemos proporcionar un criterio, basado siempre en nuestro modelo de calidad, que nos proporcione un orden total de los resúmenes. Dado que esto puede hacerse de muchas formas distintas, en la práctica es posible diseñar múltiples algoritmos Greedy para afrontar nuestro problema.

Los algoritmos que hemos diseñado, y que presentamos a continuación, se basan en el siguiente criterio de calidad, que permite la ordenación total de los resúmenes:

- En primer lugar, nuestros algoritmos buscarán soluciones dentro del subespacio de búsqueda compuesto por resúmenes cuya cobertura de los datos sea del 100 %, y tales que la exactitud de cada una de las sentencias del resumen sea superior a un umbral τ especificado previamente. Esto último implica asimismo que la exactitud total de los resúmenes considerados, calculada como el promedio de la exactitud de las sentencias que lo forman, será mayor que τ.

- Los resúmenes del subespacio anterior se ordenarán considerando en primer lugar el objetivo de brevedad, es decir, un resumen será mejor que otro si es más breve. Este objetivo es especialmente importante en aplicaciones en las que se van a obtener un gran número de resúmenes correspondientes a distintas series de datos temporales, como es el caso de la consulta a cubos OLAP con dimensión tiempo.

- A igual brevedad se considerará mejor el resumen con mayor especificidad, ya que se entiende que la exactitud es suficientemente buena al superar el umbral τ. En cualquier caso, a igual especificidad, se considerará el resumen con mayor exactitud. No se tiene en cuenta la cobertura ya que esta es del 100 % en todos los resúmenes.

Esta ordenación total obtenida a partir de nuestro modelo multi-dimensional ha sido incluida en los algoritmos Greedy diseñados de la siguiente forma:

1. La restricción de cobertura máxima queda garantizada por la condición de parada: el algoritmo añadirá nuevas sentencias al resumen hasta que se cumpla esta restricción.

2. La restricción de exactitud mayor que τ se garantiza considerando, a la hora de añadir una sentencia nueva al resumen parcial, solamente aquellas sentencias cuya exactitud sea mayor que dicho parámetro.

3. El criterio prioritario de máxima brevedad se tiene en cuenta buscando añadir en cada momento al resumen una sentencia que haga referencia al periodo de tiempo más extenso de entre aquellos que no han sido cubiertos aún por el resumen. Esta decisión es la que identifica nuestro enfoque como Greedy, ya que buscamos cubrir el conjunto completo de datos con el mínimo número de sentencias. Para ello, sacamos partido de la estructura jerárquica presente en la definición de la dimensión temporal, con el fin de acotar eficientemente la exploración del subespacio de búsqueda que anteriormente hemos descrito. Se irán considerando etiquetas de los niveles en los que hay menos etiquetas, cada una de ellas cubriendo los mayores intervalos temporales.

   Para éstas se intenta buscar una sentencia cuantificada, añadiendo un cuantificador y una etiqueta que describa la variable, que maximicen el grado de cumplimiento de la sentencia. Con el fin de intentar asegurar la brevedad en la medida de lo posible, evitaremos bajar de nivel durante la exploración de la jerarquía, permitiendo el uso de cuantificadores menos estrictos y de agrupaciones de etiquetas que describan la variable. Solamente si no es posible encontrar una sentencia adecuada para uno de dichos periodos de tiempo, se opta por explorar periodos de tiempo de granularidad más fina en la jerarquía, y que cubran el mismo periodo de tiempo (las etiquetas hijas). Por ejemplo, siempre es preferible (en cuanto al criterio de brevedad) obtener una sentencia de resumen que describa el comportamiento global durante el verano con baja especificidad (*"La mayoría de los días de verano la temperatura es alta o muy alta"*) que tener cuatro sentencias que describan con mayor especificidad el comportamiento en cada uno de los meses de verano (*"La mayoría de los días de Junio la temperatura es alta"*, *"La mayoría de los días de Julio la temperatura es muy alta"*, *"La mayoría de los días de Agosto la temperatura es muy alta"* y *"La mayoría de los días de Septiembre la temperatura es alta"*). Más nivel de detalle en esta explicación se aporta en la siguiente subsección (Subsección 4.2.1).

4. Tanto los cuantificadores menos estrictos como las etiquetas obtenidas por la unión de otras reducen la especificidad, y por tanto solo se recurre a ellas cuando no ha sido posible encontrar sentencias más específicas con un mínimo de exactitud. A la hora de buscar una sentencia menos específica que cubra el intervalo temporal considerado, podemos decantarnos por considerar un cuantificador menos estricto o una etiqueta menos específica. Estas dos opciones han dado lugar a dos algoritmos Greedy distintos, cada uno de los cuales sigue una de estas estrategias, y que describimos en el siguiente apartado. Como hemos indicado en el punto anterior, a igualdad de especificidad, el criterio de máxima exactitud se busca seleccionando, de entre todas las sentencias con la misma especificidad, aquella con mayor exactitud en cada paso, siguiendo de nuevo la estrategia Greedy.

Puede verse que, a la hora de añadir una nueva sentencia al resumen, se incorpora aquella sentencia que proporciona una mayor calidad al resumen parcial en ese momento, siguiendo la estrategia Greedy. Como es habitual en los algoritmos Greedy, esta estrategia no nos garantiza obtener el mejor resumen en términos del criterio de calidad que hemos especificado, pero sí una optimal, y de una manera rápida. En el siguiente apartado describimos nuestros dos algoritmos con más detalle.

4.2.1. Estrategias

Según nuestro modelo propuesto para el resumen lingüístico de series de datos temporales, consideramos T la dimensión temporal y D_T su dominio descrito en el nivel más fino de granularidad, y V la variable bajo estudio con D_V como dominio básico. La serie temporal TS definida en D_T se representa como $TS = \{< t_1, v_1 > , \ldots, < t_m, v_m >\}$ donde cada v_i es un valor en el dominio D_V.

Además, consideramos que el marco lingüístico está formado por una partición $E = \{E_1, \ldots, E_s\}$ de V donde cada E_i es una etiqueta lingüística y una partición jerárquica en T como un conjunto de niveles $L = \{L_1, \ldots, L_n\}$ donde n es el número de niveles de la jerarquía, y donde cada nivel L_i se encuentra particionado como $L_i = \{D_{i,1}, \ldots, D_{i,p_i}\}$. Asimismo, para todo $A = \bigcup_{E_i \in E'}$ con $E' \subseteq E$, y para todo $D_{i,j}$, definimos $D_{i,j}^{TS}(< t, v >) = D_{i,j}(t)$ y $A^{TS}(< t, v >) = A(v)$ como los subconjuntos difusos de la serie de datos temporal inducidos por $D_{i,j}$ y A, respectivamente.

Recordemos también que las etiquetas hijas de otra etiqueta se definen como $hijas(D_{n,j}) = \emptyset$ para todo j. En otro caso, $hijas(D_{i,j}) = \{D_{i-1,k}, k \in \{1..p_{i-1}\} | D_{i-1,k} \cap D_{i,j} \neq \emptyset\}$.

Contamos finalmente con un subconjunto de una familia coherente de cuantificadores $\{Q_1, \ldots, Q_{qmax}\}$, con $Q_{i+1} \subset Q_i$, $Q_1 \subseteq \exists$, y $\forall \subseteq Q_{qmax}$, y que por tanto están ordenados en orden creciente de especificidad y a la vez, de menos a más estricto en cuanto a la restricción que representan sobre el cardinal.

Con el fin de cubrir un periodo de tiempo $D_{i,j}$ determinado, la primera estrategia explora el espacio de búsqueda de sentencias cuantificadas de la forma *"Q de $D_{i,j}^{TS}$ son A^{TS}"*, mediante el uso de una filosofía *primero en profundidad* con respecto de A. En cambio, en la segunda estrategia se explora primero en profundidad con respecto a Q. Veámoslo con más detalle en las siguientes subsecciones.

Primera estrategia: preferencia por cuantificadores más específicos

La primera estrategia (Algoritmo 1) explora todas las posibles etiquetas A antes de probar con un cuantificador menos estricto de la familia. La exploración se comienza con las etiquetas A más específicas y desde el cuantificador más estricto, para a continuación ir probando otras alternativas.

Con el fin de permitir que el usuario introduzca sus preferencias, que serán tenidas en cuenta durante el proceso de búsqueda, se han introducido dos parámetros: $Glim$, para controlar hasta qué punto se desea agrupar (número máximo de etiquetas que se agrupan, por ejemplo con $Glim = 3$ se podrán usar las combinaciones *medio*, *medio o bajo*, *medio o bajo o muy bajo*, mientras que con $Glim = 2$ sólo se podrán usar las dos primeras, es decir etiquetas simples y agrupaciones de dos etiquetas); y $Qlim$, para controlar hasta qué punto está permitido trabajar con cuantificadores menos estrictos (por ejemplo, si $Qlim = 2$ se trabajará con los dos cuantificadores más estrictos, en cambio si $Qlim = 3$ se trabajará con los tres más estrictos).

Además de los anteriores, el usuario deberá determinar también un umbral τ que marcará el grado de cumplimiento mínimo que se exige a las sentencias cuantificadas que componen el resumen.

Con la intención de obtener resúmenes que sean lo más breves posible, la exploración de la jerarquía comienza por el nivel más alto, o más abstracto, de la misma. Al contener este nivel los periodos $D_{i,j}$ más amplios, cubrirán un periodo de tiempo mayor y, por lo tanto, es la decisión óptima en este paso de cara a obtener resúmenes de mayor brevedad, aunque no necesariamente la óptima de cara a la mejor solución global.

Podemos generalizar lo anteriormente comentado al respecto de los parámetros Glim y Qlim, definiéndolos en cada nivel de la jerarquía. Así, para cada nivel L_i de la jerarquía se tiene un valor para el $Qlim_i$ que marcará el cuantificador menos estricto a tener en cuenta antes de tomar en consideración la exploración de una agrupación diferente en A^{TS}. También se cuenta con un límite $Glim_i$ para indicar el número máximo de etiquetas E_i agregadas en una sentencia del resumen.

El conjunto $ParaResumir$ es una colección ordenada de periodos de tiempo para los cuales todavía no se tienen sentencias cuantificadas que los resuman. La inicialización de $ParaResumir$ con el conjunto de etiquetas del nivel L_1, como veremos, garantiza la restricción de cobertura que rige en nuestra estrategia Greedy. Por el contrario, el conjunto $Resumidos$ es la colección de periodos de tiempo que sí han sido resumidos. Por último, $Resumen$ contendrá las sentencias cuantificadas que componen el resumen.

Si es posible obtener un grado de cumplimiento mayor o igual que τ para un cierto periodo de tiempo $D_{i,j}$, usando un cuantificador Q y una etiqueta simple (línea 11), el algoritmo compone una sentencia de resumen para dicho periodo (Q_p *de* $D_{i,j}^{TS}$ *son* A^{TS} en línea 12).

Si por el contrario esto no es posible, el algoritmo lo intenta con la unión de diferentes subconjuntos de etiquetas (línea 9): parejas, tríos, cuartetos, etcétera, hasta que el grado de cumplimiento de la sentencia sea mayor que o igual a τ. El tamaño del

Algoritmo 1 : Primera estrategia.

Entrada: Una serie de tiempo TS.

Una partición jerárquica difusa de la dimensión temporal D_T.

Una partición difusa de la variable, V, $E = \{E_1, \ldots, E_s\}$.

Un subconjunto totalmente ordenado de una familia coherente de cuantificadores $\{Q_1, \ldots, Q_{qmax}\}$.

Un umbral τ de mínimo grado de cumplimiento para las sentencias cuantificadas.

Para cada nivel i, número máximo de: a) cuantificadores a usar ($Qlim_i$), y b) etiquetas que agrupar ($Glim_i$).

Salida: Un resumen de TS compuesto por un conjunto de sentencias cuantificadas.

1: $ParaResumir \leftarrow L_1$;
2: $Resumen \leftarrow \emptyset$; $Resumidos \leftarrow \emptyset$;
3: **mientras** $ParaResumir \neq \emptyset$ **hacer**
4: Toma $D_{i,j} \in ParaResumir$
5: $ParaResumir \leftarrow ParaResumir \backslash \{D_{i,j}\}$;
6: $p \leftarrow qmax$; $cubierto \leftarrow$ **falso** ;
7: **mientras** $p > (qmax - Qlim_i)$ y *no cubierto* **hacer**
8: $k \leftarrow 1$;
9: **mientras** $k \leq Glim_i$ y *no cubierto* **hacer**
10: Sea $A \leftarrow argmax^*_{B \in C_k} GD_{Q_p}(B^{TS}/D^{TS}_{i,j})$;
11: **si** $GD_{Q_p}(A^{TS}/D^{TS}_{i,j}) \geq \tau$ **entonces**
12: $Resumen \leftarrow Resumen \cup \{Q_p$ de $D^{TS}_{i,j}$ son $A^{TS}\}$;
13: $Resumidos \leftarrow Resumidos \cup (D_{i,j})$;
14: $cubierto \leftarrow$ **cierto** ;
15: **fin si**
16: $k \leftarrow k + 1$;
17: **fin mientras**
18: $p \leftarrow p - 1$;
19: **fin mientras**
20: **si** *no cubierto* y $i < n$ **entonces**
21: $ParaResumir \leftarrow ParaResumir \cup hijas(D_{i,j})$;
22: **si no**, **si** $i = n$ **entonces**
23: $Resumen \leftarrow Resumen \cup \{D^{TS}_{i,j}$ es altamente variable$\}$;
24: **fin si**
25: **fin mientras**

{La función *argmax** es una modificación de la función *argmax* de manera que en lugar de devolver el conjunto de argumentos que dan lugar al máximo valor para la función, se devuelva el primero de ellos que fue encontrado. En este caso devolverá como A la que se encontró en primer lugar de entre las posibles combinaciones de E_i que maximicen el valor GD para una sentencia formada por el cuantificador Q_p y el periodo de tiempo $D^{TS}_{i,j}$.}
{El orden de exploración de las etiquetas E_i viene establecido por el usuario desde el proceso de definición de las mismas.}
$\{C_k = \{\cup_{E_h \in F} E_h \mid F \subseteq E, |F| = k\}\}$
$\{hijas(D_{n,j}) = \emptyset$ para todo j. En otro caso, $hijas(D_{i,j}) = \{D_{i-1,k}, k \in \{1..p_{i-1}\} | D_{i-1,k} \cap D_{i,j} \neq \emptyset$ y $\neg \exists D \in ParaResumir \cup Resumidos$, $(D_{i-1,k} \cap D_{i,j}) \subseteq D\}$.}

subconjunto viene dado por el índice k, siendo $Glim_i$ el máximo valor posible para el nivel.

Cuando se ha encontrado un resumen adecuado para un cierto periodo de tiempo decimos que dicho periodo ha sido cubierto (línea 14).

Si, entre todas las combinaciones posibles de Q y A, el algoritmo no encuentra ninguna adecuada, se volverá a repetir el proceso, pero esta vez con un cuantificador Q menos estricto, hasta que $Qlim_i$ sea alcanzado.

Si, a pesar de todo, al final del proceso no se ha encontrado ningún resumen adecuado para un determinado periodo de tiempo $D_{i,j}$, el algoritmo tratará de obtenerlo con las etiquetas que presentan intersección no vacía en el nivel inferior, es decir, las etiquetas hijas $hijas(D_{i,j})$ (línea 21). Si $hijas(D_{i,j}) = \emptyset$ se añadirá al resumen una sentencia indicando la alta variabilidad observada en los datos para dicho periodo ($D_{i,j}^{TS}$ *es altamente variable* en la línea 23).

Segunda estrategia: preferencia por términos A^{TS} más específicos

Como estrategia alternativa a la presentada anteriormente, la que se muestra a continuación presta más atención a encontrar sentencias con términos lingüísticos para A^{TS} más específicos. El Algoritmo 2 refleja esta estrategia.

Como se puede observar, en este segundo algoritmo existe un intercambio entre las líneas 7 y 9 (intercambio entre bucles). El resto de líneas son esencialmente iguales, de modo que se podría decir que el mecanismo es casi el mismo. En este caso, si no es posible obtener un grado de cumplimiento mayor que o igual a τ para un cierto periodo de tiempo usando un cuantificador y una etiqueta simple, el procedimiento evalua de nuevo la sentencia pero esta vez con un cuantificador menos estricto antes de considerar grupos de etiquetas más grandes para A^{TS}.

Discusión

Ambas estrategias buscan la mayor *brevedad* con el máximo de *cobertura*. Recordemos que ambos criterios de calidad se consiguen gracias a la forma en la que se explora la estructura jerárquica presente en la dimensión temporal.

Del mismo modo ambas estrategias buscan maximizar la *exactitud* de las sentencias y, por lo tanto, del resumen final. Tanto la primera estrategia como la segunda buscan la combinación de cuantificador y etiqueta, o disyunción de etiquetas, que maximicen el grado de cumplimiento de la sentencia cuantificada para un cierto periodo.

Sin embargo, como característica diferenciadora entre ambas estrategias, encontramos la manera en la que éstas intentan alcanzar la máxima *especificidad*. Recordemos que dicha medida de calidad depende de la especificidad del cuantificador junto con la

Algoritmo 2 : Segunda estrategia.

Entrada: Una serie de tiempo TS.

Una partición jerárquica difusa de la dimensión temporal D_T.

Una partición difusa de la variable, V, $E = \{E_1, \ldots, E_s\}$.

Un subconjunto totalmente ordenado de una familia coherente de cuantificadores $\{Q_1, \ldots, Q_{qmax}\}$.

Un umbral τ de mínimo grado de cumplimiento para las sentencias cuantificadas.

Para cada nivel i, número máximo de: a) cuantificadores a usar ($Qlim_i$), y b) etiquetas que agrupar ($Glim_i$).

Salida: Un resumen de TS compuesto por un conjunto de sentencias cuantificadas.

1: $ParaResumir \leftarrow L_1$;
2: $Resumen \leftarrow \emptyset$; $Resumidos \leftarrow \emptyset$;
3: **mientras** $ParaResumir \neq \emptyset$ **hacer**
4: Toma $D_{i,j} \in ParaResumir$
5: $ParaResumir \leftarrow ParaResumir \backslash \{D_{i,j}\}$;
6: $k \leftarrow 1$; *cubierto* $\leftarrow$ **falso** ;
7: **mientras** $k \leq Glim_i$ y *no cubierto* **hacer**
8: $p \leftarrow qmax$;
9: **mientras** $p > (qmax - Qlim_i)$ y *no cubierto* **hacer**
10: Sea $A \leftarrow argmax^*_{B \in C_k} GD_{Q_p}(B^{TS}/D^{TS}_{i,j})$
11: **si** $GD_{Q_p}(A^{TS}/D^{TS}_{i,j}) \geq \tau$ **entonces**
12: $Resumen \leftarrow Resumen \cup \{Q_p$ de $D^{TS}_{i,j}$ son $A^{TS}\}$;
13: $Resumidos \leftarrow Resumidos \cup (D_{i,j})$;
14: *cubierto* $\leftarrow$ **cierto** ;
15: **fin si**
16: $p \leftarrow p - 1$
17: **fin mientras**
18: $k \leftarrow k + 1$;
19: **fin mientras**
20: **si** *no cubierto* y $i > n$ **entonces**
21: $ParaResumir \leftarrow ParaResumir \cup hijas(D_{i,j})$.
22: **si no**, **si** $i = n$ **entonces**
23: $Resumen \leftarrow Resumen \cup \{D^{TS}_{i,j}$ es altamente variable$\}$
24: **fin si**
25: **fin mientras**

especificidad de las etiquetas de descripción de la variable. Cada una de las estrategias intenta maximizar en mayor medida uno de los dos componentes anteriores.

Mientras que en la primera estrategia se trata de obtener resúmenes que contengan cuantificadores más restrictivos, maximizando la especificidad de los mismos, en la segunda estrategia se intentan obtener resúmenes en los que aparezcan grupos de etiquetas más pequeños que describan la variable bajo estudio, es decir, maximizando la especificidad del término de resumen. De este modo la primera estrategia opta por sentencias del tipo *"Todos los días de verano la temperatura es alta o muy alta"*, la segunda opta por sentencias del tipo *"Al menos el 80 % de los días de verano la temperatura es muy alta"*, siempre que sea posible. Donde *Todos* es más específico que *Al menos el 80 %*, y *muy alta* es más específico que *alta o muy alta*.

Son las preferencias del *diseñador*, respecto a en qué término prefiere una carga mayor de especificidad, las que han quedado plasmadas en los algoritmos finales dada su naturaleza Greedy. Escoger entre una de las dos estrategias es una elección importante que debe tomar el *usuario*. La elección deberá hacerse de acuerdo a las preferencias que tenga en ese momento por una cierta tipología de resúmenes: si se prefieren cuantificadores más estrictos, Algoritmo 1, o si por el contrario, se prefieren términos lingüísticos más precisos, Algoritmo 2.

Debemos puntualizar que ninguna estrategia se considera mejor o peor que la otra en términos absolutos y, que la elección de una de ellas en detrimento de la otra se realizará de acuerdo a razones subjetivas del usuario. En la mayoría de las ocasiones, las dos estrategias Greedy darán, para un mismo conjunto de datos, resúmenes muy similares, no siendo extraordinario que puedan dar el mismo resultado.

Los resultados obtenidos dependerán, además de la estrategia elegida, del contexto lingüístico que se ha construido, de la familia de cuantificadores definida, el umbral τ, los límites impuestos a cada nivel por el usuario, y por supuesto de la serie de datos en sí.

El resumen final estará compuesto por un conjunto de sentencias cuantificadas. Como se comentó anteriormente, y siguiendo una de las fases en la generación del lenguaje natural (NLG), existe la posibilidad de llevar a cabo un proceso posterior sobre dichas sentencias. Dicho proceso fusiona varias sentencias en una si éstas poseen partes comunes. Existen diversos criterios que marcarán el proceso de fusión (se pueden fusionar sentencias si cubren distintos periodos de tiempo con la misma etiqueta o si cubren distintos periodos de tiempo con el mismo cuantificador, entre otros). De esta forma se obtendrá un párrafo que se asemejará bastante al que podría haber creado un usuario humano.

Por ejemplo, si tenemos en el resumen las sentencias,

La mayoría de los días en clima cálido, el flujo de pacientes es bajo o muy bajo.
La mayoría de los días en Marzo, el flujo de pacientes es bajo o muy bajo.

podemos fusionar dichas sentencias eliminando la repetición de partes comunes de la siguiente forma,

La mayoría de los días en clima cálido y Marzo, el flujo de pacientes es bajo o muy bajo.

para que el texto se reduzca y, por lo tanto, se aumente la legibilidad.

Este proceso es especialmente importante en el caso de las sentencias que indican una alta variabilidad en determinados periodos de tiempo ya que, por el diseño de los algoritmos, puede ocurrir que un periodo de tiempo muy extenso presente alta variabilidad y que, en lugar de tener una sola sentencia para indicar este hecho, tengamos un conjunto de sentencias indicando este hecho para todas las etiquetas hijas del periodo en el nivel más específico de la jerarquía de tiempo.

Complejidad algorítmica de las estrategias

Anteriormente, en la Sección 4.1 se realizó un estudio del tamaño del espacio de búsqueda para determinar la complejidad del problema. En esta sección, aunque también se trata la complejidad, se hace desde otro punto de vista. En esta ocasión lo que trata de determinarse es la complejidad del algoritmo usado para la exploración del espacio de búsqueda.

Podemos definir el concepto de algoritmo como una descripción precisa de una secuencia de pasos que se deben seguir para alcanzar la solución a un problema dado. La complejidad del algoritmo, más comúnmente denominada complejidad algorítmica, representa la cantidad de recursos que necesita un algoritmo para resolver un problema por lo que nos permitirá determinar la eficiencia del mismo.

En general, los criterios que se emplean para evaluar dicha complejidad no proporcionan medidas absolutas sino medidas relativas al tamaño del problema. En el caso de los algoritmos Greedy el espacio de búsqueda explorado se encuentra fuertemente acotado por la dinámica de exploración y los distintos parámetros. Esto hace que aunque el espacio de búsqueda sea amplio, gracias a la incorporación de conocimiento, el espacio realmente explorado es menor, y por tanto la complejidad en términos de tamaño será menor. Veamos el proceso de forma más detallada para la primera de las estrategias Greedy (Algoritmo 1).

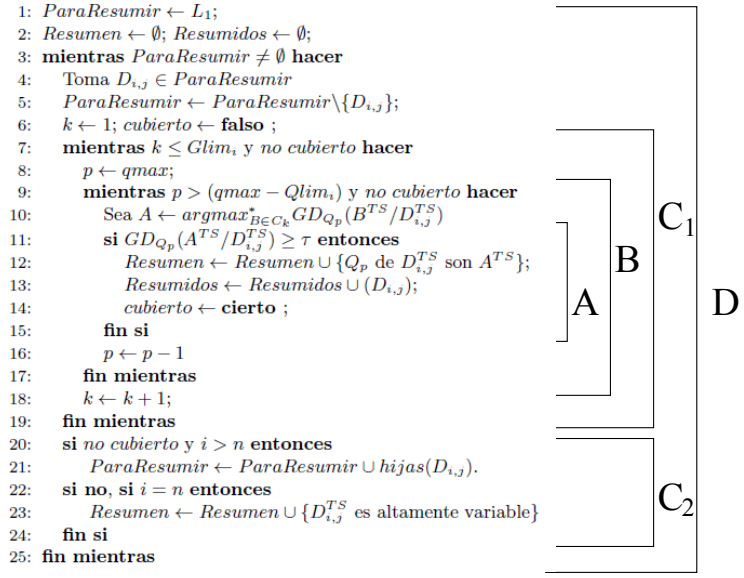


Figura 4.2: Bloques de código para el estudio de la complejidad del Algoritmo 1.

Para ilustrar el cálculo paso a paso nos ayudaremos de un gráfico en el que de forma visual se identifican los diferentes bloques a los que se les calculará la complejidad y que nos llevarán a obtener la complejidad total. En la Figura 4.2 podemos ver dicha partición.

Como siempre en estos casos comenzaremos analizando por la estructura más interna para acabar por el más externa (es decir, desde el bloque A hasta el D).

Bloque A, líneas de la 11 a la 15 inclusive: **Si** para la comprobación del valor obtenido del GD con el umbral de mínimo cumplimiento. La condición se efectúa en un orden O(1). En el peor de los casos, la condición es verdadera y se ejecutan las tres líneas siguientes (12, 13 y 14) cada una de ellas con un coste asociado de O(1). El coste total de la operación es de O(1).

Bloque B, líneas de la 8 a la 17 inclusive: **Mientras** que cicla entre las diferentes posibilidades para describir la variable para un cuantificador y un periodo determinados. El número de sentencias a comprobar es el número de posibles combinaciones de k etiquetas E_i, que se expresa de la forma $\frac{s!}{(s-k)!k!}$, donde s es el número total de etiquetas. La evaluación de una sentencia mediante GD puede hacerse en $O(1)$ si se utiliza un número fijo de alfa-cortes. En el peor de los casos el bucle (línea 9) se tendrá que ejecutar $Glim_i$ veces, de modo que finalmente para el trozo de código completo

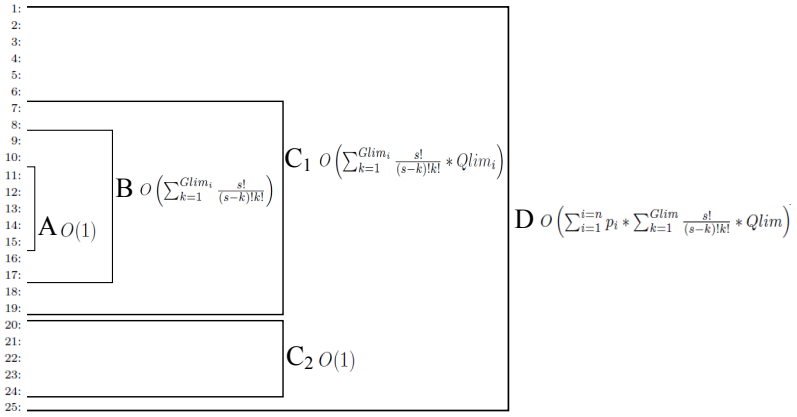


Figura 4.3: Complejidad de cada uno de los bloques de código del Algoritmo 1.

tenemos que la eficiencia es de $O\left(\sum_{k=1}^{Glim_i} \frac{s!}{(s-k)!k!}\right)$ con $\sum_{k=1}^{Glim_i} \frac{s!}{(s-k)!k!} \leq 2^s$, siendo $O(2^s)$ si y solo si $Glim_i = s$.

Bloque C_1, líneas de la 7 a la 19 inclusive: **Mientras** que cicla entre los diferentes cuantificadores. En el peor de los casos se tendrán que explorar todos los posibles cuantificadores, cuyo número viene marcado por $Qlim_i$. Podemos decir que la eficiencia del siguiente código es $O\left(\sum_{k=1}^{Glim_i} \frac{s!}{(s-k)!k!} * Qlim_i\right)$.

Bloque C_2, líneas de la 20 a la 24 inclusive: Tanto la evaluación de la expresión, como la parte del **si** y la parte del **si no** poseen una eficiencia de $O(1)$; de modo que la eficiencia de la estructura completa es de $O(1)$.

Bloque D, líneas de la 1 a la 25 inclusive: Podemos decir que el peor de los casos se ejecutará una vez por cada etiqueta existente en la jerarquía; es decir $O(\sum_{i=1}^{i=n} p_i)$.

Finalmente tendremos que la eficiencia del algoritmo completo en el peor de los casos es de $O\left(\sum_{i=1}^{i=n} p_i * \sum_{k=1}^{Glim} \frac{s!}{(s-k)!k!} * Qlim\right)$, siendo $Glim = \max_{i=1}^{n} Glim_i$ y $Qlim = \max_{i=1}^{n} Qlim_i$.

Llegados a este punto vamos a recuperar la figura que representa los bloques de código pero con una pequeña variación. En esta ocasión no aparece el código sino los bloques con la complejidad asociada tal y como se ha ido calculando en párrafos anteriores (ver Figura 4.3).

Este orden de complejidad puede expresarse de forma menos precisa pero más concisa como $O\left(\sum_{i=1}^{i=n} p_i * \frac{s!}{(s-Glim)!(Glim-1)!} * Qlim\right)$ dado que $\sum_{k=1}^{Glim} \frac{s!}{(s-k)!k!} < Glim * \frac{s!}{(s-Glim)!Glim!} = \frac{s!}{(s-Glim)!(Glim-1)!}$. De forma aún menos precisa pero más

concisa, tenemos $O\left(\sum_{i=1}^{i=n} p_i * 2^s * Qlim\right)$. Esta última expresión resulta útil para dejar claro que el algoritmo es bastante eficiente, ya que como dijimos al describir el marco lingüístico, habitualmente $s \in \{3, 5, 7\}$ y $Qlim$ suele ser un número incluso menor a s. Además, como hemos indicado, esta última expresión es una cota muy superior al número real de operaciones en el peor caso.

El estudio de la complejidad para el Algoritmo 2 correspondiente a la segunda de las estrategias Greedy da como resultado la misma expresión. Recordemos que la diferencia entre ambos algoritmos era la colocación de los bucles que recorrían las etiquetas y los cuantificadores. Los bucles anidados multiplican sus complejidades de modo que lo que tendríamos serían dos formas diferentes de llegar hasta el mismo resultado, cada una con distinto orden en los factores de la multiplicación.

Como se puede ver, el uso de la estrategia Greedy propuesta, elimina del espacio de búsqueda la exponenciación señalada en la Ecuación 4.3 gracias a un barrido lineal de, en el peor de los casos, todas las etiquetas de la dimensión tiempo.

4.2.2. Ilustración del comportamiento de los algoritmos

A continuación, y a través de un ejemplo concreto, se realizará una ilustración de los distintos comportamientos de los algoritmos para un problema dado, así como la influencia de los diferentes parámetros que el usuario puede ajustar para adecuar el resumen a sus necesidades.

Ejemplo: Centro de salud C_A

Consideramos un almacén de datos con información relativa a centros médicos en un territorio específico. Contamos con un cubo de datos existente en dicho almacén con información relacionada con la afluencia de pacientes de acuerdo a diferentes dimensiones, siendo éstas *centro médico*, *sexo del paciente* y *tiempo*. Si aplicamos una serie de operaciones OLAP sobre dicho cubo podemos obtener una serie de datos que describe *el flujo de pacientes a un centro C_X a lo largo del tiempo*. En nuestro caso tenemos una serie de 365 datos representando "la afluencia masculina a un centro C_A durante un año completo".

Una vez que tenemos en nuestro poder la serie que queremos resumir, debemos establecer el *contexto o marco lingüístico* que nos permita transformar los datos numéricos en lenguaje natural. Para ello, la dimensión que representa a la variable bajo estudio se ha particionado haciendo uso de cinco etiquetas que describen el flujo de pacientes. Se puede ver el resultado de dicha partición en la Tabla 4.1.

En este caso se ha establecido 500 como la cota máxima, es decir, el máximo número de pacientes varones que pueden asistir al centro C_A a lo largo de un día. Para problemas en los que los límites tiendan a infinito se debe realizar un estudio de

Etiqueta	Definición
muy baja	(0, 0, 90, 110)
baja	(90, 110, 190, 210)
media	(190, 210, 290, 310)
alta	(290, 310, 390, 410)
muy alta	(390, 410, 500, 500)

Tabla 4.1: Partición del dominio de la variable para el ejemplo C_A.

valores, bien analizando un conjunto lo suficientemente grande de valores anteriores o hallando los valores máximo y mínimo del conjunto actual y usándolos para establecer las cotas, entre otros.

La dimensión temporal, en cambio, está descrita mediante una jerarquía de tres niveles, cada uno de ellos con particiones de diferente granularidad.

Las particiones de mayor y media granularidad se componen de etiquetas que simbolizan conceptos meteorológicos. En el primer nivel se ha optado por dos etiquetas que dividen el tiempo en dos grandes periodos como son días del año con clima extremo (temperaturas extremas ya sea frías o cálidas) o días del año con clima templado (temperaturas medias). En el segundo nivel, se ha ampliado el número de etiquetas usadas a cuatro. Se han dividido los días de clima extremo en días cálidos o días fríos, y los días templados en periodos de transición entre los periodos anteriores. En estos niveles es muy intuitivo el uso de particiones difusas ya que representan de forma adecuada que los cambios en la temperatura son graduales y no se producen de forma abrupta entre un periodo y otro; otro tanto cabe decir del cambio de estación, entendido de nuevo como un cambio general en el clima. También es razonable considerar periodos difusos dado que no es habitual que se produzcan grandes cambios en la afluencia de personas a un centro de un dia para otro en el límite crisp habitual del cambio entre estaciones.

La partición de grano más fino está compuesta por etiquetas lingüísticas que representan a los doce meses convencionales. Al igual que en el caso anterior la introducción de las fronteras difusas nos aporta flexibilidad a la hora de reflejar comportamientos. En general la afluencia a un centro de salud no varía de forma drástica el 1 de Febrero con respecto al 31 de Enero. El uso de esta representación nos evita la fuerte dependencia de los resúmenes con respecto a los límites rigurosos presentes en los meses convencionales.

La forma de establecer las etiquetas dependerá de la semántica que quiera usar la persona en el resumen, es decir, del uso que desea hacer del lenguaje. Dependiendo de la situación concreta podría interesar contar con transiciones rigurosas o con transiciones suaves en diversa medida. El uso de las segundas favorece la transición suave

Nivel	Etiqueta	Definición
1	clima extremo	(1, 1, 58, 60) $\bigcup$ (138, 140, 258, 260) $\bigcup$ (320, 322, 365, 365)
	clima templado	(58, 60, 138, 140) $\bigcup$ (258, 260, 320, 322)
2	clima frío	(1, 1, 58, 60) $\bigcup$ (320, 322, 365, 365)
	clima cálido	(138, 140, 258, 260)
	clima de cálido a frío	(58, 60, 138, 140)
	clima de frío a cálido	(258, 260, 320, 322)
3	Enero	(1, 1, 30, 32)
	Febrero	(30, 32, 58, 60)
	Marzo	(58, 60, 89, 91)
	Abril	(89, 91, 119, 121)
	Mayo	(119, 121, 150, 152)
	Junio	(150, 152, 180, 182)
	Julio	(180, 182, 211, 213)
	Agosto	(211, 213, 242, 244)
	Septiembre	(242, 244, 272, 274)
	Octubre	(272, 274, 303, 305)
	Noviembre	(303, 305, 333, 335)
	Diciembre	(333, 335, 365, 365)

Tabla 4.2: Partición de la dimensión temporal para el ejemplo C_A.

Cuantificador	Definición
La mayoría	(0, 0.7, 0.9, 1)
Aproximadamente más del 70 %	(0, 0.6, 0.8, 1)
Aproximadamente más de la mitad	(0, 0.4, 0.6, 1)

Tabla 4.3: Cuantificadores para el ejemplo C_A.

durante la exploración del espacio de soluciones.

En la Tabla 4.2 están representadas las etiquetas lingüísticas utilizadas en las diferentes particiones que componen la jerarquía temporal.

Como resumen, la Figura 4.4 muestra una representación gráfica de la serie temporal y del marco lingüístico que se aplicará en el proceso de transformación.

Además del contexto lingüístico se debe definir el conjunto de cuantificadores que se usará durante el proceso. La definición de los mismos se puede ver en la Tabla 4.3.

Por último, se debe proporcionar el umbral τ y los límites, que en este caso son $\tau = 0.7$ y $Qlim_i = Glim_i = 2$ para todos los niveles i de la dimensión temporal; es decir, se usan dos cuantificadores desde el más estricto y están permitidas las agrupaciones de hasta dos etiquetas.

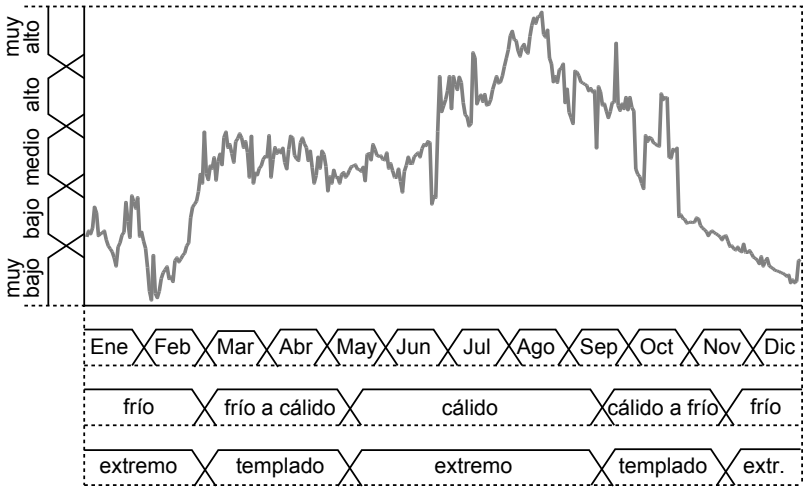


Figura 4.4: Flujo de pacientes masculinos al centro de salud C_A durante un año.

Primera estrategia

Una vez presentado el problema y su contexto, así como los parámetros que guiarán la exploración, pasemos a ilustrar cada uno de los pasos de dicha exploración que se producen al hacer uso de la primera estrategia para la resolución. A lo largo de la sección se presentarán tanto tablas como gráficas que mostrarán las distintas fases por las que va pasando el algoritmo.

A continuación, se muestra una aclaración que nos ayudará a comprender mejor lo que se intenta mostrar en las distintas tablas de evolución:

- p es el índice del cuantificador en la familia de cuantificadores coherentes.
- k es el número de etiquetas E_i agregadas en la sentencia.
- Q es la expresión lingüística para el cuantificador Q_p.
- A es la expresión lingüística para el valor de la variable en la sentencia.
- El resto de las columnas se corresponden con los periodos de tiempo del nivel. Los números almacenados en cada celda representan los grados de cumplimiento de la sentencia correspondiente formada por el cuantificador, la agrupación de etiquetas y el periodo de tiempo indicado. Los números en negrita se usan para resaltar las sentencias seleccionadas por el algoritmo para añadirlas al resumen final. Las celdas vacías representan sentencias que no han sido exploradas por el algoritmo debido a la partición jerárquica del dominio temporal.

Con respecto a las figuras insertadas en las tablas, se marcará mediante un sombreado en la zona del gráfico la intersección entre el soporte del periodo analizado y el del término utilizado para tratar de describirlo. Debemos decir que aunque en ambos casos la definición se hace mediante etiquetas lingüísticas con transiciones graduales esta circunstancia no se ha reflejado en el sombreado para favorecer la simplicidad de la representación. Si se hubiera tenido en cuenta debería quedar reflejado mediante la degradación de color en los bordes.

Como se ha descrito en la Sección 4.2.1, el Algoritmo 1 explora todas las posibles uniones de etiquetas E_i para obtener el término A^{TS} antes de explorar soluciones que hagan uso de un cuantificador menos estricto, comenzando por el más estricto de entre los que aporta el usuario. Esto es, busca sentencias con cuantificadores tan estrictos como sea posible siempre igualando o superando el umbral τ.

La exploración del algoritmo comienza por los periodos de tiempo con granularidad menos fina, es decir aquellos de L_1, de modo que:

$$ParaResumir = \{clima\ extremo,\ clima\ templado\}.$$

Se toma el primer periodo que vamos a resumir, es decir ***clima extremo***, se fija el cuantificador más estricto, *La mayoría*, y se intenta con todas las combinaciones de los términos que describen la variable en un orden creciente de cardinalidad. En las Tablas 4.4, 4.5, 4.6, 4.7 y 4.8 se muestra el proceso de evaluación de las sentencias producidas de la evaluación del periodo *clima extremo* usando el cuantificador *La mayoría* y las etiquetas simples en A. Como se puede apreciar en la Figura 4.8, no se ha obtenido ninguna combinación que de un buen resultado, de modo que se pasa a probar con combinaciones de etiquetas. Las Tablas 4.9 y 4.10 representan una pequeña muestra de esta misma evaluación pero usando parejas de etiquetas.

Se prueba con todas las parejas de etiquetas, es decir, se alcanza el límite $Glim_1$ del nivel, y a pesar de ello no se ha encontrado una sentencia que supere o iguale a τ, de modo que se repite todo el proceso con un cuantificador menos estricto, en este caso, *Aprox. más del 70 %*. El resultado de dichas evaluaciones podemos verlo en la Tabla 4.11.

Como podemos ver en la Tabla 4.11, se ha probado con todos los cuantificadores posibles ($Qlim_1$) y no se ha encontrado ninguna sentencia satisfactoria; se desecha el periodo *clima extremo* y se añaden a $ParaResumir$ sus hijos en el nivel inferior de la jerarquía, es decir con periodos con granularidad más fina. De modo que temenos

$$ParaResumir = \{clima\ templado,\ clima\ frío,\ clima\ cálido\}.$$

p	k	Q	A	clima extremo	clima templado
3	1	La mayoría	muy bajo bajo medio alto muy alto	0	
	2		muy bajo o bajo muy bajo o medio muy bajo o alto muy bajo o muy alto bajo o medio bajo o alto bajo o muy alto medio o alto medio o muy alto alto o muy alto		
2	1	Aprox. más del 70 %	muy bajo bajo medio alto muy alto		
	2		muy bajo o bajo muy bajo o medio muy bajo o alto muy bajo o muy alto bajo o medio bajo o alto bajo o muy alto medio o alto medio o muy alto alto o muy alto		

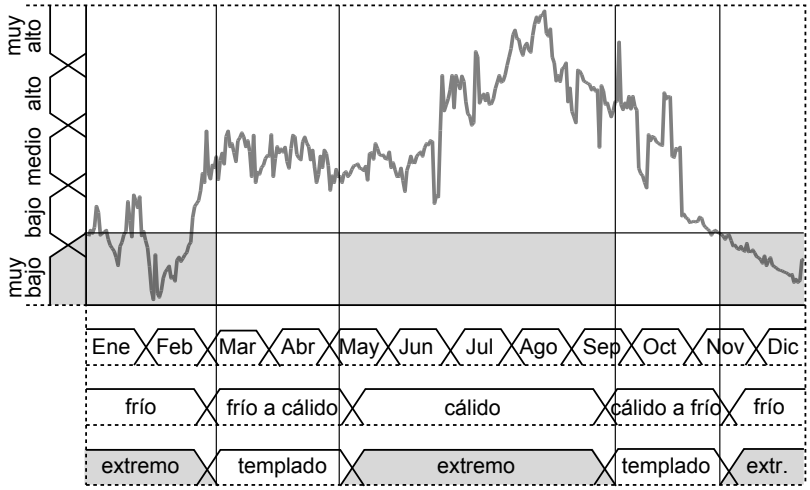


Tabla 4.4: Exploración del Algoritmo 1 para el problema del centro de salud C_A. En este paso se explora el primer periodo *clima extremo* (del nivel L_1) con la combinación *La mayoría* y *muy bajo*. Como se aprecia en la figura correspondiente existen puntos de la secuencia para los que la combinación se cumple, pero al mismo tiempo existen otros muchos para los que no es verdadera. El resultado de la evaluación de la sentencia cuantificada correspondiente es el valor 0, por lo tanto podemos asegurar que la sentencia no se encontrará en el resumen final.

p	k	Q	A	clima extremo	clima templado
3	1	La mayoría	muy bajo bajo medio alto muy alto	0 0	
	2		muy bajo o bajo muy bajo o medio muy bajo o alto muy bajo o muy alto bajo o medio bajo o alto bajo o muy alto medio o alto medio o muy alto alto o muy alto		
2	1	Aprox. más del 70 %	muy bajo bajo medio alto muy alto		
	2		muy bajo o bajo muy bajo o medio muy bajo o alto muy bajo o muy alto bajo o medio bajo o alto bajo o muy alto medio o alto medio o muy alto alto o muy alto		

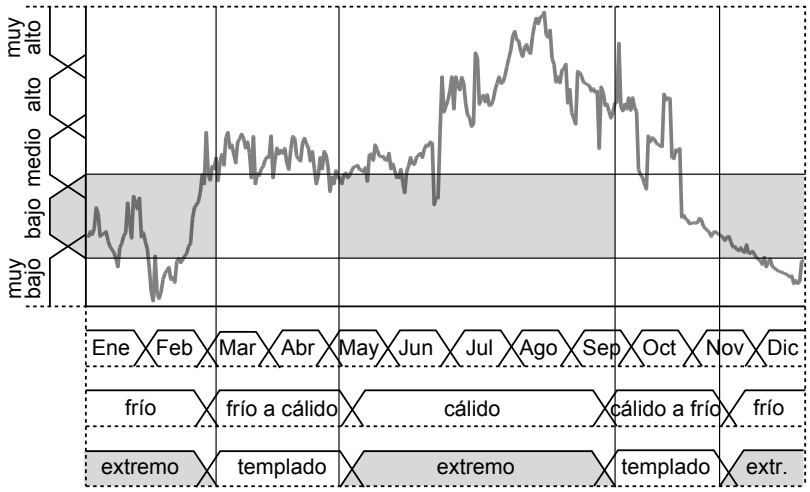


Tabla 4.5: Exploración del Algoritmo 1 para el problema del centro de salud C_A. En este paso se explora el primer periodo *clima extremo* con la combinación *La mayoría* y *bajo*. En esta ocasión el resultado de la evaluación de la sentencia cuantificada correspondiente es de nuevo 0.

p	k	Q	A	clima extremo	clima templado
3	1	La mayoría	muy bajo bajo medio alto muy alto	0 0 0	
	2		muy bajo o bajo muy bajo o medio muy bajo o alto muy bajo o muy alto bajo o medio bajo o alto bajo o muy alto medio o alto medio o muy alto alto o muy alto		
2	1	Aprox. más del 70 %	muy bajo bajo medio alto muy alto		
	2		muy bajo o bajo muy bajo o medio muy bajo o alto muy bajo o muy alto bajo o medio bajo o alto bajo o muy alto medio o alto medio o muy alto alto o muy alto		

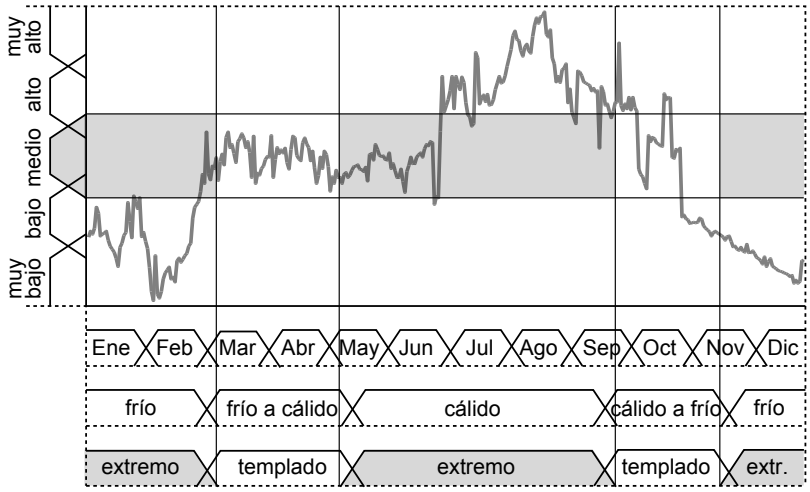


Tabla 4.6: Exploración del Algoritmo 1 para el problema del centro de salud C_A. En este paso se explora el primer periodo *clima extremo* con la combinación *La mayoría* y *medio*. Como resultado para la evaluación de la correspondiente sentencia, es decir *"La mayoría de los días en clima extremo, el flujo de pacientes es medio"* obtenemos un 0.

p	k	Q	A	clima extremo	clima templado
3	1	La mayoría	muy bajo bajo medio alto muy alto	0 0 0 0	
	2		muy bajo o bajo muy bajo o medio muy bajo o alto muy bajo o muy alto bajo o medio bajo o alto bajo o muy alto medio o alto medio o muy alto alto o muy alto		
2	1	Aprox. más del 70 %	muy bajo bajo medio alto muy alto		
	2		muy bajo o bajo muy bajo o medio muy bajo o alto muy bajo o muy alto bajo o medio bajo o alto bajo o muy alto medio o alto medio o muy alto alto o muy alto		

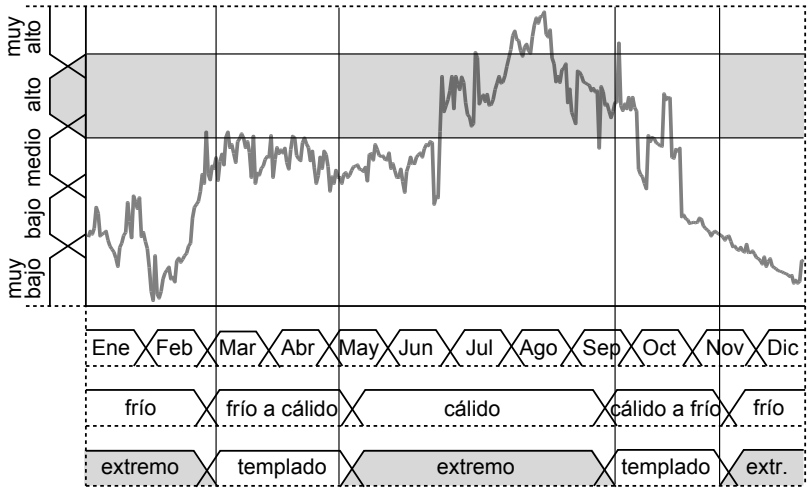


Tabla 4.7: Exploración del Algoritmo 1 para el problema del centro de salud C_A. En este paso se explora el primer periodo *clima extremo* con la combinación *La mayoría* y *alto*. Como resultado para la evaluación de la correspondiente sentencia, es decir *"La mayoría de los días en clima extremo, el flujo de pacientes es alto"* obtenemos un 0.

p	k	Q	A	clima extremo	clima templado
3	1	La mayoría	muy bajo	0	
			bajo	0	
			medio	0	
			alto	0	
			muy alto	0	
	2		muy bajo o bajo		
			muy bajo o medio		
			muy bajo o alto		
			muy bajo o muy alto		
			bajo o medio		
			bajo o alto		
			bajo o muy alto		
			medio o alto		
			medio o muy alto		
			alto o muy alto		
2	1	Aprox. más del 70 %	muy bajo		
			bajo		
			medio		
			alto		
			muy alto		
	2		muy bajo o bajo		
			muy bajo o medio		
			muy bajo o alto		
			muy bajo o muy alto		
			bajo o medio		
			bajo o alto		
			bajo o muy alto		
			medio o alto		
			medio o muy alto		
			alto o muy alto		

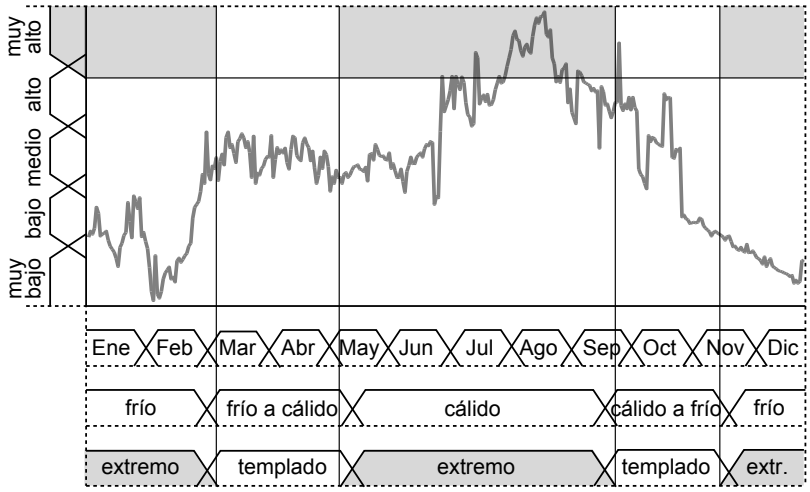


Tabla 4.8: Exploración del Algoritmo 1 para el problema del centro de salud C_A. En este paso se explora el primer periodo *clima extremo* con la combinación *La mayoría* y *muy alto*. Como resultado para la evaluación de la correspondiente sentencia, es decir *"La mayoría de los días en clima extremo, el flujo de pacientes es muy alto"* obtenemos un 0.

p	k	Q	A	clima extremo	clima templado
3	1	La mayoría	muy bajo	0	
			bajo	0	
			medio	0	
			alto	0	
			muy alto	0	
	2		muy bajo o bajo	0	
			muy bajo o medio		
			muy bajo o alto		
			muy bajo o muy alto		
			bajo o medio		
			bajo o alto		
			bajo o muy alto		
			medio o alto		
			medio o muy alto		
			alto o muy alto		
2	1	Aprox. más del 70 %	muy bajo		
			bajo		
			medio		
			alto		
			muy alto		
	2		muy bajo o bajo		
			muy bajo o medio		
			muy bajo o alto		
			muy bajo o muy alto		
			bajo o medio		
			bajo o alto		
			bajo o muy alto		
			medio o alto		
			medio o muy alto		
			alto o muy alto		

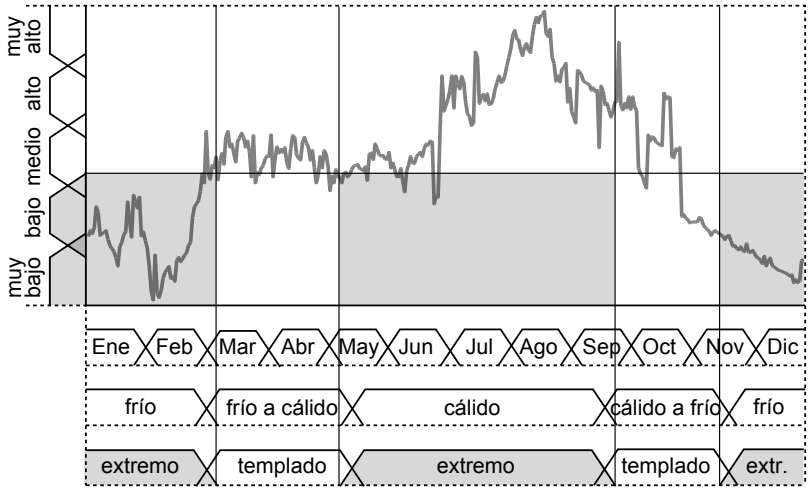


Tabla 4.9: Exploración del Algoritmo 1 para el problema del centro de salud C_A. En este paso se explora el primer periodo *clima extremo* con la combinación *La mayoría* y *muy bajo o bajo*. Como resultado para la evaluación de la correspondiente sentencia, es decir *"La mayoría de los días en clima extremo, el flujo de pacientes es muy bajo o bajo"* obtenemos un 0.

p	k	Q	A	clima extremo	clima templado
3	1	La mayoría	muy bajo	0	
			bajo	0	
			medio	0	
			alto	0	
			muy alto	0	
	2		muy bajo o bajo	0	
			muy bajo o medio	0	
			muy bajo o alto		
			muy bajo o muy alto		
			bajo o medio		
			bajo o alto		
			bajo o muy alto		
			medio o alto		
			medio o muy alto		
			alto o muy alto		
2	1	Aprox. más del 70 %	muy bajo		
			bajo		
			medio		
			alto		
			muy alto		
	2		muy bajo o bajo		
			muy bajo o medio		
			muy bajo o alto		
			muy bajo o muy alto		
			bajo o medio		
			bajo o alto		
			bajo o muy alto		
			medio o alto		
			medio o muy alto		
			alto o muy alto		

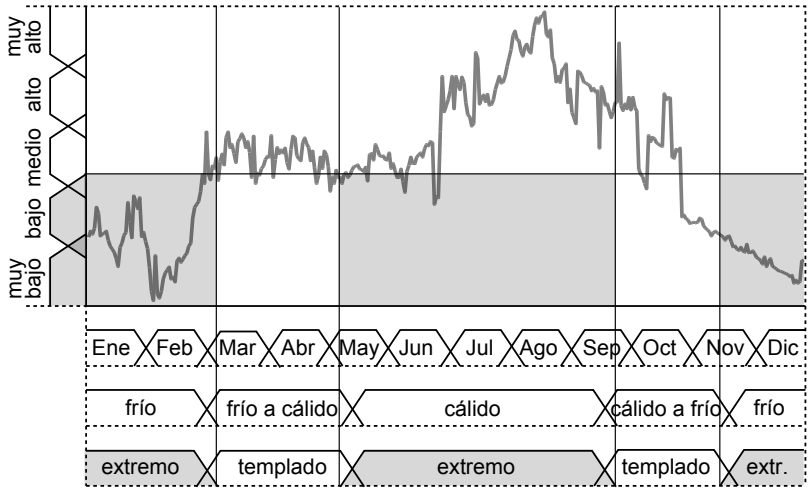


Tabla 4.10: Exploración del Algoritmo 1 para el problema del centro de salud C_A. En este paso se explora el primer periodo *clima extremo* con la combinación *La mayoría* y *muy bajo o medio*. Como resultado para la evaluación de la correspondiente sentencia, es decir *"La mayoría de los días en clima extremo, el flujo de pacientes es muy bajo o medio"* obtenemos un 0.

p	k	Q	A	clima extremo	clima templado
3	1	La mayoría	muy bajo	0	
			bajo	0	
			medio	0	
			alto	0	
			muy alto	0	
	2		muy bajo o bajo	0	
			muy bajo o medio	0	
			muy bajo o alto	0	
			muy bajo o muy alto	0	
			bajo o medio	0	
			bajo o alto	0	
			bajo o muy alto	0	
			medio o alto	0	
			medio o muy alto	0	
			alto o muy alto	0	
2	1	Aprox. más del 70 %	muy bajo	0	
			bajo	0	
			medio	0	
			alto	0	
			muy alto	0	
	2		muy bajo o bajo	0	
			muy bajo o medio	0	
			muy bajo o alto	0	
			muy bajo o muy alto	0	
			bajo o medio	0	
			bajo o alto	0	
			bajo o muy alto	0	
			medio o alto	0	
			medio o muy alto	0	
			alto o muy alto	0	

Tabla 4.11: Exploración del Algoritmo 1 para el problema del centro de salud C_A. Vemos como después de realizar el estudio de todas las combinaciones posibles no se ha generado ninguna sentencia que sea válida en relación con el umbral τ. Esto provoca la inserción en la cola de exploración de los hijos de *clima extremo*, es decir, las etiquetas *clima frío* y *clima cálido* que serán analizadas en pasos posteriores.

Aunque los periodos de transición *clima frío a cálido* y *clima cálido a frío* también se consideran etiquetas hijas de *clima extremo* (ya que su intersección no es vacía) éstas no han sido incluidas en la cola ya que son generalizadas por *clima templado* que todavía no ha sido analizada.

Ya en el siguiente paso, se explora el periodo ***clima templado*** que es el siguiente en la cola *ParaResumir*. Como en el caso anterior probamos en primer lugar con el cuantificador más estricto y etiquetas simples (Tablas 4.12, 4.13, 4.14, 4.15 y 4.16) y, también esta vez con combinaciones de etiquetas (algunos ejemplo en Tablas 4.17 y 4.18). Como no se ha tenido éxito se repite de nuevo el proceso para el siguiente cuantificador disponible, es decir *Aprox. más del 70 %*.

Afortunadamente, una de las combinaciones con $p = 2$ y $k = 2$ genera una sentencia con un grado de cumplimiento mayor que τ, que es añadida al resumen final. Si éste no hubiera sido el caso, deberían añadirse a la cola los hijos de la etiqueta, pero el proceso no se lleva a cabo ya que los hijos ya se encuentran en la cola desde

p	k	Q	A	clima extremo	clima templado
3	1	La mayoría	muy bajo	0	0
			bajo	0	
			medio	0	
			alto	0	
			muy alto	0	
	2		muy bajo o bajo	0	
			muy bajo o medio	0	
			muy bajo o alto	0	
			muy bajo o muy alto	0	
			bajo o medio	0	
			bajo o alto	0	
			bajo o muy alto	0	
			medio o alto	0	
			medio o muy alto	0	
			alto o muy alto	0	
2	1	Aprox. más del 70 %	muy bajo	0	
			bajo	0	
			medio	0	
			alto	0	
			muy alto	0	
	2		muy bajo o bajo	0	
			muy bajo o medio	0	
			muy bajo o alto	0	
			muy bajo o muy alto	0	
			bajo o medio	0	
			bajo o alto	0	
			bajo o muy alto	0	
			medio o alto	0	
			medio o muy alto	0	
			alto o muy alto	0	

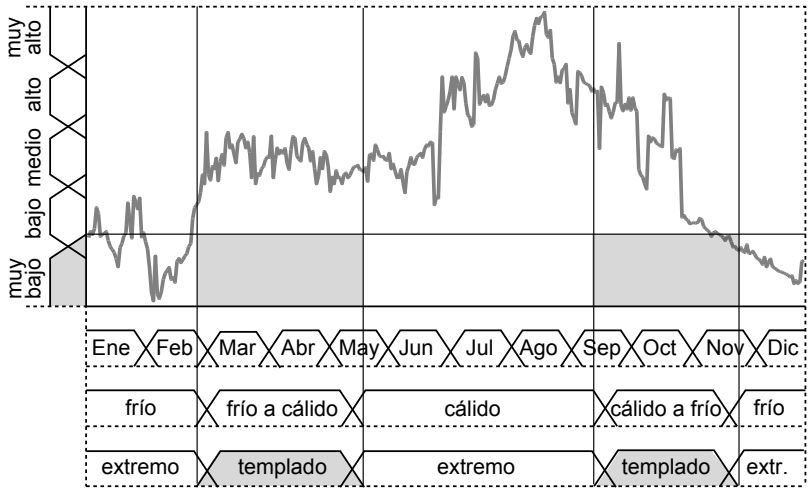


Tabla 4.12: Exploración del Algoritmo 1 para el problema del centro de salud C_A. En esta ocasión se analiza la combinación *clima templado* con el cuantificador *La mayoría* y la descripción *muy bajo*. El resultado no es satisfactorio de modo que la sentencia no aparecerá en el resumen final.

p	k	Q	A	clima extremo	clima templado
3	1	La mayoría	muy bajo	0	0
			bajo	0	0
			medio	0	
			alto	0	
			muy alto	0	
	2		muy bajo o bajo	0	
			muy bajo o medio	0	
			muy bajo o alto	0	
			muy bajo o muy alto	0	
			bajo o medio	0	
			bajo o alto	0	
			bajo o muy alto	0	
			medio o alto	0	
			medio o muy alto	0	
			alto o muy alto	0	
2	1	Aprox. más del 70 %	muy bajo	0	
			bajo	0	
			medio	0	
			alto	0	
			muy alto	0	
	2		muy bajo o bajo	0	
			muy bajo o medio	0	
			muy bajo o alto	0	
			muy bajo o muy alto	0	
			bajo o medio	0	
			bajo o alto	0	
			bajo o muy alto	0	
			medio o alto	0	
			medio o muy alto	0	
			alto o muy alto	0	

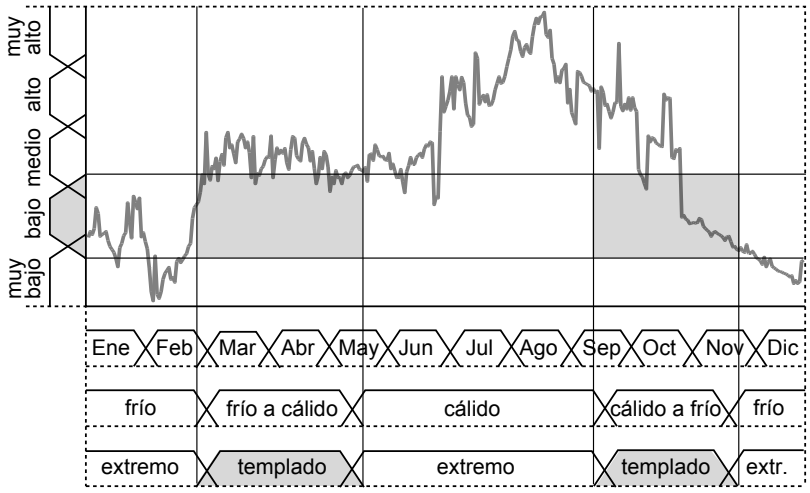


Tabla 4.13: Exploración del Algoritmo 1 para el problema del centro de salud C_A. Se analiza la combinación *clima templado* con el cuantificador *La mayoría* y la descripción *bajo*. De nuevo el resultado no es satisfactorio.

p	k	Q	A	clima extremo	clima templado
3	1	La mayoría	muy bajo	0	0
			bajo	0	0
			medio	0	0
			alto	0	
			muy alto	0	
	2		muy bajo o bajo	0	
			muy bajo o medio	0	
			muy bajo o alto	0	
			muy bajo o muy alto	0	
			bajo o medio	0	
			bajo o alto	0	
			bajo o muy alto	0	
			medio o alto	0	
			medio o muy alto	0	
			alto o muy alto	0	
2	1	Aprox. más del 70 %	muy bajo	0	
			bajo	0	
			medio	0	
			alto	0	
			muy alto	0	
	2		muy bajo o bajo	0	
			muy bajo o medio	0	
			muy bajo o alto	0	
			muy bajo o muy alto	0	
			bajo o medio	0	
			bajo o alto	0	
			bajo o muy alto	0	
			medio o alto	0	
			medio o muy alto	0	
			alto o muy alto	0	

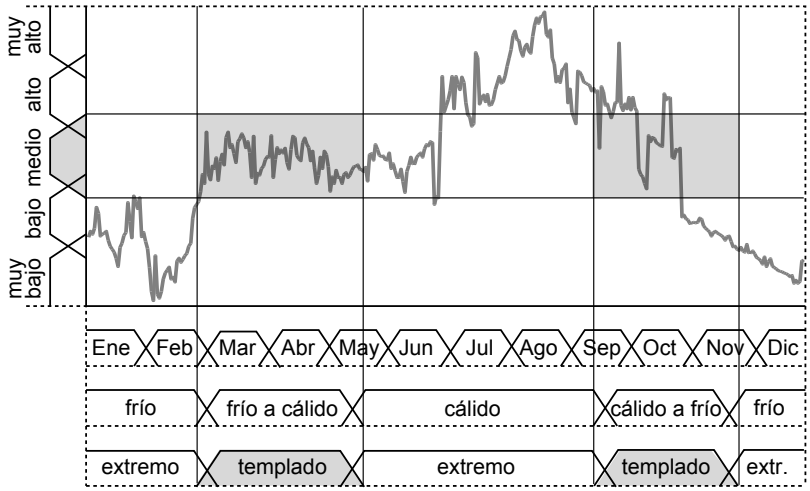


Tabla 4.14: Exploración del Algoritmo 1 para el problema del centro de salud C_A. Se analiza la combinación *clima templado* con el cuantificador *La mayoría* y la descripción *medio*, que da lugar a la sentencia cuantificada *La mayoría de los días en clima templado, el flujo de pacientes es medio*. Vemos en la figura que la sentencia no describe bien los puntos involucrados de modo que la sentencia no aparecerá en el resumen final.

p	k	Q	A	clima extremo	clima templado
3	1	La mayoría	muy bajo	0	0
			bajo	0	0
			medio	0	0
			alto	0	0
			muy alto	0	
	2		muy bajo o bajo	0	
			muy bajo o medio	0	
			muy bajo o alto	0	
			muy bajo o muy alto	0	
			bajo o medio	0	
			bajo o alto	0	
			bajo o muy alto	0	
			medio o alto	0	
			medio o muy alto	0	
			alto o muy alto	0	
2	1	Aprox. más del 70 %	muy bajo	0	
			bajo	0	
			medio	0	
			alto	0	
			muy alto	0	
	2		muy bajo o bajo	0	
			muy bajo o medio	0	
			muy bajo o alto	0	
			muy bajo o muy alto	0	
			bajo o medio	0	
			bajo o alto	0	
			bajo o muy alto	0	
			medio o alto	0	
			medio o muy alto	0	
			alto o muy alto	0	

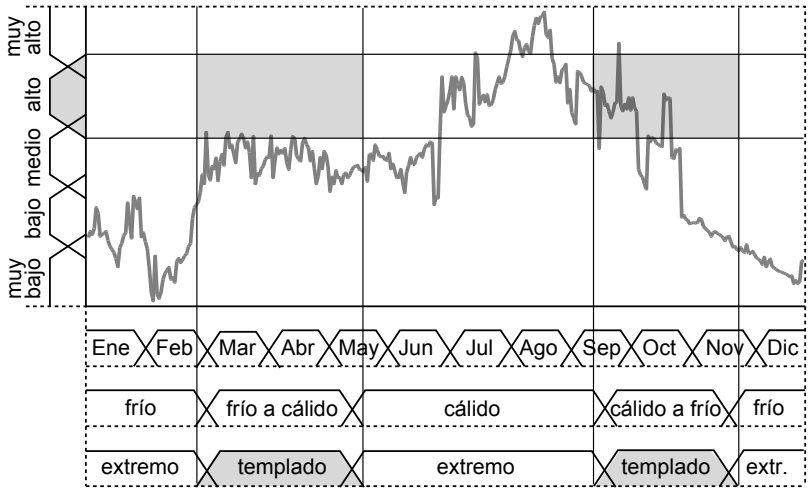


Tabla 4.15: Exploración del Algoritmo 1 para el problema del centro de salud C_A. El análisis de la sentencia cuantificada *La mayoría de los días en clima templado, el flujo de pacientes es alto* da como resultado un 0.

p	k	Q	A	clima extremo	clima templado
3	1	La mayoría	muy bajo	0	0
			bajo	0	0
			medio	0	0
			alto	0	0
			muy alto	0	0
	2		muy bajo o bajo	0	
			muy bajo o medio	0	
			muy bajo o alto	0	
			muy bajo o muy alto	0	
			bajo o medio	0	
			bajo o alto	0	
			bajo o muy alto	0	
			medio o alto	0	
			medio o muy alto	0	
			alto o muy alto	0	
2	1	Aprox. más del 70 %	muy bajo	0	
			bajo	0	
			medio	0	
			alto	0	
			muy alto	0	
	2		muy bajo o bajo	0	
			muy bajo o medio	0	
			muy bajo o alto	0	
			muy bajo o muy alto	0	
			bajo o medio	0	
			bajo o alto	0	
			bajo o muy alto	0	
			medio o alto	0	
			medio o muy alto	0	
			alto o muy alto	0	

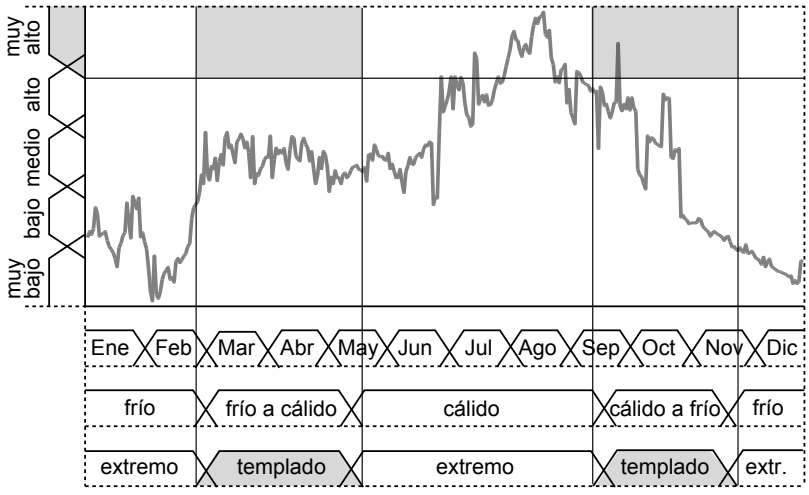


Tabla 4.16: Exploración del Algoritmo 1 para el problema del centro de salud C_A. Por último, el análisis de la combinación expuesta no supera el umbral. En el siguiente paso, se deberá probar con combinaciones de etiquetas para la descripción.

p	k	Q	A	clima extremo	clima templado
3	1	La mayoría	muy bajo	0	0
			bajo	0	0
			medio	0	0
			alto	0	0
			muy alto	0	0
	2		muy bajo o bajo	0	0
			muy bajo o medio	0	
			muy bajo o alto	0	
			muy bajo o muy alto	0	
			bajo o medio	0	
			bajo o alto	0	
			bajo o muy alto	0	
			medio o alto	0	
			medio o muy alto	0	
			alto o muy alto	0	
2	1	Aprox. más del 70 %	muy bajo	0	
			bajo	0	
			medio	0	
			alto	0	
			muy alto	0	
	2		muy bajo o bajo	0	
			muy bajo o medio	0	
			muy bajo o alto	0	
			muy bajo o muy alto	0	
			bajo o medio	0	
			bajo o alto	0	
			bajo o muy alto	0	
			medio o alto	0	
			medio o muy alto	0	
			alto o muy alto	0	

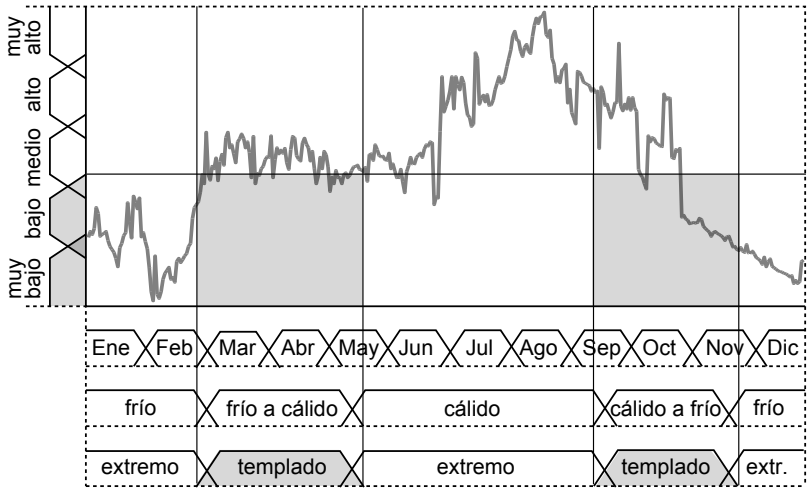


Tabla 4.17: Exploración del Algoritmo 1 para el problema del centro de salud C_A. La combinación *muy bajo o bajo* no da buen resultado en la descripción de los datos.

p	k	Q	A	clima extremo	clima templado
3	1	La mayoría	muy bajo	0	0
			bajo	0	0
			medio	0	0
			alto	0	0
			muy alto	0	0
	2		muy bajo o bajo	0	0
			muy bajo o medio	0	0
			muy bajo o alto	0	
			muy bajo o muy alto	0	
			bajo o medio	0	
			bajo o alto	0	
			bajo o muy alto	0	
			medio o alto	0	
			medio o muy alto	0	
			alto o muy alto	0	
2	1	Aprox. más del 70 %	muy bajo	0	
			bajo	0	
			medio	0	
			alto	0	
			muy alto	0	
	2		muy bajo o bajo	0	
			muy bajo o medio	0	
			muy bajo o alto	0	
			muy bajo o muy alto	0	
			bajo o medio	0	
			bajo o alto	0	
			bajo o muy alto	0	
			medio o alto	0	
			medio o muy alto	0	
			alto o muy alto	0	

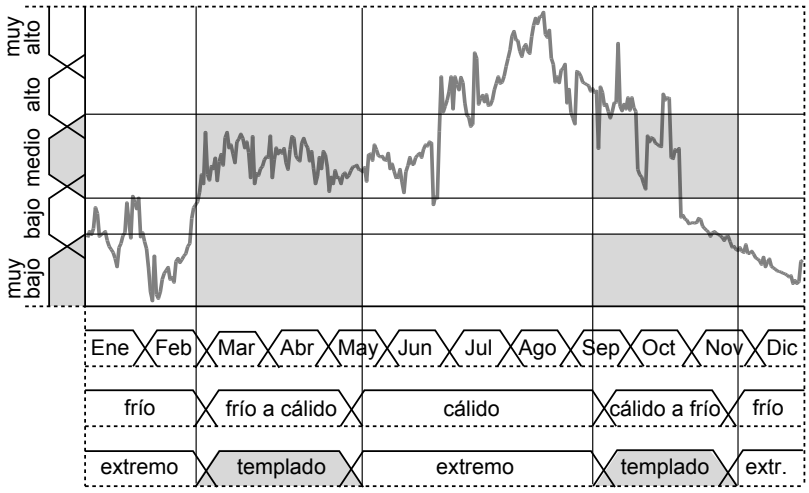


Tabla 4.18: Exploración del Algoritmo 1 para el problema del centro de salud C_A. Tampoco la disyunción entre *muy bajo* y *medio* ofrece buenos resultados para describir el periodo *clima templado* con el cuantificador *La mayoría*.

el paso anterior. La Tabla 4.19 muestra los grados de cumplimiento de las sentencias cuantificadas exploradas por el algoritmo en este nivel de la jerarquía temporal. Además, gracias a la imagen adjunta podemos comprobar que *"Aproximadamente más del 70 % de los días en clima templado, el flujo de pacientes es alto o medio"* y que además lo hace con grado de cumplimiento igual a 1.

p	k	Q	A	clima extremo	clima templado
3	1	La mayoría	muy bajo	0	0
			bajo	0	0
			medio	0	0
			alto	0	0
			muy alto	0	0
	2		muy bajo o bajo	0	0
			muy bajo o medio	0	0
			muy bajo o alto	0	0
			muy bajo o muy alto	0	0
			bajo o medio	0	0.56
			bajo o alto	0	0
			bajo o muy alto	0	0
			medio o alto	0	0.68
			medio o muy alto	0	0
			alto o muy alto	0	0
2	1	Aprox. más del 70 %	muy bajo	0	0
			bajo	0	0
			medio	0	0.35
			alto	0	0
			muy alto	0	0
	2		muy bajo o bajo	0	0
			muy bajo o medio	0	0.35
			muy bajo o alto	0	0
			muy bajo o muy alto	0	0
			bajo o medio	0	0.98
			bajo o alto	0	0
			bajo o muy alto	0	0
			medio o alto	0	**1**
			medio o muy alto	0	0.39
			alto o muy alto	0	0

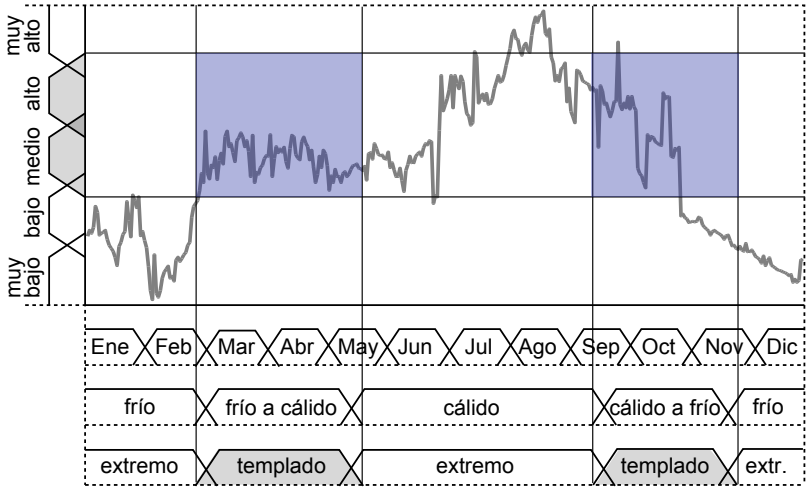


Tabla 4.19: Exploración del Algoritmo 1 para el problema del centro de salud C_A. Una vez finalizada la exploración del nivel L_1 obtenemos una sentencia cuantificada que describe el periodo *clima templado* de la siguiente forma *"Aproximadamente más del 70 % de los días en clima templado, el flujo de pacientes es alto o medio"*. Como dijimos anteriormente no ha sido posible encontrar una sentencia que describa de forma adecuada el periodo *clima extremo*, debido a ello trataremos de describirlo a través del análisis de sus hijos en el nivel L_2.

En este momento la cola contiene ya únicamente etiquetas del nivel L_2 como vemos en:

$$ParaResumir = \{clima\ frío,\ clima\ cálido\}.$$

El algoritmo continúa con la siguiente etiqueta de la cola, ***clima frío***. En primer lugar, se realizará la evaluación con el cuantificador más estricto y las etiquetas sin agrupar. Al igual que en los casos anteriores, podemos ver la evolución a través de las Tablas 4.20, 4.21, 4.22, 4.23 y 4.24.

Como la exploración ha finalizado sin resultados adecuados, se procede a probar con las disyunciones de las etiquetas. En esta ocasión si se encuentra una sentencia que supera el umbral de cumplimiento, de modo que, como consecuencia, se añade una nueva sentencia al resumen final, se detiene la exploración de esta etiqueta y se continúa con la siguiente en la cola. En la Tabla 4.25 se muestran los resultados de las evaluaciones de todas las parejas posibles y se muestra la figura de la combinación que ofrece la combinación que supera el umbral de cumplimiento.

En estos momentos la cola tendría el siguiente aspecto:

$$ParaResumir = \{clima\ cálido\}.$$

De modo que el algoritmo continúa con la etiqueta ***clima cálido***. En esta ocasión no vamos a repetir el proceso paso a paso. La Tabla 4.26 muestra los valores de todas las combinaciones exploradas para la etiqueta mencionada así como la visión global del comportamiento del algoritmo para el análisis de las etiquetas del nivel L_2 de la jerarquía.

Finalmente, el algoritmo analiza las etiquetas del nivel L_3 de la jerarquía temporal (el de granularidad más fina) que se encuentran en $ParaResumir$. En este caso, todos los periodos de tiempo son descritos con sentencias del primer cuantificador explorado, *La mayoría* (ver Tabla 4.27).

Como vemos el análisis del periodo *clima cálido* finaliza sin una sentencia adecuada que lo describa con el grado de cumplimiento requerido, de modo que el algoritmo debe seguir profundizando en la jerarquía explorando los hijos del periodo, de modo que tenemos:

$$ParaResumir = \{mayo,\ junio,\ julio,\ agosto,\ septiembre\}.$$

p	k	Q	A	frío	frío a cálido	cálido	cálido a frío
3	1	La mayoría	muy bajo bajo medio alto muy alto	0			
	2		muy bajo o bajo muy bajo o medio muy bajo o alto muy bajo o muy alto bajo o medio bajo o alto bajo o muy alto medio o alto medio o muy alto alto o muy alto				
2	1	Aprox. más del 70 %	muy bajo bajo medio alto muy alto				
	2		muy bajo o bajo muy bajo o medio muy bajo o alto muy bajo o muy alto bajo o medio bajo o alto bajo o muy alto medio o alto medio o muy alto alto o muy alto				

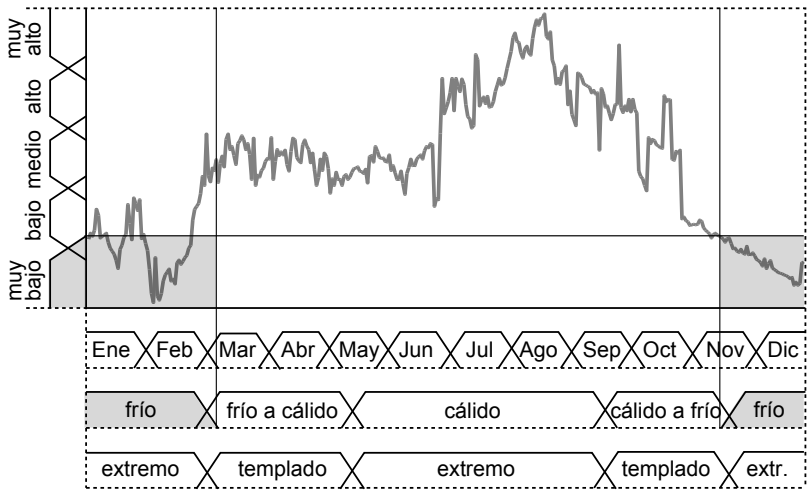


Tabla 4.20: Exploración del Algoritmo 1 para el problema del centro de salud C_A. Se comienza el análisis de los periodos del nivel L_2 con la etiqueta *clima frío*. Se usa el cuantificador más estricto *La mayoría* y la etiqueta *muy bajo*. El nivel de cumplimiento de la sentencia generada para el conjunto de datos que poseemos es 0.

p	k	Q	A	frío	frío a cálido	cálido	cálido a frío
3	1	La mayoría	muy bajo bajo medio alto muy alto	0 0			
	2		muy bajo o bajo muy bajo o medio muy bajo o alto muy bajo o muy alto bajo o medio bajo o alto bajo o muy alto medio o alto medio o muy alto alto o muy alto				
2	1	Aprox. más del 70 %	muy bajo bajo medio alto muy alto				
	2		muy bajo o bajo muy bajo o medio muy bajo o alto muy bajo o muy alto bajo o medio bajo o alto bajo o muy alto medio o alto medio o muy alto alto o muy alto				

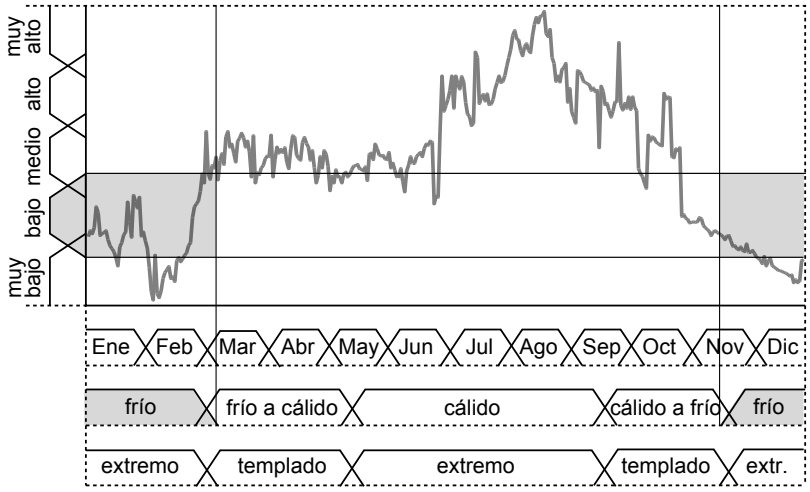


Tabla 4.21: Exploración del Algoritmo 1 para el problema del centro de salud C_A. Se continúa la exploración esta vez con la etiqueta *bajo*. De nuevo el resultado no es bueno.

p	k	Q	A	frío	frío a cálido	cálido	cálido a frío
3	1	La mayoría	muy bajo bajo medio alto muy alto	0 0 0			
	2		muy bajo o bajo muy bajo o medio muy bajo o alto muy bajo o muy alto bajo o medio bajo o alto bajo o muy alto medio o alto medio o muy alto alto o muy alto				
2	1	Aprox. más del 70 %	muy bajo bajo medio alto muy alto				
	2		muy bajo o bajo muy bajo o medio muy bajo o alto muy bajo o muy alto bajo o medio bajo o alto bajo o muy alto medio o alto medio o muy alto alto o muy alto				

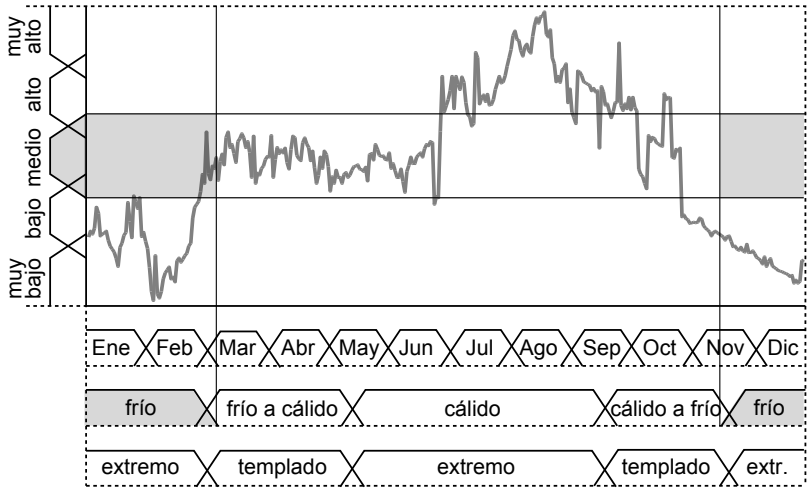


Tabla 4.22: Exploración del Algoritmo 1 para el problema del centro de salud C_A. La sentencia *"La mayoría de los días de clima frío, el flujo de pacientes es medio"* tiene un grado de cumplimiento igual a 0.

p	k	Q	A	frío	frío a cálido	cálido	cálido a frío
3	1	La mayoría	muy bajo bajo medio alto muy alto	0 0 0 0			
	2		muy bajo o bajo muy bajo o medio muy bajo o alto muy bajo o muy alto bajo o medio bajo o alto bajo o muy alto medio o alto medio o muy alto alto o muy alto				
2	1	Aprox. más del 70 %	muy bajo bajo medio alto muy alto				
	2		muy bajo o bajo muy bajo o medio muy bajo o alto muy bajo o muy alto bajo o medio bajo o alto bajo o muy alto medio o alto medio o muy alto alto o muy alto				

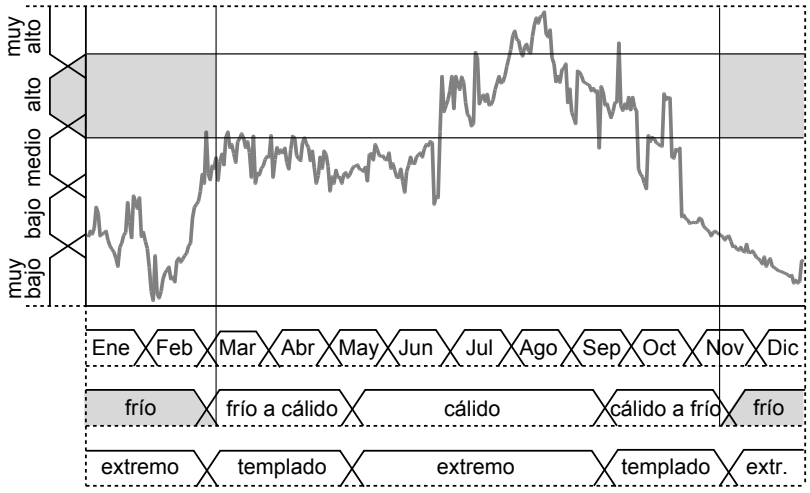


Tabla 4.23: Exploración del Algoritmo 1 para el problema del centro de salud C_A. La tabla muestra el resultado de la evaluación de la sentencia cuantificada *"La mayoría de los días de clima frío, el flujo de pacientes es alto"*.

p	k	Q	A	frío	frío a cálido	cálido	cálido a frío
3	1	La mayoría	muy bajo bajo medio alto muy alto	0 0 0 0 0			
	2		muy bajo o bajo muy bajo o medio muy bajo o alto muy bajo o muy alto bajo o medio bajo o alto bajo o muy alto medio o alto medio o muy alto alto o muy alto				
2	1	Aprox. más del 70 %	muy bajo bajo medio alto muy alto				
	2		muy bajo o bajo muy bajo o medio muy bajo o alto muy bajo o muy alto bajo o medio bajo o alto bajo o muy alto medio o alto medio o muy alto alto o muy alto				

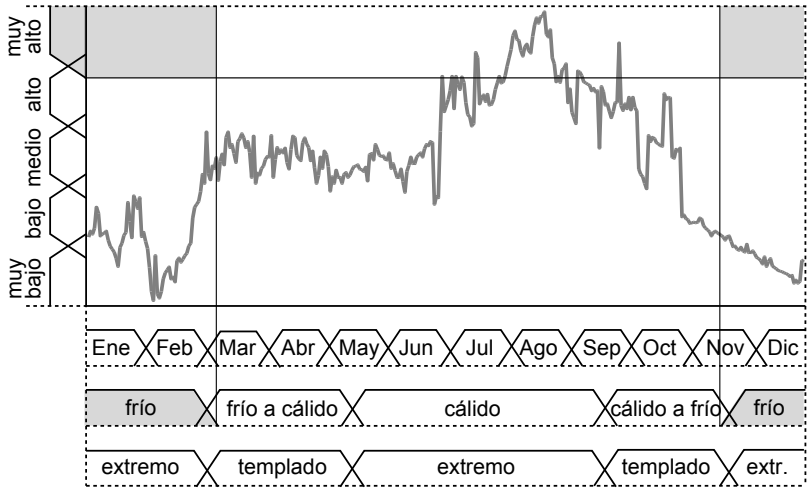


Tabla 4.24: Exploración del Algoritmo 1 para el problema del centro de salud C_A. La tabla muestra el resultado de la evaluación de la sentencia cuantificada *"La mayoría de los días de clima frío, el flujo de pacientes es muy alto"*.

p	k	Q	A	frío	frío a cálido	cálido	cálido a frío
3	1	La mayoría	muy bajo	0			
			bajo	0			
			medio	0			
			alto	0			
			muy alto	0			
	2		muy bajo o bajo	**0.89**			
			muy bajo o medio	0			
			muy bajo o alto	0			
			muy bajo o muy alto	0			
			bajo o medio	0			
			bajo o alto	0			
			bajo o muy alto	0			
			medio o alto	0			
			medio o muy alto	0			
			alto o muy alto	0			
2	1	Aprox. más del 70 %	muy bajo				
			bajo				
			medio				
			alto				
			muy alto				
	2		muy bajo o bajo				
			muy bajo o medio				
			muy bajo o alto				
			muy bajo o muy alto				
			bajo o medio				
			bajo o alto				
			bajo o muy alto				
			medio o alto				
			medio o muy alto				
			alto o muy alto				

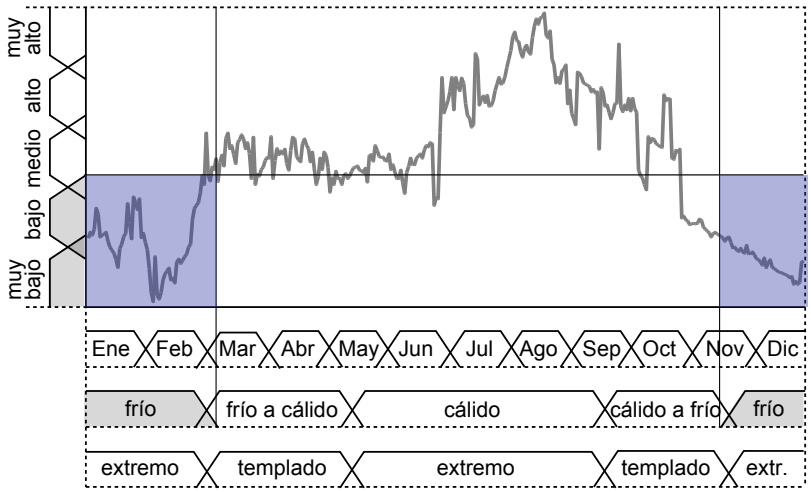


Tabla 4.25: Exploración del Algoritmo 1 para el problema del centro de salud C_A. De entre todas las parejas de etiquetas que se han probado con $p = 3$ y $k = 2$, vemos que la única que ofrece un buen resultado, que además superan el umbral establecido, es *muy bajo o bajo*. Como consecuencia se genera la sentencia *"La mayoría de los días en clima frío, el flujo de pacientes es muy bajo o bajo"* que se añade al resumen final.

p	k	Q	A	frío	frío a cálido	cálido	cálido a frío
3	1	La mayoría	muy bajo	0		0	
			bajo	0		0	
			medio	0		0	
			alto	0		0	
			muy alto	0		0	
	2		muy bajo o bajo	**0.89**		0	
			muy bajo o medio	0		0	
			muy bajo o alto	0		0	
			muy bajo o muy alto	0		0	
			bajo o medio	0		0	
			bajo o alto	0		0	
			bajo o muy alto	0		0	
			medio o alto	0		0.09	
			medio o muy alto	0		0	
			alto o muy alto	0		0	
2	1	Aprox. más del 70 %	muy bajo			0	
			bajo			0	
			medio			0	
			alto			0	
			muy alto			0	
	2		muy bajo o bajo			0	
			muy bajo o medio			0	
			muy bajo o alto			0	
			muy bajo o muy alto			0	
			bajo o medio			0	
			bajo o alto			0	
			bajo o muy alto			0	
			medio o alto			0.57	
			medio o muy alto			0	
			alto o muy alto			0.19	

Tabla 4.26: Exploración del Algoritmo 1 para el problema del centro de salud C_A. Representación de la exploración de soluciones para las etiquetas temporales del nivel L_2. Sólo se ha descrito con éxito el periodo *clima frío*, para describir el periodo *clima cálido* se deben analizar las etiquetas hijas que se encuentran en el nivel L_3.

p	k	Q	A	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
3	1	La mayoría	muy bajo					0	0	0	0	0			
			bajo					0	0	0	0	0			
			medio					**1**	**0.73**	0	0	0			
			alto					0	0	**1**	0	**0.83**			
			muy alto					0	0	0	0.61	0			
	2		muy bajo o bajo								0				
			muy bajo o medio								0				
			muy bajo o alto								0				
			muy bajo o muy alto								0.61				
			bajo o medio								0				
			bajo o alto								0				
			bajo o muy alto								0.61				
			medio o alto								0				
			medio o muy alto								0.61				
			alto o muy alto								**1**				
2	1	Aprox. más del 70 %	muy bajo												
			bajo												
			medio												
			alto												
			muy alto												
	2		muy bajo o bajo												
			muy bajo o medio												
			muy bajo o alto												
			muy bajo o muy alto												
			bajo o medio												
			bajo o alto												
			bajo o muy alto												
			medio o alto												
			medio o muy alto												
			alto o muy alto												

Tabla 4.27: Exploración del Algoritmo 1 para el problema del centro de salud C_A. Representación de la exploración de soluciones para las etiquetas temporales del nivel L_2. Como se ve, todos los periodos se han podido describir adecuadamente sin tener que utilizar para ello el cuantificador menos estricto.

La Figura 4.5 muestra sombreadas las porciones del gráfico en las que se explican convenientemente los puntos de la serie temporal una vez completado el proceso de exploración de la jerarquía. Como podemos apreciar los periodos de granularidad más fina presentes en la cola han sido descritos de forma satisfactoria, añadiendo cinco nuevas sentencias al resumen final.

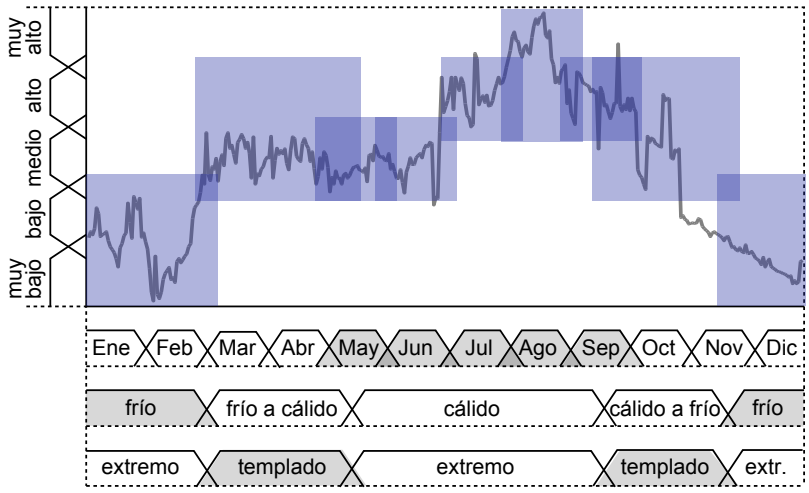


Figura 4.5: C_A: elecciones de descripción del Algoritmo 1.

Una vez finalizado el proceso, si recopilamos todas las sentencias que hemos ido obteniendo, tendremos el siguiente resumen:

"Aprox. más del 70 % de los días en clima templado, el flujo de pacientes es alto o medio (1)
La mayoría de los días en clima frío, el flujo de pacientes es muy bajo o bajo (0.89)
La mayoría de los días en Mayo, el flujo de pacientes es medio (1)
La mayoría de los días en Junio, el flujo de pacientes es medio (0.73)
La mayoría de los días en Julio, el flujo de pacientes es alto (1)
La mayoría de los días en Agosto, el flujo de pacientes es alto o muy alto (1)
La mayoría de los días en Septiembre, el flujo de pacientes es alto (0.83)"

Observamos que además de la sentencia obtenida, entre paréntesis se muestra el grado de cumplimiento de la misma. Después de pasar por la fase de post-proceso, el conjunto de sentencias se convertirá en un párrafo legible de la forma:

"Aprox. más del 70 % de los días en clima templado, el flujo de pacientes es alto o medio. La mayoría de los días en clima frío, es muy bajo o bajo; en Mayo y Junio, medio; en Julio y Septiembre, alto; y en Agosto, alto o muy alto".

Segunda estrategia

El enfoque que se sigue en esta segunda estrategia es diferente. El Algoritmo 2 explora todos los posibles cuantificadores antes de explorar soluciones que impliquen agrupaciones de etiquetas. De modo que busca sentencias con A^{TS} tan específicos como sea posible siempre que se supere o iguale el umbral τ.

En esta ocasión, y para no ser redundantes en ello, para la representación de la exploración nos serviremos únicamente de las tablas generadas y de una figura final donde se muestra el resultado.

La exploración de este algoritmo, al igual que en caso anterior, comienza en el nivel L_1 de la jerarquía temporal, de modo que tenemos $ParaResumir = \{$*clima extremo, clima templado*$\}$. Toma el primer periodo, *clima extremo*, y busca sentencias en las que A^{TS} sea una etiqueta simple. Se fija el cuantificador más estricto (*La mayoría*) y se prueba con todas las etiquetas simples de la partición de la variable. Si no hay éxito en el proceso, se exploran las etiquetas de nuevo pero esta vez con un cuantificador menos estricto. A pesar de ello no se encuentra una sentencia adecuada y se ha alcanzado el límite impuesto por $Qlim_1$, entonces el algoritmo repite el proceso pero esta vez agrupando etiquetas E_i. De nuevo no se tiene éxito. Como se ha llegado a $Glim_1$ y no se ha tenido éxito, el algoritmo introduce en la cola a los hijos del periodo. La cola queda de la siguiente forma $ParaResumir = \{$*clima templado, clima frío, clima cálido*$\}$.

A continuación, se explora la siguiente etiqueta en la cola, *clima templado*. En este caso, una de las combinaciones con $k = 2$ supera el umbral τ, y el estudio del periodo finaliza. La Tabla 4.28 muestra los grados de cumplimiento de las sentencias cuantificadas exploradas por el algoritmo en este nivel de la jerarquía temporal.

En el siguiente paso, el algoritmo se centra en la exploración de los periodos del nivel L_2 que todavía quedan en la cola $ParaResumir$.

Comienza por el periodo *clima frío*. La exploración de sentencias para dicho periodo tiene éxito con el cuantificador *La mayoría* y la agrupación *muy bajo o bajo* (ver Tabla 4.29).

En el último paso, el algoritmo analiza las etiquetas del primer nivel en la jerarquía que aún quedan en la cola $ParaResumir$. En este caso, todos los periodos de tiempo generan sentencias adecuadas con el cuantificador más estricto *La mayoría* excepto Agosto para el cual se necesita un cuantificador menos estricto (ver la Tabla 4.30).

p	k	Q	A	clima extremo	clima templado
1	3	La mayoría	muy bajo	0	0
			bajo	0	0
			medio	0	0
			alto	0	0
			muy alto	0	0
	2	Aprox. más del 70 %	muy bajo	0	0
			bajo	0	0
			medio	0	0.35
			alto	0	0
			muy alto	0	0
2	3	La mayoría	muy bajo o bajo	0	0
			muy bajo o medio	0	0
			muy bajo o alto	0	0
			muy bajo o muy alto	0	0
			bajo o medio	0	0.56
			bajo o alto	0	0
			bajo o muy alto	0	0
			medio o alto	0	0.68
			medio o muy alto	0	0
			alto o muy alto	0	0
	2	Aprox. más del 70 %	muy bajo o bajo	0	0
			muy bajo o medio	0	0.35
			muy bajo o alto	0	0
			muy bajo o muy alto	0	0
			bajo o medio	0	0.98
			bajo o alto	0	0
			bajo o muy alto	0	0
			medio o alto	0	**1**
			medio o muy alto	0	0.39
			alto o muy alto	0	0

Tabla 4.28: Exploración de las etiquetas del nivel L_1 por el Algoritmo 2 para el problema C_A.

k	p	Q	A	frío	frío a cálido	cálido	cálido a frío
1	3	La mayoría	muy bajo	0		0	
			bajo	0		0	
			medio	0		0	
			alto	0		0	
			muy alto	0		0	
	2	Aprox. más del 70 %	muy bajo	0.02		0	
			bajo			0	
			medio			0	
			alto			0	
			muy alto			0	
2	3	La mayoría	muy bajo o bajo	**0.89**		0	
			muy bajo o medio	0		0	
			muy bajo o alto	0		0	
			muy bajo o muy alto	0		0	
			bajo o medio	0		0	
			bajo o alto	0		0	
			bajo o muy alto	0		0	
			medio o alto	0		0.09	
			medio o muy alto	0		0	
			alto o muy alto	0		0	
	2	Aprox. más del 70 %	muy bajo o bajo			0	
			muy bajo o medio			0	
			muy bajo o alto			0	
			muy bajo o muy alto			0	
			bajo o medio			0	
			bajo o alto			0	
			bajo o muy alto			0	
			medio o alto			0.57	
			medio o muy alto			0	
			alto o muy alto			0.19	

Tabla 4.29: Exploración de las etiquetas del nivel L_2 por el Algoritmo 2 para el problema C_A.

p	k	Q	A	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
1	3	La mayoría	muy bajo					0	0	0	0	0			
			bajo					0	0	0	0	0			
			medio					**1**	**0.73**	0	0	0			
			alto					0	0	**1**	0	**0.83**			
			muy alto					0	0	0	0.61	0			
	2	Aprox. más del 70 %	muy bajo								0				
			bajo								0				
			medio								0				
			alto								0				
			muy alto								**1**				
2	3	La mayoría	muy bajo o bajo												
			muy bajo o medio												
			muy bajo o alto												
			muy bajo o muy alto												
			bajo o medio												
			bajo o alto												
			bajo o muy alto												
			medio o alto												
			medio o muy alto												
			alto o muy alto												
	2	Aprox. más del 70 %	muy bajo o bajo												
			muy bajo o medio												
			muy bajo o alto												
			muy bajo o muy alto												
			bajo o medio												
			bajo o alto												
			bajo o muy alto												
			medio o alto												
			medio o muy alto												
			alto o muy alto												

Tabla 4.30: Exploración de las etiquetas del nivel L_3 por el Algoritmo 2 para el problema C_A.

La Figura 4.6 muestra gráficamente el resultado de la exploración.

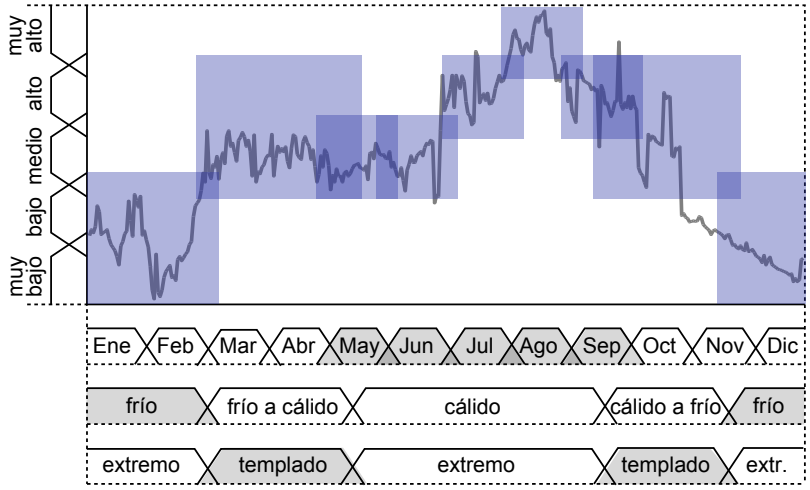


Figura 4.6: C_A: elecciones de descripción del Algoritmo 2.

Una vez finalizado el proceso, si recopilamos todas las sentencias que hemos ido obteniendo, tendremos el siguiente resumen:

> "Aprox. más del 70 % de los días en clima templado, el flujo de pacientes es alto o medio (1)
> La mayoría de los días en clima frío, el flujo de pacientes es muy bajo o bajo (0.89)
> La mayoría de los días en Mayo, el flujo de pacientes es medio (1)
> La mayoría de los días en Junio, el flujo de pacientes es medio (0.73)
> La mayoría de los días en Julio, el flujo de pacientes es alto (1)
> Aprox. más del 70 % de los días en Agosto, el flujo de pacientes es muy alto (1)
> La mayoría de los días en Septiembre, el flujo de pacientes es alto (0.83)"

Después de pasar por el post-proceso, el conjunto de sentencias se convertirá en un párrafo legible de la forma:

"Aprox. más del 70 % de los días en clima templado, el flujo de pacientes es alto o medio; y en Agosto, alto. La mayoría de los días en clima frío, es muy bajo o bajo; en Mayo y Junio, medio; y en Julio y Septiembre, alto".

Discusión

Como se puede observar, los resultados obtenidos usando las distintas estrategias son iguales, excepto por la sentencia que se obtiene para resumir el periodo *Agosto*. Veámoslo gráficamente mediante las figuras 4.7 y 4.8. En ellas se ha marcado de forma más clara el periodo con sentencias que difieren.

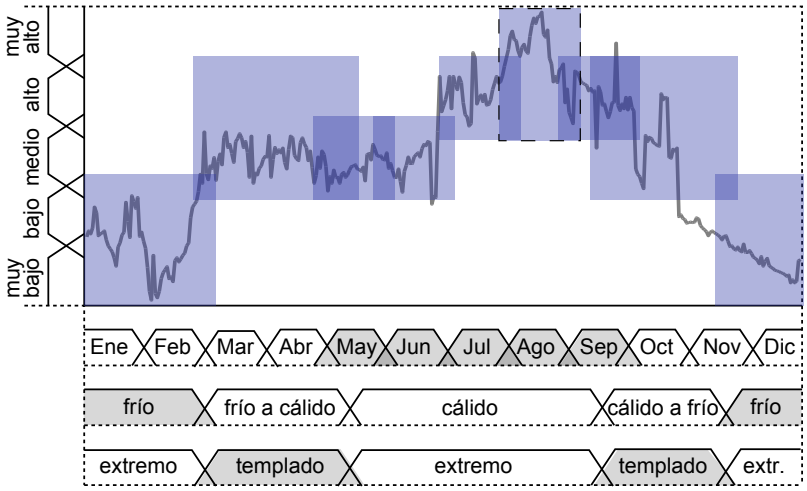


Figura 4.7: Elecciones de descripción del Algoritmo 1 para el problema C_A.

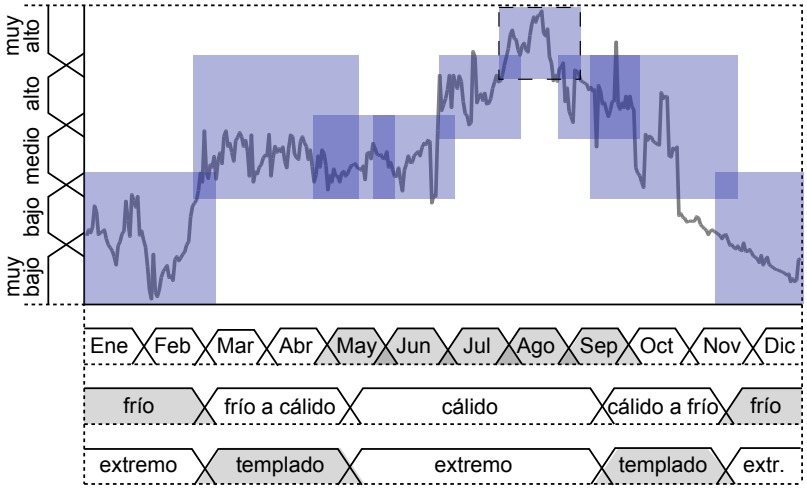


Figura 4.8: Elecciones de descripción del Algoritmo 2 para el problema C_A.

Mientras que con la primera estrategia (Algoritmo 1) obteníamos,

La mayoría *de los días en Agosto, el flujo de pacientes es* ***alto o muy alto***

usando la segunda estrategia (Algoritmo 2) tenemos,

Aprox. más del 70 % *de los días en Agosto, el flujo de pacientes es* ***muy alto***

Notamos que la primera estrategia se ha decantado por el cuantificador más estric-

to aún si para ello ha debido hacer agrupaciones, mientras que la segunda estrategia se decanta por un cuantificador menos estricto para poder conservar la especificidad de los términos usados para describir la variable.

En ambos casos las soluciones son de bondad parecida en relación con los criterios de calidad definidos para nuestro modelo de resumen y evaluación del resumen. Con respecto a la brevedad ambas estrategias describen el año con sólo 7 sentencias. La cobertura es total en ambos casos para toda la línea temporal. Los grados de exactitud son idénticos, ya que si bien el resumen difiere en una sentencia, dicha sentencia en ambos casos posee el mismo grado de cumplimiento. La especificidad también es bastante parecida ya que ambas estrategias intentan maximizarla pero prestando mayor atención a una componente distinta en cada ocasión.

Volvemos a remarcar que ninguna estrategia es mejor que la otra, son similares pero con diferentes matices semánticos durante la búsqueda. Diferentes usuarios preferirán las mismas o diferentes estrategias, o un mismo usuario puede preferir una estrategia frente a la otra dependiendo del problema o el contexto.

4.2.3. Efectos de los parámetros en la búsqueda

En la Sección 4.2.1 se han presentado dos estrategias Greedy para la obtención de resúmenes lingüísticos de series de datos mediante uso del ordenador. La Sección 4.2.2 se ha dedicado a ilustrar el comportamiento de los algoritmos asociados a dichas estrategias.

Hemos visto que cada estrategia se inclina más hacia un tipo de resúmenes, pero el resumen final no sólo depende de la estrategia la serie de datos, sino que también depende de la familia de cuantificadores que usemos, del umbral τ, o los límites fijados para cada nivel en la jerarquía.

La dependencia de los cuantificadores es obvia ya que la familia de cuantificadores debe reflejar el vocabulario referente a cantidades vagas o difusas que el usuario desea usar en su resumen.

Con respecto a los otros parámetros, los cambios de éstos pueden repercutir en el número de sentencias del resumen final, la complejidad de las mismas, en lo referente a agrupaciones de etiquetas, o en la cantidad de puntos afectados por la sentencia.

La cobertura total está asegurada a través de la forma en la que se realiza la exploración, de modo que las sentencias cubren todos los puntos de la serie temporal.

En general, las sentencias que componen el resumen deben tener un grado de cumplimiento mayor o igual que el umbral τ definido, de modo que valores menores que dicho τ se traducirán en un resumen más breve, aunque, como contrapartida, se disminuirá la precisión de las sentencias (siempre y cuando el resto de parámetros

continúe igual).

Con respecto al límite para el cuantificador ($Qlim_i$) en los distintos niveles i, éste indica el cuantificador menos estricto que se tendrá en cuenta para la realización de los resúmenes. Si su valor crece, más cuantificadores serán considerados, de modo que el resumen será más breve, pero puede que también menos específico (si el resto de parámetros continúa igual). El límite de agrupación ($Glim_i$) del nivel i indica el máximo número de etiquetas E_i que se está dispuesto a agregar. Si el valor del límite crece, el resumen será más corto pero también más complejo.

Tengamos en cuenta que es inmediato el razonamiento que nos indica que si usamos $Qlim_i = Glim_i = 1\ \forall i$, los Algoritmos 1 y 2 son equivalentes a trabajar con un sólo cuantificador Q (el más estricto de la familia) y etiquetas simples en A, de modo que ambos arrojarán el mismo resultado para un mismo conjunto de datos si el resto de parámetros son iguales.

A continuación se muestran algunos experimentos en los que, basándonos en el contexto y los datos definidos en 4.2.2, tratamos de ilustrar las consecuencias al realizar cambios en los valores de los distintos parámetros.

Cambios en el umbral τ

Comenzamos ilustrando las repercusiones que se originan al realizar cambios en el umbral establecido para el grado de cumplimiento de las sentencias que formarán el resumen. Manteniendo los límites como $Qlim_i = Glim_i = 2$ pero incrementando el valor del umbral de $\tau = 0.7$ a $\tau = 0.9$, y usando el Algoritmo 1 obtenemos el resumen,

> "Aprox. más del 70 % de los días en clima templado, el flujo de pacientes es alto o medio (1)
> Aprox. más del 70 % de los días en clima frío, el flujo de pacientes es muy bajo o bajo (1)
> La mayoría de los días en Mayo, el flujo de pacientes es medio (1)
> La mayoría de los días en Junio, el flujo de pacientes es medio o bajo (1)
> La mayoría de los días en Julio, el flujo de pacientes es alto (1)
> La mayoría de los días en Agosto, el flujo de pacientes es muy alto (1)
> La mayoría de los días en Septiembre, el flujo de pacientes es alto (0.99)"

Como primera conclusión podríamos decir que el número de sentencias se ha mantenido igual (idéntica brevedad), pero ahora todas las sentencias poseen un cumplimiento mayor a 0.9 (mayor exactitud). Fijémonos ahora en las sentencias de forma individual. En la segunda sentencia, para poder subir el cumplimiento de 0.89 (0.89 $<$ 0.9) a 1, se ha pasado de un cuantificador *La mayoría* a *Al menos el 70 %*. La cuarta sentencia también ha experimentado cambios para incrementar su grado de cumplimiento de 0.73 a 1. En este caso, la etiqueta *medio* se ha agrupado con *bajo* manteniendo el cuantificador más estricto *La mayoría* de acuerdo con las prioridades

del algoritmo. La última modificación sucede en la última sentencia donde *alto* pasa a ser *alto o muy alto* para subir de 0.83 a 0.99. Estas tres últimas modificaciones reflejan una pérdida en la especificidad global del resumen.

Si por el contrario el algoritmo usado es el Algoritmo 2 tenemos,

> "Aprox. más del 70 % de los días en clima templado, el flujo de pacientes es alto o medio (1)
> Aprox. más del 70 % de los días en clima frío, el flujo de pacientes es muy bajo o bajo (1)
> La mayoría de los días en Mayo, el flujo de pacientes es medio (1)
> Aprox. más del 70 % de los días en Junio, el flujo de pacientes es medio (1)
> La mayoría de los días en Julio, el flujo de pacientes es alto (1)
> Aprox. más del 70 % de los días en Agosto, el flujo de pacientes es muy alto (1)
> Aprox. más del 70 % de los días en Septiembre, el flujo de pacientes es alto (1)"

De nuevo, el número de sentencias continúa igual y las únicas que han experimentado cambios son aquellas cuyo grado de cumplimiento era menor que 0.9. Todos los cambios, sentencias 2, 4, y 7 se han debido al uso de un cuantificador menos estricto. En la segunda sentencia se ha pasado de *La mayoría* a *Al menos el 70 %*. El mismo cambio han experimentado las sentencias 4 y 7 en lugar del agrupamiento aplicado por el Algoritmo 1. Se mantiene la brevedad, se incrementa la exactitud, pero sin embargo, se decrementa la especificidad.

Sigamos manteniendo los límites, pero esta vez decrementemos el umbral de $\tau = 0.7$ a $\tau = 0.5$. En esta ocasión ambos algoritmos han encontrado la misma solución,

> "La mayoría de los días en clima templado, el flujo de pacientes es medio o alto (0.68)
> La mayoría de los días en clima frío, el flujo de pacientes es muy bajo o bajo (0.89)
> Aprox. más del 70 % de los días en tiempo cálido, el flujo de pacientes es medio o alto (0.56)

Este ejemplo ilustra cómo la disminución del umbral repercute en la obtención de un resumen más breve pero menos exacto, como se indica mediante el grado de cumplimiento asociado a las sentencias.

Cambios en el límite $Glim_i$

Continuemos con nuestro ejemplo; si se establece el límite de agrupamiento como $Glim_i = 1$ y mantenemos $Qlim_i = 2$ y $\tau = 0.7$, tenemos un resumen de la siguiente forma para ambos algoritmos,

"La mayoría de los días en clima de frío a cálido, el flujo de pacientes es medio (1)
La mayoría de los días en Enero, el flujo de pacientes es bajo (0.81)
En Febrero, el flujo de pacientes es altamente variable
La mayoría de los días en Noviembre, el flujo de pacientes es bajo (0.89)
La mayoría de los días en Diciembre, el flujo de pacientes es muy bajo (0.90)
La mayoría de los días en Mayo, el flujo de pacientes es medio (1)
La mayoría de los días en Junio, el flujo de pacientes es medio (0.73)
La mayoría de los días en Julio, el flujo de pacientes es alto (1)
Aprox. más del 70 % de los días en Agosto, el flujo de pacientes es muy alto (1)
La mayoría de los días en Septiembre, el flujo de pacientes es alto (0.83)
En Octubre, el flujo de pacientes es altamente variable"

Como era de esperar, el resumen es menos breve. Ha pasado de 7 a 11 sentencias debido al hecho de que se le ha impuesto que sólo pueda usar etiquetas simples al construir las sentencias que serán menos complejas. Por el contrario, si ponemos el $Glim_i = 3\ \forall i$ y mantenemos los valores de $Qlim_i = 2\ \forall i$ y $\tau = 0.7$, obtenemos,

"La mayoría de los días en clima templado, el flujo de pacientes es alto, medio o bajo (1)
La mayoría de los días en clima frío, el flujo de pacientes es medio, bajo o muy bajo (0.89)
La mayoría de los días en clima cálido, el flujo de pacientes es muy alto, alto o medio (1)"

dicho resumen es más corto (de 7 a 3 sentencias), pero está compuesto por sentencias con una estructura más compleja, de menor especificidad.

Cambios en el límite $Qlim_i$

Finalmente, probamos a cambiar el valor de $Qlim_i$. Si cambiamos el valor a $Qlim_i = 1\ \forall i$ y mantenemos el resto de parámetros como inicialmente, $Glim_i = 2\ \forall i$ y $\tau = 0.7$, obtenemos,

"La mayoría de los días en clima frío, el flujo de pacientes es bajo o muy bajo (0.89)
La mayoría de los días en clima frío a cálido, el flujo de pacientes es medio (1)
La mayoría de los días en Mayo, el flujo de pacientes es medio (1)
La mayoría de los días en Junio, el flujo de pacientes es medio (0.73)
La mayoría de los días en Julio, el flujo de pacientes es alto (1)
La mayoría de los días en Agosto, el flujo de pacientes e alto o muy alto (1)
La mayoría de los días en Septiembre, el flujo de pacientes es alto (0.83)
La mayoría de los días en Octubre, el flujo de pacientes es medio o alto (1)
La mayoría de los días en Noviembre, el flujo de pacientes es bajo (0.89)"

Debido a que somos más rígidos con respecto al cuantificador que se puede usar, sólo *La mayoría*, el número de sentencias aumenta de 7 a 9.

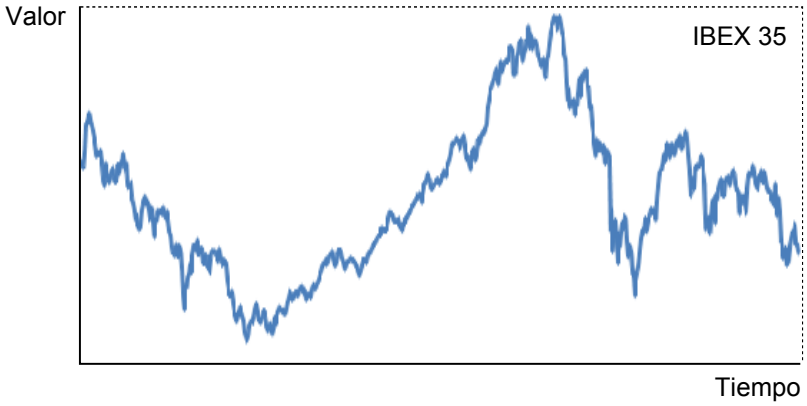


Figura 4.9: Valor de cotización del IBEX 35 en el periodo 2000-2011.

4.2.4. Ejemplo: IBEX35

En la presente sección presentaremos otro ejemplo para ilustrar el funcionamiento de las diferentes estrategias Greedy a la hora de realizar resumen lingüístico de series de datos temporales. Al contrario que en el caso anterior, en esta ocasión, los datos que utilizaremos han sido obtenidos de una base de datos real.

En esta ocasión contamos con un cubo de datos en el que se almacenan los valores de los principales índices en bolsa de una serie de países a lo largo del tiempo. Algunos de los valores almacenados son el FTSE 100 de Gran Bretaña, el CAC 40 de Francia o el DAX 30 de Alemania entre otros (información obtenida de la web Yahoo finance [46]). En el caso de España se encuentra almacenado el valor del IBEX 35 o índice selectivo de la Bolsa de Madrid (información obtenida a través de la web de la Bolsa de Madrid [32]).

Mediante operaciones OLAP se ha extraído una serie de datos que representa el valor del IBEX 35 desde el año 2000 al 2011 ambos incluidos, a razón de una medida por semana, lo que hace un total de 621 medidas. Los valores de la serie se encuentran representados en la Figura 4.9.

A continuación definiremos el marco lingüístico adecuado para el problema. La dimensión que representa a la variable bajo estudio se ha particionado haciendo uso de once etiquetas diferentes que describen el valor de cotización en tramos de mil unidades. Para construir la partición ha sido necesario hallar el valor mínimo y valor máximo que toma la serie en el periodo de tiempo que deseamos resumir; esto nos ayudará a hacer una partición ajustada al rango de los posibles valores en un periodo determinado. Para más información acerca de esta partición véase la Tabla 4.31.

En este caso se han usado cuatro niveles de granularidad diferentes al construir la

Etiqueta	Definición
entre 5000 y 6000	(4800, 5000, 6000, 6200)
entre 6000 y 7000	(5800, 6000, 7000, 7200)
entre 7000 y 8000	(6800, 7000, 8000, 8200)
entre 8000 y 9000	(7800, 8000, 9000, 9200)
entre 9000 y 10000	(8800, 9000, 10000, 10200)
entre 10000 y 11000	(9800, 10000, 11000, 11200)
entre 11000 y 12000	(10800, 11000, 12000, 12200)
entre 12000 y 13000	(11800, 12000, 13000, 13200)
entre 13000 y 14000	(12800, 13000, 14000, 14200)
entre 14000 y 15000	(13800, 14000, 15000, 15200)
entre 15000 y 16000	(14800, 15000, 16000, 16200)

Tabla 4.31: Partición del dominio de la variable para el ejemplo $IBEX35$.

jerarquía en la dimensión temporal. Comenzaremos la descripción desde el nivel más general y continuaremos hasta alcanzar el más preciso. El nivel con mayor abstracción está formado por dos etiquetas lingüísticas que agrupan las medidas por décadas (década de los 00s, década de los 10s). La segunda partición se compone de cuatro etiquetas con un mayor nivel de granularidad que describen las décadas con mayor detalle (a comienzos de los 00s, a mediados de los 00s, ...). El tercer nivel de abstracción lo forman doce etiquetas lingüísticas que representan los años de una forma difusa. Por último, el cuarto nivel, aquel con menor nivel de abstracción, está compuesto por veinticuatro etiquetas que describen a los años en función de sus semestres (es decir, la primera mitad de 2004 o la segunda mitad de 2005).

No volveremos sobre la discusión que ya se ha hecho acerca de la idoneidad del uso de conjuntos lingüísticos para describir las transiciones que tienen lugar al cambiar de periodos en la dimensión temporal. Los valores concretos de las etiquetas anteriormente mencionadas se representan en la Tabla 4.32.

En la Figura 4.10 se puede apreciar la serie de datos IBEX35 pero esta vez con el contexto lingüístico definido en las fases previas. En el eje de ordenadas de la gráfica podemos ver la partición de conjuntos difusos que describen a la variable con sus etiquetas lingüísticas asociadas. En el eje de abscisas se representa el tiempo. Por cuestiones de espacio no aparecen los nombres de las etiquetas de la partición del nivel número 4.

Con respecto a los cuantificadores, en esta ocasión se ha optado por la familia representada en la Tabla 4.33. En esta ocasión no sólo se ve aumentado el número de cuantificadores sino que se han considerado cuantificadores más estrictos. Ahora el cuantificador *La mayoría* es bastante más estricto que en el ejemplo anterior.

Por último ajustaremos los parámetros de manejo del algoritmo. En esta ocasión al

Nivel	Etiqueta	Definición
1	década de los 00s	(1, 1, 523, 528)
	década de los 10s	(518, 524, 621, 621)
2	comienzos de los 00s	(1, 1, 209, 214)
	mediados de los 00s	(205, 210, 366, 371)
	finales de los 00s	(361, 367, 523, 528)
	comienzos de los 10s	(518, 524, 621, 621)
3	2000	(1, 1, 53, 58)
	2001	(48, 54, 105, 110)
	2002	(101, 106, 157, 163)
	2003	(153, 158, 209, 215)
	2004	(204, 210, 262, 267)
	2005	(257, 263, 314, 319)
	2006	(309, 315, 366, 371)
	2007	(361, 367, 418, 423)
	2008	(414, 419, 470, 476)
	2009	(466, 471, 523, 528)
	2010	(518, 524, 575, 580)
	2011	(570, 576, 621, 621)
4	primera mitad del 2000	(1, 1, 26, 30)
	segunda mitad del 2000	(24, 27, 53, 56)
	primera mitad del 2001	(50, 54, 79, 82)
	segunda mitad del 2001	(76, 80, 105, 108)
	primera mitad del 2002	(102, 106, 131, 134)
	segunda mitad del 2002	(128, 132, 157, 160)
	primera mitad del 2003	(155, 158, 183, 186)
	segunda mitad del 2003	(181, 184, 209, 212)
	primera mitad del 2004	(207, 210, 235, 238)
	segunda mitad del 2004	(233, 236, 262, 265)
	primera mitad del 2005	(259, 263, 287, 290)
	segunda mitad del 2005	(285, 288, 314, 317)
	primera mitad del 2006	(311, 315, 340, 343)
	segunda mitad del 2006	(337, 341, 366, 369)
	primera mitad del 2007	(363, 367, 392, 395)
	segunda mitad del 2007	(390, 393, 418, 421)
	primera mitad del 2008	(416, 419, 444, 447)
	segunda mitad del 2008	(442, 445, 470, 473)
	primera mitad del 2009	(468, 471, 496, 499)
	segunda mitad del 2009	(494, 497, 523, 526)
	primera mitad del 2010	(521, 524, 548, 551)
	segunda mitad del 2010	(546, 549, 575, 578)
	primera mitad del 2011	(572, 576, 600, 603)
	segunda mitad del 2011	(598, 601, 621, 621)

Tabla 4.32: Partición de la dimensión temporal para el ejemplo $IBEX35$.

Figura 4.10: Valor de cotización del IBEX 35 en el periodo 2000-2011.

Cuantificador	Definición
La mayoría	(0, 0.8, 0.9, 1)
Al menos el 80 %	(0, 0.7, 0.8, 1)
Al menos el 70 %	(0, 0.6, 0.7, 1)
Al menos el 60 %	(0, 0.5, 0.6, 1)

Tabla 4.33: Cuantificadores para el ejemplo $IBEX35$.

umbral τ se le ha dado un valor de 0.8 (también más estricto que en el ejemplo anterior donde recibía un valor del 0.7). Con respecto a los límites, $Qlim_i = 3$ y $Glim_i = 2$ para todos los niveles i; de modo que se permite el uso de los cuantificadores *La mayoría, al menos el 80 % y al menos el 70 %* y las agrupaciones por parejas en las etiquetas que describen la variable.

A continuación hemos aplicado las diferentes estrategias Greedy para el resumen. Junto con los resumenes en formato textual se adjunta la representación mediante sombreado de las soluciones (4.11 y 4.12).

Aplicando el Algoritmo 1 con los parámetros especificados anteriormente obtenemos la siguiente solución:

"Al menos el 70 % de los días de la década de los 10s, el valor se sitúa entre 10000 y 11000 o entre 9000 y 10000 (1)
Al menos el 70 % de los días del 2000, el valor se sitúa entre 11000 y 12000 o entre 10000 y 11000 (1)
Al menos el 80 % de los días del 2001, el valor se sitúa entre 9000 y 10000 o entre 8000 y 9000 (1)
Al menos el 80 % de los días del 2002, el valor se sitúa entre 8000 y 9000 o entre 6000 y 7000 (0.86)
La mayoría de los días del 2003, el valor se sitúa entre 7000 y 8000 o entre 6000 y 7000 (0.99)
La mayoría de los días del 2004, el valor se sitúa entre 8000 y 9000 o entre 7000 y 8000 (1)
La mayoría de los días del 2005, el valor se sitúa entre 10000 y 11000 o entre 11000 y 12000 (1)
Al menos el 70 % de los días del 2006, el valor se sitúa entre 14000 y 15000 o entre 11000 y 12000 (0.89)
La mayoría de los días del 2007, el valor se sitúa entre 15000 y 16000 o entre 14000 y 15000 (1)
Al menos el 80 % de los días de la primera mitad del 2008, el valor se sitúa entre 13000 y 14000 o entre 12000 y 13000 (1)
Al menos el 80 % de los días de la segunda mitad del 2008, el valor se sitúa entre 11000 y 12000 o entre 9000 y 10000 (0.96)
Al menos el 80 % de los días de la primera mitad del 2009, el valor se sitúa entre 9000 y 10000 o entre 10000 y 11000 (0.84)
La mayoría de los días de la segunda mitad del 2009, el valor se sitúa entre 11000 y 12000 o entre 10000 y 11000 (0.81)"

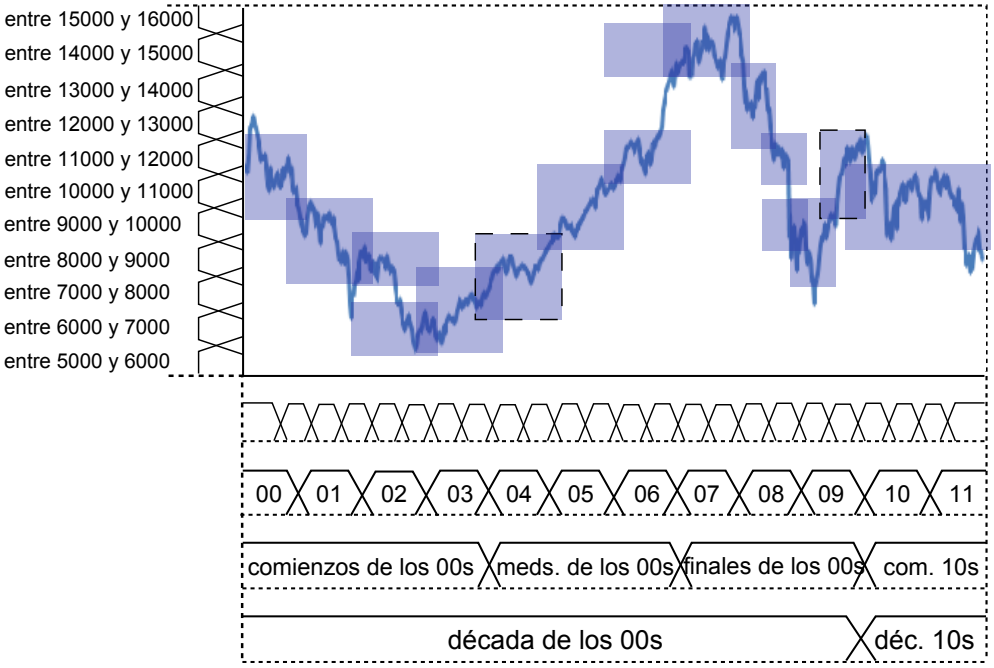


Figura 4.11: Elecciones de descripción del Algoritmo 1 para el problema $IBEX35$.

Si en cambio usamos el Algoritmo 2 obtenemos:

"Al menos el 70 % de los días de la década de los 10s, el valor se sitúa entre 10000 y 11000 o entre 9000 y 10000 (1)
Al menos el 70 % de los días del 2000, el valor se sitúa entre 11000 y 12000 o entre 10000 y 11000 (1)
Al menos el 80 % de los días del 2001, el valor se sitúa entre 9000 y 10000 o entre 8000 y 9000 (1)
Al menos el 80 % de los días del 2002, el valor se sitúa entre 8000 y 9000 o entre 6000 y 7000 (0.86)
La mayoría de los días del 2003, el valor se sitúa entre 7000 y 8000 o entre 6000 y 7000 (0.99)
Al menos el 70 % de los días del 2004, el valor se sitúa entre 8000 y 9000 (0.95)
La mayoría de los días del 2005, el valor se sitúa entre 10000 y 11000 o entre 11000 y 12000 (1)
Al menos el 70 % de los días del 2006, el valor se sitúa entre 14000 y 15000 o entre 11000 y 12000 (0.89)
La mayoría de los días del 2007, el valor se sitúa entre 15000 y 16000 o entre 14000 y 15000 (1)
Al menos el 80 % de los días de la primera mitad del 2008, el valor se sitúa entre 13000 y 14000 o entre 12000 y 13000 (1)
Al menos el 80 % de los días de la segunda mitad del 2008, el valor se sitúa entre 11000 y 12000 o entre 9000 y 10000 (0.96)
Al menos el 80 % de los días de la primera mitad del 2009, el valor se sitúa entre 9000 y 10000 o entre 10000 y 11000 (0.84)
Al menos el 70 % de los días de la segunda mitad del 2009, el valor se sitúa entre 11000 y 12000 (1)"

Como podemos observar los resultados son muy similares, pero si prestamos un poco más de atención veremos que existen diferencias entre ellos debido a las particularidades semánticas que introduce cada estrategia.

Para encontrar la primera de las diferencias debemos mirar las sentencias que utilizan ambas estrategias para describir el año *2004*. Mientras que la primera estrategia elige:

La mayoría de los días del 2004, el valor se sitúa **entre 8000 y 9000 o entre 7000 y 8000**

la segunda estrategia se decanta por:

Al menos el 70 % de los días del 2004, el valor se sitúa **entre 8000 y 9000**

En ambas sentencias vemos claramente reflejado el espíritu de cada una de las estrategias, mientras la primera ha preferido un cuantificador más estricto, la segunda ha optado por un término de descripción de la variable menos abstracto.

La segunda de las diferencias tiene lugar cuando se describe el periodo *segunda mitad del 2009*. La sentencia que encontramos en el resumen de la primera estrategia es:

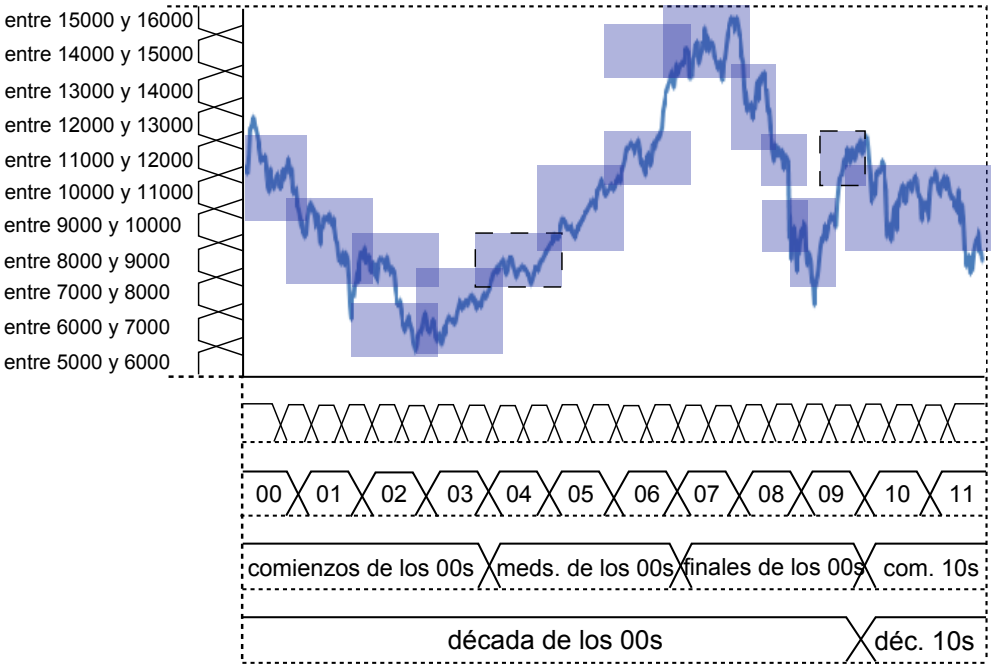


Figura 4.12: Elecciones de descripción del Algoritmo 2 para el problema $IBEX35$.

La mayoría de los días de la segunda mitad del 2009, el valor se sitúa **entre 11000 y 12000 o entre 10000 y 11000**

mientras que la que encontramos en el resumen de la segunda estrategia es:

Al menos el 70 % de los días de la segunda mitad del 2009, el valor se sitúa **entre 11000 y 12000**

De nuevo en esta ocasión observamos la predilección de cada una de las estrategias por insertar los términos más precisos en diferentes componentes de la sentencia cuantificada.

4.3. Estudio de técnicas evolutivas

Como ya se ha comentado, los algoritmos Greedy ofrecen muy buenas soluciones en un espacio acotado de tiempo que no resulta demasiado gravoso para el usuario.

Sin embargo, con el objetivo de comprobar la bondad de las soluciones que se encuentran usando este algoritmo en sus dos versiones, hemos optado por implementar un algoritmo evolutivo [17, 18]. Este tipo de algoritmos exploran de forma más extensiva el espacio de soluciones. De este modo nos ofrecen una muestra lo bastante

grande de las mismas sin llegar a ser tan extensa como si lo hiciéramos de manera exhaustiva.

Además, entre la gran familia que compone la computación evolutiva, existe cierta rama que se encarga de la optimización de problemas multi-objetivo. Este tipo de algoritmos son muy adecuados para conseguir optimizar los diversos criterios de calidad asociados a un resumen. En general este tipo de algoritmos ofrecen como resultado final una serie de soluciones óptimas.

4.3.1. Algoritmos evolutivos

Podemos decir que los algoritmos evolutivos [35, 43] son métodos estocásticos de optimización y búsqueda de soluciones inspirados por la *Teoría de la evolución de Darwin* [28], en particular por los procesos biológicos que permiten a las poblaciones de organismos adaptarse a su entorno. Es decir, los individuos mejor adaptados a su entorno serán los que tengan más probabilidades de supervivencia, y por tanto más posibilidades de transmitir su material genético a posteriores generaciones.

En este tipo de algoritmos se mantiene un conjunto o población, de entidades o individuos, que representan posibles soluciones en el espacio de búsqueda, las cuales se cruzan entre ellas y compiten entre sí, de tal manera que las más aptas son capaces de prevalecer a lo largo del tiempo evolucionando hacia mejores soluciones cada vez.

La evolución se consigue aplicando de forma iterativa una serie de operadores estocásticos conocidos como *mutación*, *recombinación* o *cruce* y *selección*. La mutación realiza cambios aleatorios en las soluciones; la recombinación descompone dos soluciones distintas y mezcla sus partes aleatoriamente para crear nuevas soluciones, y la selección replica individuos de la población con buenas cualidades teniendo en cuenta la calidad de las mismas.

La población inicial puede ser establecida mediante un proceso aleatorio o puede ser instanciada con soluciones encontradas mediante otros mecanismos de búsqueda local, si éstas están disponibles. El resultado final tiende a encontrar, si se le da el tiempo necesario, soluciones óptimas globales al problema de la misma forma en la que los organismos en la naturaleza se adaptan a su entorno.

Habitualmente, la mayoría de los problemas implican la existencia de diversos objetivos que deben ser optimizados de forma simultánea. Sin embargo, en la práctica esto no es sencillo, o incluso puede ser imposible, ya que pueden entrar en conflicto entre ellos. Con el fin de poder enfrentase a este tipo de problemática se han propuesto algoritmos evolutivos multi-objetivo (MOEAs) que usan diversas técnicas [47].

El problema de encontrar un buen resumen lingüístico de un conjunto de datos puede ser naturalmente formulado como un problema de optimización multi-criterio,

donde distintas medidas de calidad deben ser maximizadas, tal como hemos visto al presentar nuestra propuesta de modelo de calidad. Esto significa que, en general, no es viable obtener el *mejor* resumen lingüístico posible. Sin embargo, al usuario le serán ofrecidas una serie de soluciones que presentan diferentes combinaciones para los criterios de calidad.

Un algoritmo evolutivo multi-objetivo muy popular y efectivo es el llamado NSGA-II [33]. Este algoritmo trabaja por medio de la ordenación de las soluciones candidatas en frentes de Pareto, de modo que las mejores soluciones se encontrarán en el primer frente. Además, aplica una técnica basada en nichos[1] y elitismo para mejorar la población completa del frente de Pareto.

Se ha adoptado dicho algoritmo y se ha modificado con la finalidad de adecuarlo a las particularidades de nuestro problema, es decir, el resumen lingüístico. Dichas adaptaciones consisten básicamente en la definición de una serie de operadores genéticos específicos.

4.3.2. Presentación de la propuesta sobre NSGA-II

A continuación se presentarán las diferentes decisiones que se han tomado a la hora de definir el algoritmo genético que usaremos para explorar el espacio de búsqueda de soluciones.

Representación de las soluciones

La primera tarea que debemos afrontar cuando decidimos trabajar con algoritmos genéticos es la definición de la representación de las soluciones. El uso de la memoria en la representación es esencial para obtener un desarrollo correcto en términos de memoria y tiempo. Los siguientes pasos se verán altamente influenciados por las decisiones que tomemos en esta parte del diseño. El conocimiento a fondo de la representación elegida es fundamental para poder diseñar los operadores genéticos de forma correcta. Del mismo modo es muy importante tanto para la inicialización de la población como para la evaluación de objetivos y restricciones en la población.

Un resumen lingüístico se encuentra representado mediante un cromosoma de longitud variable, dividido en componentes lógicas, genes, que representarán las sentencias cuantificadas del resumen. Dichos genes están compuestos a su vez por componentes lógicas que representan los componentes de la sentencia.

De este modo podemos decir que un resumen o solución es un cromosoma; que las sentencias *"Q de D son A"* que componen el resumen serán los genes del cromosoma, y que los genes almacenan las componentes Q, D y A de las sentencias. La Figura 4.13 nos ilustra de forma gráfica la representación seleccionada.

[1] niching

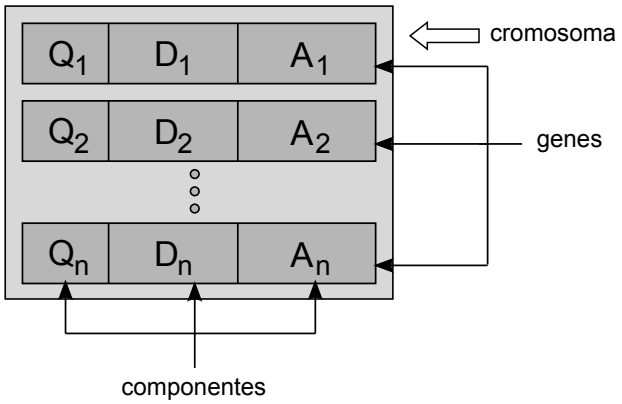


Figura 4.13: Representación de una solución.

Objetivos

Los objetivos, es decir, lo que deseamos conseguir o premiar, son los objetivos de calidad definidos en la Sección 3.5: cobertura, $c_d(s)$, brevedad, $b(s)$, especificidad, $p(s)$ y exactitud, $a_d(s)$, del resumen.

Un problema de NSGA-II citado en [1] es la mala escalabilidad inherente a los algoritmos de optimización multi-objetivo en relación con el número de objetivos evaluados. Aunque la buena actuación de los MOEAs cuando el problema cuenta con dos o tres objetivos a optimizar está suficientemente demostrada, es cierto que conforme aumenta ese número estas técnicas funcionan cada vez peor.

La mala actuación de los MOEAs convencionales se achaca a varios aspectos que se deben tener en cuenta, por ejemplo la creciente complejidad que siempre existe en espacios de búsqueda con un alto número de dimensiones, el uso de operadores de selección y mutación que no tienen en cuenta las características de este tipo de espacios y los tamaños de población inapropiados para llevar a cabo una búsqueda evolutiva en un espacio multi-dimensional.

Hemos realizado pruebas teniendo en cuenta los cuatro objetivos iniciales pero hemos observado que cuanto mayor es el problema más se acusa la mala escalabilidad. Como veremos, aunque las pruebas en el ámbito del problema del centro de salud sí han funcionado correctamente, cuando pasamos al problema del IBEX-35 observamos una mala convergencia de las soluciones a lo largo del tiempo. Esto se debe al mayor tamaño de la serie, pero sobre todo a la mayor complejidad del contexto lingüístico en que se pueden dar muchas más combinaciones.

Por fortuna, podemos poner remedio a este problema gracias a la generalidad de nuestro modelo de calidad. Nada hay en contra de que se adapten los objetivos a

nuestras necesidades. En este caso concreto se ha realizado una fusión de objetivos con el fin de reducir su número de cuatro a tres. Veremos más detalles a este respecto durante la exposición de la experimentación realizada en la Sección 4.3.3.

Restricciones

Hasta ahora en la memoria siempre se ha hablado de objetivos de calidad que deseamos alcanzar, pero para trabajar con algoritmos genéticos también puede ser necesario, como es el caso, definir restricciones. Si vemos a los objetivos como aquello deseable, debemos considerar las restricciones como aquellos comportamientos que queremos evitar.

Las restricciones asociadas a nuestro problema son:

- Inclusión: en un determinado resumen, el mismo periodo de tiempo no deberá de estar descrito por más de una sentencia si una de las etiquetas utilizadas es una generalización de otra,
- Umbral: la exactitud de las sentencias del resumen deberá ser siempre mayor o igual al umbral de tolerancia aportado por el usuario,
- *Qlim*: representa el cuantificador menos estricto que se puede usar en una sentencia,
- *Glim*: representa el máximo agrupamiento permitido entre términos lingüísticos de descripción de la variable.

Como podemos observar dichas restricciones no son nuevas para nosotros. Ya hemos dicho que el Greedy lleva a cabo una especie de búsqueda dirigida, y la forma en que la dirige es mediante el diseño que se ha hecho y los parámetros que introduce el usuario. Debido a la forma de explorar del algoritmo Greedy es imposible que en un resultado final aparezca más de una sentencia describiendo el mismo periodo de tiempo. Con respecto al umbral y los límites, tienen una gran influencia en el diseño, ya que el algoritmo Greedy nunca permitirá que en la solución final aparezcan conductas incorrectas, es decir, una sentencia con grado de cumplimiento menor que τ o con un Q no permitido para ese nivel de la jerarquía temporal.

La filosofía del algoritmo evolutivo es totalmente diferente ya que genera soluciones que, en principio, pueden ostentar una calidad baja o, incluso, no cumplir con alguna de las restricciones planteadas. Dichas soluciones se mantienen en la población para dar diversidad a la búsqueda. Es posible que estas soluciones, en principio malas, muten o se combinen con otras dando lugar a buenos individuos en generaciones posteriores.

Pero aunque deban ser mantenidas, no debemos olvidar que no son buenas y una manera de hacérselo saber al algoritmo es permitir que se penalicen estos comportamientos. La solución no se elimina, pero hay que avisar que no es una buena solución.

Inicialización

La inicialización de la primera población se llevará a cabo de forma aleatoria. La longitud de cada solución será obtenida mediante una distribución exponencial, mientras que los componentes Q, D y A se extraerán de una distribución uniforme.

El uso de una población inicial obtenida mediante procesos aleatorios no es una práctica inusual cuando se trabaja con algoritmos evolutivos. El objetivo es mantener la heterogeneidad mediante una amplia muestra de posibles soluciones que luego se irán mejorando con el tiempo.

Operadores

Con respecto a los operadores genéticos debemos destacar que hemos trabajado con un tipo de recombinación y varios tipos de mutaciones.

La *recombinación* toma dos resúmenes de la población y produce dos nuevos resúmenes mediante un cruce uniforme. Cada sentencia de cada resumen original (padre) va a un resumen generado (hijo) u otro con la misma probabilidad, es decir, $p = 1/2$.

El algoritmo NSGA-II construye dos permutaciones con los individuos de la población para decidir qué dos individuos se cruzarán entre sí. Para decidir si efectivamente el cruce se llevará a cabo se dispone de un parámetro que establece la probabilidad de cruce o p_c. Si los individuos se cruzan, dos nuevas soluciones se añadirán a la nueva población. Si no, serán replicados ellos mismos en la nueva población.

Se han diseñado e implementado cuatro operadores de mutación: uno clásico y tres específicos para el problema, que hemos llamado *inteligentes*, que llevan a cabo manipulaciones significativas sobre las sentencias que conforman el resumen. Dichos operadores los hemos denominado *cover*, *split* y *merge*.

La *mutación clásica* efectúa pequeñas mutaciones en las componentes Q, D y A de los genes con una probabilidad p_m. De modo que para cada componente de cada sentencia del resumen se aplica la probabilidad para ver si resultará mutado o no.

La *mutación cover* intenta garantizar la cobertura completa del conjunto de datos por parte del resumen. Para lograrlo, se buscan periodos de tiempo sin describir y se cubren con la sentencia más adecuada. Es decir, se selecciona la etiqueta $D_{i,j}$ que mejor cubra el periodo y se utilizan las componentes Q y A que maximicen el grado de cumplimiento de la sentencia cuantificada resultante (teniendo en cuenta los límites

Qlim y *Glim* del nivel i). Debido a que el objetivo de cobertura es tan importante, esta mutación se llevará a cabo siempre, de modo que $p_{m_{cover}} = 1$.

Un esbozo en forma de pseudo-código puede verse en el Algoritmo 3, donde la función *buscarHueco (tiempo, desde, inicio, longitud)* busca un periodo no cubierto en la dimensión temporal *tiempo* a partir de *desde*. Si se encuentra un hueco *inicio* y *longitud* devuelven el punto de inicio y la longitud del mismo respectivamente. La función *buscarEtiqueta (tiempo, inicio, longitud)* busca la etiqueta que mejor cubra el hueco encontrado. Finalmente la función *añadirEtiquetasComoGenes (individuo, etiquetasSeleccionadas)* se llama si el número de etiquetas seleccionadas es mayor que cero, para añadir dichas etiquetas al individuo mediante genes.

Algoritmo 3 : pseudo-código de la mutación *cover*.

```
1: etiquetasSeleccionadas ← ∅;
2: numEtiquetasSeleccionadas ← 0;
3: continuar ← cierto ;
4: random ← rnd(0, 1);
5: si random ≤ p_mi entonces
6:    mientras continuar hacer
7:       buscarHueco(tiempo, desde, inicio, longitud);
8:       si inicio = longitudTiempo and longitud = 0 entonces
9:          continuar ← falso ;
10:      si no
11:         etiqueta ← buscarEtiqueta(tiempo, inicio, longitud);
12:         si etiqueta no en etiquetasSeleccionadas entonces
13:            etiquetasSeleccionadas ← etiquetasSeleccionadas ∪ etiqueta;
14:            numEtiquetasSeleccionadas ← numEtiquetasSeleccionadas + 1;
15:         fin si
16:         si inicio + longitud ≥ longitudTiempo entonces
17:            continuar ← falso ;
18:         fin si
19:      fin si
20:      desde ← inicio + longitud;
21:   fin mientras
22:   si numEtiquetasSeleccionadas > 0 entonces
23:      anadirEtiquetasComoGenes(individuo, etiquetasSeleccionadas);
24:   fin si
25: fin si
```

La *mutación split* selecciona aleatoriamente una sentencia del resumen e intenta sustituirla por otras. Para ello se toma el periodo $D_{i,j}$ de la sentencia seleccionada y se hallan sus hijos si los tiene. En caso afirmativo, la sentencia seleccionada se eliminará del resumen (mediante *borrarEtiquetasComoGenes*) y se insertará una sentencia por cada periodo hijo (en el caso de que no se encuentren ya en el resultado). Las

componentes Q y A se seleccionan de forma que maximicen el grado de cumplimiento de la sentencia. La probabilidad de que una sentencia sea sometida al proceso de split es de $p_{m_{split}}$. Ver el Algoritmo 4, donde la función *seleccionaPeriodoTemporal (i)* nos da el periodo de tiempo en la posición i.

Algoritmo 4 : pseudo-código de la mutación *split*.

```
1: random ← rnd(0, 1);
2: si random ≤ p_m entonces
3:    split ← falso ;
4:    i ← numPeriodosHV;
5:    mientras i < longitudIndividuo y split = falso  hacer
6:       etiquetasSeleccionadas ← ∅;
7:       numEtiquetasSeleccionadas ← 0;
8:       padre ← seleccionaPeriodoTemporal(i);
9:       si (padre.nivel < numNivelesJeraquia − 1) entonces
10:         contador ← 0;
11:         para todo numNivelesJeraquia hacer
12:            l ← seleccionaPeriodoTemporal(contador);
13:            si l esHijoDe padre y l no en etiquetasSeleccionadas y l no en individuo
               entonces
14:               etiquetasSeleccionadas ← etiquetasSeleccionadas ∪ l;
15:               numEtiquetasSeleccionadas ← numEtiquetasSeleccionadas + 1;
16:            fin si
17:            contador ← contador + 1;
18:         fin para
19:      fin si
20:      i ← i + 1;
21:   fin mientras
22:   si numEtiquetasSeleccionadas > 0 entonces
23:      anadirEtiquetasComoGenes(individuo, etiquetasSeleccionadas);
24:      borrarEtiquetasComoGenes(individuo, padre);
25:   fin si
26: fin si
```

Por último, la *mutación merge* se podría considerar la opuesta de la anterior. De forma aleatoria se selecciona un número de sentencias que se intentarán unir entre ellas. Si es posible unirlas, es decir, tienen un padre común al que representan completamente, se sigue con el proceso. Se eliminarán las sentencias y se introducirá una nueva sentencia que contenga el periodo $D_{i,j}$ padre y los Q y A que maximicen el grado de cumplimiento de la nueva sentencia. Como se puede deducir, la probabilidad con la que se encontrarán al azar un número adecuado de etiquetas que puedan ser satisfactoriamente unidas es muy baja, por este motivo se establecerá la probabilidad como $p_{m_{merge}} = 1$, para que se intente siempre.

El pseudo-código de la mutación *merge* se puede ver en el Algoritmo 5, donde

la función *merge (random, etiquetasSeleccionadas, etiquetaParaAñadir)* devuelve un valor booleano que nos indica si ha sido posible fusionar las etiquetas *etiquetasSeleccionadas* obteniendo la nueva etiqueta *etiquetaParaAñadir*.

Algoritmo 5 : pseudo-código de la mutación *merge*.

```
1: aux ← 0;
2: etiquetasSeleccionadas ← ∅;
3: si longitudIndividuo − numPeriodosHV > 1 entonces
4:    random ← rnd(0, longitudIndividuo − numPeriodosHV − 1);
5:    si random > 1 entonces
6:       i ← 0;
7:       mientras i < random hacer
8:          aux ← rnd(numPeriodosHV, longitudIndividuo − 1);
9:          si aux no en etiquetasSeleccionadas entonces
10:            etiquetasSeleccionadas ← etiquetasSeleccionadas ∪ aux;
11:            i ← i + 1;
12:         fin si
13:      fin mientras
14:      etiquetaParaAnadir ← ∅;
15:      m ← merge(random, etiquetasSeleccionadas, etiquetaParaAnadir);
16:      si i > 0 and m = true entonces
17:         anadirEtiquetasComoGenes(individuo, etiquetaParaAnadir);
18:         borrarEtiquetasComoGenes(individuo, etiquetasSeleccionadas);
19:      fin si
20:   fin si
21: fin si
```

Como se puede comprobar, existe una notable componente heurística en los operadores específicos, de ahí el sobrenombre de *inteligentes*. Este mecanismo es una forma de asegurar que las sentencias nuevas no violan ninguna restricción, y que además tengan un nivel de calidad lo más aceptable posible.

4.3.3. Experimentación

En esta sección veremos los resultados obtenidos al aplicar la técnica evolutiva sobre los ejemplos introducidos anteriormente. En primer lugar se presentan unas consideraciones que se deben tener en cuenta al enfrentarse a la experimentación con algoritmos evolutivos. Seguiremos con el resumen de series que reflejan la afluencia de pacientes a un centro de salud, para a continuación pasar al resumen de los datos financieros dados por el IBEX-35.

Consideraciones previas

Ya hemos comentado que durante la definición del marco lingüístico de un problema determinado, además de las diferentes particiones, se debe dar valores a una serie de parámetros como son el umbral de cumplimiento τ y los límites *Qlim* y *Glim*. La variación de dichos parámetros permite al usuario adaptar la salida del proceso a sus necesidades en un momento determinado.

Cuando hablamos de técnicas evolutivas el número de esos parámetros crece, de modo que se nos ofrece una variedad más amplia de parámetros que en consecuencia resultan más difíciles de ajustar. Ahora, además de los parámetros puramente lingüísticos, el usuario deberá tener en cuenta otros de tipo evolutivo. Se deberá determinar el tamaño de la población y el número de generaciones, así como las probabilidades de cruce, mutación y mutación inteligente. En total 5 parámetros nuevos que añadir a los 3 existentes con la dificultad añadida que eso conlleva.

En esta memoria no hemos entrado en la realización del estudio sobre la selección de los mejores parámetros evolutivos posibles para un determinado problema. En su lugar hemos realizado una experimentación con un número elevado de combinaciones de parámetros de entre todas las posibles. Veámoslo con más detalle:

- *popSize* y *numGen* controlan el tamaño de la población y el número de generaciones. La definición de estos parámetros se debe hacer teniendo en cuenta que el número final de evaluaciones de soluciones debe ser el mismo para poder ser comparable. De este modo *popSize* va desde 100 hasta 3200 mientras que *numGen* lo hace de 800 a 25, teniendo en cuenta que en todas las combinaciones se deben realizar un total de 80000 evaluaciones (800*100 = 3200*25 = 80000).

- En general, la probabilidad de cruce p_c suele tener valores intermedios. En este caso se ha variado el valor de desde 0.5 hasta 0.7.

- La probabilidad de mutación p_m no suele ser muy alta para la gran mayoría de problemas. Se han realizado pruebas donde los valores se encuentran entre 0.01 y 0.4.

- Por último, la probabilidad de mutación inteligente p_{m_i} es algo específico de nuestro enfoque y no podemos guiarnos por pruebas anteriores. Se ha variado el valor desde 0.1 hasta 1.

Como la técnica evolutiva es estocástica, o fuertemente basada en probabilidades, se han realizado 10 ejecuciones para cada combinación de parámetros. De este modo es posible tener una idea más clara del efecto que tiene una determinada combinación de parámetros sobre la población final.

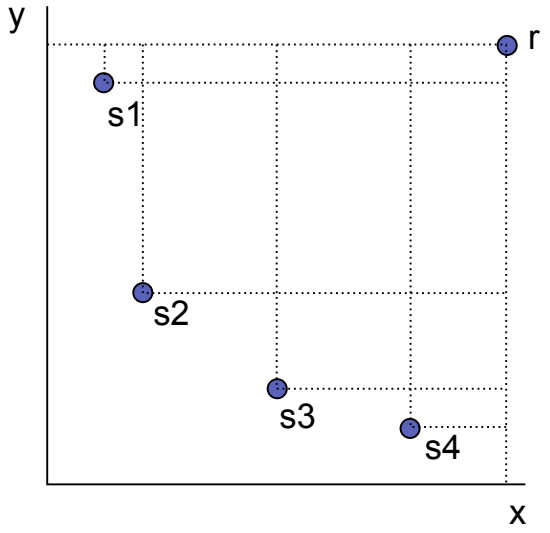


Figura 4.14: Ejemplo sencillo de frente de Pareto y su indicador de hipervolumen.

Una vez que tenemos las 10 diez ejecuciones con cada juego de parámetros deberemos saber la bondad de la combinación para poder compararla con la de otras combinaciones. Para esta tarea hemos hecho uso del *hypervolume indicator*. El indicador de hipervolumen, también conocido como *Lebesgue measure* o *S-metric* mide el volumen bajo las soluciones no dominadas de un frente de Pareto. Aunque posee algunas limitaciones, esta medida sigue siendo un referente como criterio guía a la hora de aceptar soluciones obtenidas por los algoritmos evolutivos multi-objetivo. Más información acerca de esta medida así como referencias, manual de uso y ejecutable puede encontrarse en [67].

Para calcular el indicador de hipervolumen necesitamos conocer el punto de referencia r. En la Figura 4.14 se muestra un pequeño ejemplo en el que se muestran varias soluciones optimales del frente de Pareto de un problema de dos dimensiones.

Para poder hallar el volumen necesitamos unos límites. El inferior lo constituyen las soluciones en sí mismas, pero necesitamos un límite superior que en éste caso se denomina *punto de referencia* o r. El punto de referencia debe estar situado de manera que incluya bajo él a todas las soluciones del frente de Pareto. En nuestro caso, para hallar el punto r se ha construido una población con 10.000 individuos generados al azar y se han seleccionado los mayores valores tomados por cada uno de los objetivos. Podemos comparar el punto de referencia a la peor solución posible que se ha encontrado, y por debajo de ella tenemos todas las soluciones encontradas; cuanto mayor sea la diferencia entre una solución y la peor, mejor será considerada.

Una vez que tenemos la medida de hipervolumen para cada población final se ha

calculado la media y la desviación estándar para las 10 ejecuciones de cada juego de parámetros. El juego con mayor media y menor desviación estándar tiene *más posibilidades* de ser una buena elección para los parámetros evolutivos. Para hacer un estudio más amplio de todos los juegos de soluciones, además de lo anterior se debería probar con los siguientes mejores valores y ver los resultados obtenidos. Además en muchas ocasiones la combinación con mejor media no tiene por qué tener la mejor desviación estándar.

Una vez realizadas estas aclaraciones pasamos a mostrar los resultados obtenidos en los casos prácticos. Para ello se ha seguido en ambas ocasiones una metodología concreta que pasamos a mostrar a continuación.

- Presentación del problema.
- Presentación del marco y parámetros lingüísticos.
- Presentación de los parámetros evolutivos.
- Presentación de soluciones Greedy (si es necesario).
- Presentación de algunas soluciones evolutivas encontradas.
- Comparación en términos de calidad de las soluciones.

Centro de salud C_B

Para este ejemplo recuperamos de nuevo el almacén de datos de centros de salud. Esta vez mediante una serie de operaciones OLAP se ha obtenido una nueva serie de 365 datos representando "la afluencia masculina a un centro C_B durante un año completo".

El marco lingüístico continua siendo el mismo que el usado para describir el centro C_A y que se encuentra representado en la Figura 4.4. Al pertenecer las dos series temporales al mismo almacén de datos no es necesario redefinir el marco lingüístico. La nueva serie C_B con el marco lingüístico anteriormente propuesto se puede ver en la Figura 4.15.

En esta ocasión se ha considerado una familia de cuantificadores más estricta que la usada en el ejemplo inicial. En lugar de los definidos en la Tabla 4.3, utilizaremos los definidos para el problema del IBEX-35 en la Tabla 4.33. Con respecto al resto de parámetros lingüísticos, tenemos que el umbral es $\tau = 0.8$ y los límites $Qlim = 3$ y $Glim = 2$ para todos los niveles de la jerarquía.

En cuanto a los parámetros evolutivos podemos ver los valores explorados en la Tabla 4.34.

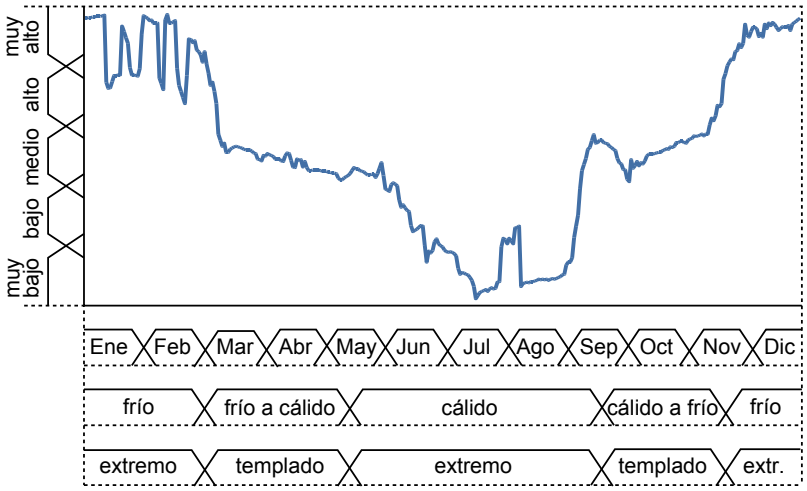


Figura 4.15: Flujo de pacientes masculinos al centro de salud C_B durante un año.

Parámetro	Valores posibles
$popSize$	100 200 400 800 1600 3200
$numGen$	800 400 200 100 50 25
p_c	0.5 0.6 0.7
p_m	0.01 0.05 0.1 0.2 0.4
p_{m_i}	0.1 0.2 1

Tabla 4.34: Parámetros evolutivos usados en la experimentación para el problema C_B.

En primer lugar veremos las soluciones obtenidas con las estrategias Greedy. Usando la primera estrategia (Algoritmo 1) tenemos el siguiente resultado:

Greedy Primera Estrategia
"La mayoría de los días con clima templado, el flujo de pacientes es medio (1)
La mayoría de los días con clima frío, el flujo de pacientes es alto o muy alto (1)
Al menos el 70 % de los días con clima cálido, el flujo de pacientes es bajo o muy bajo (0.93)"

El resultado obtenido utilizando la segunda estrategia (Algoritmo 2) es:

Greedy Segunda Estrategia
"La mayoría de los días con clima templado, el flujo de pacientes es medio (1)
Al menos el 70 % de los días con clima frío, el flujo de pacientes muy alto (1)
Al menos el 70 % de los días con clima cálido, el flujo de pacientes es bajo o muy bajo (0.93)"

Como se puede observar, también para este ejemplo los resultados Greedy son similares pero con matices semánticos que los diferencian (ver resumen para días con clima frío).

Una vez que tenemos todos los resultados obtenidos con las diferentes combinaciones de parámetros evolutivos y realizados los cálculos de hipervolumen tenemos que la que se revela como mejor combinación para este problema es *tamaño de la población* igual a 3200, *número de generaciones* igual a 25, *probabilidad de cruce* igual a 0.6, *probabilidad de mutación* igual 0.4 y *probabilidad de mutación inteligente* igual a 0.1.

De entre los individuos existentes en las poblaciones generadas utilizando dicha combinación de parámetros podemos encontrar ambas soluciones Greedy junto con otras buenas soluciones (comparables en términos de calidad). Algunos ejemplos de las soluciones encontradas que satisfacen de manera adecuada los criterios de calidad del modelo son:

Evolutiva n1
"Al menos el 70 % de los días con clima cálido, el flujo de pacientes es bajo o muy bajo (0.93)
Al menos el 80 % de los días con clima de frío a cálido, el flujo de pacientes es medio (1)
Al menos el 70 % de los días con clima frío, el flujo de pacientes es muy alto (0.99)
La mayoría de los días con clima de cálido a frío, el flujo de pacientes es medio (1) "

Evolutiva n2
"La mayoría de los días de Abril, el flujo de pacientes es medio (1)
Al menos el 70 % de los días con clima frío, el flujo de pacientes es muy alto (0.99)
Al menos el 70 % de los días de Marzo, el flujo de pacientes es medio (0.94)
Al menos 70 % de los días con clima cálido, el flujo de pacientes es bajo o muy bajo (0.92)
La mayoría de los días de Mayo, el flujo de pacientes es medio (1)
La mayoría de los días con clima de cálido a frío, el flujo de pacientes es medio (1)"

Debemos aclarar que en esta ocasión los objetivos utilizados al ejecutar el algoritmo genético han sido los cuatro inicialmente presentados en el modelo de calidad: brevedad, exactitud, cobertura y precisión, pero como ya hemos comentado, hemos realizado la fusión de dos de ellos. Con esta acción hemos pretendido mejorar el desarrollo del algoritmo evolutivo ayudando a la convergencia hacia soluciones mejores.

La fusión se ha realizado como una combinación convexa de dos objetivos, en este caso la exactitud y la precisión:

$$\text{Objetivo fusionado} = ((\beta) * a_d + (1 - \beta) * p)/2$$

donde β es un parámetro que toma valores entre (0,1) y que nos servirá para establecer un orden de importancia entre los objetivos fusionados. Dado que NSGA-II trabaja con minimización, tanto la *exactitud* como la *especificidad* son negativos, de modo que también lo será el objetivo fusionado. En esta caso concreto $\beta = 0{,}5$, con lo que nos queda como una media aritmética de las medidas. En este sentido las medidas de calidad para las soluciones presentadas anteriormente serían los presentado en la Tabla 4.35. Debemos mencionar que esta combinación o la manera de llevarla a cabo es concreta para esta experimentación y que puede variar en función de las necesidades del usuario.

Para leer la tabla debemos recordar que: la *brevedad*, b, es mejor cuanto menor es el valor, la *cobertura*, c_d y *objetivo fusionado* son mejores cuanto mayor es el valor (por ese motivo se presenta el valor absoluto, ya que para minimizarlos se les ha cambiado el signo de positivo a negativo). En la columna que muestra la cobertura, y debido al cumplimento máximo en todos los casos se proporciona el *valor de particionado*. Cuanto mayor es este valor, mayor es la cantidad de puntos en la línea temporal que se alejan del cumplimiento ideal (es decir, ser cubierto por una y solo una etiqueta), de modo que cuanto menor es el valor, mejor es la solución.

Solución	Brevedad	Cobertura (Particionado)	Obj. fusionado
Greedy Primera estrategia (1)	3	365 (0.0109589)	0.35208
Greedy Segunda estrategia (2)	3	365 (0.0109589)	0.350693
Evolutiva n1	4	365 (0.0109589)	0.356735
Evolutiva n2	6	365 (0.0493151)	0.364893

Tabla 4.35: Calidad de las soluciones encontradas para el problema C_B.

Las soluciones Greedy presentan un grado de calidad muy similar, siendo la primera algo mejor si tenemos en cuenta el objetivo fusionado. Para hacer un estudio más profundo de la calidad de las soluciones deberemos comprobar los valores para los objetivos *exactitud* y *especificidad*. En este caso $a_d(1) = 0.946664$ y $p(1) = 0.771201$, mientras que $a_d(2) = 0.92877$ y $p(2) = 0.771657$. A la vista de estos resultados podemos observar que aunque la solución dada por la primera estrategia es más exacta, la solución dada por la segunda estrategia es algo más precisa. En conclusión, podemos decir que el objetivo fusionado guiará al proceso evolutivo de forma adecuada pero, si el usuario desea obtener más información, deberá remitirse a los valores de los objetivos originales.

En cuanto a las dos soluciones evolutivas escogidas podemos decir que ambas son menos breves que las Greedy pero sin embargo presentan valores más altos en el objetivo fusionado. En concreto, la solución n2 presenta el valor más elevado para el objetivo fusionado pero a costa de empeorar el factor de particionado (siempre manteniendo la cobertura máxima).

En resumen vemos que tanto las técnicas Greedy como las evolutivas presentan buenas soluciones, todas ellas optimales y no comparables entre sí en términos de calidad. Mientras que las técnicas Greedy ofrecen soluciones únicas en un intervalo razonable de tiempo, la estrategia evolutiva obtiene un conjunto de soluciones, supuestamente más variadas, que, aunque conllevan el uso de más recursos, pueden ser consideradas en entornos en los que no se necesite un uso interactivo del sistema.

IBEX35

A continuación realizaremos la experimentación anterior pero esta vez con la serie de 621 datos financieros que nos daba el valor del IBEX-35 en la Bolsa de Madrid a lo largo del periodo 2000-2011.

Como contexto y parámetros lingüísticos utilizaremos los definidos en la Sección 4.2.4. Con respecto a los parámetros evolutivos decir que debido a la mayor complejidad del problema hemos decidido aumentar la cantidad de posibilidades en las mutaciones. Se ha introducido el valor 0.3 para la probabilidad de mutación y 0.6 y 0.8 para la de mutación inteligente. Esto dará lugar a un número mayor de combinaciones y por lo tanto a una mayor exploración del espacio de soluciones, sin llegar a hacerlo de manera exhaustiva. Para más detalle acerca de los valores que han tomado los parámetros consultar la Tabla 4.36.

Parámetro	Valores posibles
$popSize$	100 200 400 800 1600 3200
$numGen$	800 400 200 100 50 25
p_c	0.5 0.6 0.7
p_m	0.01 0.05 0.1 0.2 0.3 0.4
p_{m_i}	0.1 0.2 0.6 0.8 1

Tabla 4.36: Parámetros evolutivos usados en la experimentación para el problema $IBEX35$.

Con el fin de poder realizar una comparación entre los resultados obtenidos mediante las técnicas Greedy y la evolutiva, recordaremos a continuación los resultados obtenidos con las primera de ellas. En primer lugar se muestra el resultado obtenido a partir del uso de la primera estrategia (Algoritmo 1) para, a continuación, mostrar el resultado obtenido al aplicar la segunda de las estrategias Greedy (Algoritmo 2).

Finalmente mostraremos dos soluciones, con criterios de calidad comparables a las presentadas anteriormente, seleccionadas de entre las que se encuentran en el primer frente de Pareto de la población de soluciones obtenidas al aplicar el método evolutivo.

Greedy Primera estrategia
"Al menos el 70 % de los días de la década de los 10s, el valor se sitúa entre 10000 y 11000 o entre 9000 y 10000 (1)
Al menos el 70 % de los días del 2000, el valor se sitúa entre 11000 y 12000 o entre 10000 y 11000 (1)
Al menos el 80 % de los días del 2001, el valor se sitúa entre 9000 y 10000 o entre 8000 y 9000 (1)
Al menos el 80 % de los días del 2002, el valor se sitúa entre 8000 y 9000 o entre 6000 y 7000 (0.86)
La mayoría de los días del 2003, el valor se sitúa entre 7000 y 8000 o entre 6000 y 7000 (0.99)
La mayoría de los días del 2004, el valor se sitúa entre 8000 y 9000 o entre 7000 y 8000 (1)
La mayoría de los días del 2005, el valor se sitúa entre 10000 y 11000 o entre 11000 y 12000 (1)
Al menos el 70 % de los días del 2006, el valor se sitúa entre 14000 y 15000 o entre 11000 y 12000 (0.89)
La mayoría de los días del 2007, el valor se sitúa entre 15000 y 16000 o entre 14000 y 15000 (1)
Al menos el 80 % de los días de la primera mitad del 2008, el valor se sitúa entre 13000 y 14000 o entre 12000 y 13000 (1)
Al menos el 80 % de los días de la segunda mitad del 2008, el valor se sitúa entre 11000 y 12000 o entre 9000 y 10000 (0.96)
Al menos el 80 % de los días de la primera mitad del 2009, el valor se sitúa entre 9000 y 10000 o entre 10000 y 11000 (0.84)
La mayoría de los días de la segunda mitad del 2009, el valor se sitúa entre 11000 y 12000 o entre 10000 y 11000 (0.81)"

Greedy Segunda estrategia
"Al menos el 70 % de los días de la década de los 10s, el valor se sitúa entre 10000 y 11000 o entre 9000 y 10000 (1)
Al menos el 70 % de los días del 2000, el valor se sitúa entre 11000 y 12000 o entre 10000 y 11000 (1)
Al menos el 80 % de los días del 2001, el valor se sitúa entre 9000 y 10000 o entre 8000 y 9000 (1)
Al menos el 80 % de los días del 2002, el valor se sitúa entre 8000 y 9000 o entre 6000 y 7000 (0.86)
La mayoría de los días del 2003, el valor se sitúa entre 7000 y 8000 o entre 6000 y 7000 (0.99)
Al menos el 70 % de los días del 2004, el valor se sitúa entre 8000 y 9000 (0.95)
La mayoría de los días del 2005, el valor se sitúa entre 10000 y 11000 o entre 11000 y 12000 (1)
Al menos el 70 % de los días del 2006, el valor se sitúa entre 14000 y 15000 o entre 11000 y 12000 (0.89)
La mayoría de los días del 2007, el valor se sitúa entre 15000 y 16000 o entre 14000 y 15000 (1)
Al menos el 80 % de los días de la primera mitad del 2008, el valor se sitúa entre 13000 y 14000 o entre 12000 y 13000 (1)
Al menos el 80 % de los días de la segunda mitad del 2008, el valor se sitúa entre 11000 y 12000 o entre 9000 y 10000 (0.96)
Al menos el 80 % de los días de la primera mitad del 2009, el valor se sitúa entre 9000 y 10000 o entre 10000 y 11000 (0.84)
Al menos el 70 % de los días de la segunda mitad del 2009, el valor se sitúa entre 11000 y 12000 (1)"

Evolutiva n1
"Al menos el 70 % de los días de comienzos de la década 00s, el valor se sitúa entre 10000 y 11000 o entre 9000 y 10000 (1)
La mayoría de los días de 2003, el valor se sitúa entre 7000 y 8000 o entre 6000 y 7000 (0.99)
Al menos el 70 % de los días de la primera mitad de 2004, el valor se sitúa entre 8000 y 9000 (0.94)
La mayoría de los días de la primera mitad de 2005, el valor se sitúa entre 9000 y 10000 (1)
Al menos el 70 % de los días de 2006, el valor se sitúa entre 14000 y 15000 o entre 11000 y 12000 (0.89)
Al menos el 70 % de la segunda mitad de 2004, el valor se sitúa entre 8000 y 9000 (0.97)
La mayoría de los días de 2007, el valor se sitúa entre 15000 y 16000 o entre 14000 y 15000 (1)
Al menos el 80 % de los días de 2002, el valor se sitúa entre 8000 y 9000 o entre 6000 y 7000 (0.86)
Al menos el 80 % de los días de 2001, el valor se sitúa entre 9000 y 10000 o entre 8000 y 9000 (1)
Al menos el 70 % de los días de 2000, el valor se sitúa entre 11000 y 12000 o entre 10000 y 11000 (1)
Al menos el 80 % de los días de la segunda mitad de 2008, el valor se sitúa entre 11000 y 12000 o entre 9000 y 10000 (0.96)
Al menos el 80 % de los días de la segunda mitad de 2005, el valor se sitúa entre 10000 y 11000 (1)
Al menos el 80 % de los días de la primera mitad de 2008, el valor se sitúa entre 13000 y 14000 o entre 12000 y 13000 (1)
Al menos el 70 % de los días de la primera mitad de 2009, el valor se sitúa entre 9000 y 10000 o entre 8000 y 9000 (1)
Al menos el 70 % de los días de la segunda mitad de 2009, el valor se sitúa entre 11000 y 12000 (1)"

Evolutiva n2
"Al menos el 70 % de los días de comienzos de la década 00s, el valor se sitúa entre 10000 y 11000 o entre 9000 y 10000 (1)
La mayoría de los días de 2003, el valor se sitúa entre 7000 y 8000 o entre 6000 y 7000 (0.99)
Al menos el 70 % de los días de la primera mitad de 2004, el valor se sitúa entre 8000 y 9000 (0.94)
La mayoría de los días de la primera mitad de 2005, el valor se sitúa entre 9000 y 10000 (1)
Al menos el 70 % de los días de 2006, el valor se sitúa entre 14000 y 15000 o entre 11000 y 12000 (0.89)
Al menos el 70 % de la segunda mitad de 2004, el valor se sitúa entre 8000 y 9000 (0.97)
La mayoría de los días de 2007, el valor se sitúa entre 15000 y 16000 o entre 14000 y 15000 (1)
Al menos el 80 % de los días de 2002, el valor se sitúa entre 8000 y 9000 o entre 6000 y 7000 (0.86)
Al menos el 80 % de los días de 2001, el valor se sitúa entre 9000 y 10000 o entre 8000 y 9000 (1)
Al menos el 70 % de los días de 2000, el valor se sitúa entre 11000 y 12000 o entre 10000 y 11000 (1)
Al menos el 80 % de los días de la segunda mitad de 2008, el valor se sitúa entre 11000 y 12000 o entre 9000 y 10000 (0.96)
Al menos el 80 % de los días de la segunda mitad de 2005, el valor se sitúa entre 10000 y 11000 (1)
Al menos el 80 % de los días de la primera mitad de 2008, el valor se sitúa entre 13000 y 14000 o entre 12000 y 13000 (1)
Al menos el 80 % de los días de la primera mitad de 2009, el valor se sitúa entre 9000 y 10000 o entre 8000 y 9000 (0.84)
Al menos el 70 % de los días de la segunda mitad de 2009, el valor se sitúa entre 11000 y 12000 (1)"

Solución	Brevedad	Cobertura (Particionado)	Obj. fusionado
Greedy Primera estrategia (1)	13	621 (0.141707)	0.380442
Greedy Segunda estrategia (2)	13	621 (0.141707)	0.37772
Evolutiva n1	15	621 (0.140097)	0.381818
Evolutiva n2	15	621 (0.140097)	0.380242

Tabla 4.37: Calidad de las soluciones encontradas para el problema $IBEX35$.

Con respecto a los resultados mostrados podemos decir que ambas estrategias Greedy dan soluciones que son similares. Si nos fijamos en los objetivos de la tabla podemos ver que incluso la solución dada por la primera estrategia domina a la solución dada por la segunda. Como ya hemos comentado, esto puede ser engañoso, al ser el objetivo fusionado una combinación convexa de otros dos. Por ejemplo para la solución dada por la primera estrategia tenemos que $a_d(1) = 0.963$ y $p(1) = 0.797611$, mientras que para la segunda tenemos que $a_d(2) = 0.967$ y $p(2) = 0.788054$. Aunque a simple vista parecía que la primera dominaba a la segunda, si descomponemos el objetivo en sus componentes vemos que la primera es mejor con respecto a la precisión pero es peor con respecto a la exactitud. Todo dependerá de cómo hagamos la com-

binación, y si la hacemos como la hemos propuesto aquí, dependerá de la magnitud de los valores y del β escogido.

En cuanto a los resultados evolutivos podemos decir que aunque son menos breves poseen un factor de particionado más bajo y por tanto mejor. Con respecto al valor del objetivo fusionado vemos que es bastante bueno, en concreto la solución n1 supera a ambas soluciones Greedy.

En este ejemplo, volvemos a ver que la estrategia Greedy ofrece resultados de calidad aceptable en un tiempo menor. Aunque no es objeto de estudio de esta tesis entrar en el mejor ajuste de parámetros para la ejecución del algoritmo evolutivo, debemos destacar que dicha tarea consume grandes cantidades de tiempo y recursos, lo que la hacen inviable en entornos en los que el usuario necesite un uso interactivo del sistema.

4.4. Conclusiones

En este capítulo se ha realizado un breve estudio de la complejidad del problema y el espacio de búsqueda asociado al mismo. Dicha complejidad crece cuando nos enfrentamos a problemas reales y se hace imposible explorar el espacio de manera exhaustiva con el fin de encontrar la mejor solución.

En situaciones en las que no es indispensable encontrar la mejor solución o en las que los requerimientos de tiempo no lo permiten, se hacen necesarias técnicas que nos permitan encontrar soluciones lo bastante buenas o satisfactorias.

Existen numerosos enfoques cuando tratamos de encontrar soluciones optimales para un determinado problema de búsqueda, que al fin y al cabo, puede ser tratado como un problema global de búsqueda de óptimos para cumplimiento de unos determinados objetivos, en este caso de calidad.

Para evitar hacer una exploración exhaustiva, en primer lugar hemos optado por un algoritmo Greedy que busque soluciones lo bastante buenas. El diseño de un algoritmo Greedy (y por tanto la personalidad del diseñador) tiene una gran influencia a la hora de establecer el comportamiento del algoritmo, por ello decimos que el Greedy está muy ligado al criterio del diseñador. En este caso, para nosotros, eso se convierte en una ventaja porque junto con la definición del contexto lingüístico nos permite introducir conocimiento del entorno en el algoritmo, de forma que la exploración asegure el buen cumplimiento de los objetivos de calidad del problema.

Debido a la fuerza de los algoritmos evolutivos a la hora de realizar buenas exploraciones de un amplio espacio de soluciones, nos hemos decidido a aplicarlo en nuestro problema. Además, gracias a ese tipo especial de algoritmos evolutivos llamados multi-objetivo, se nos brinda una herramienta que parece adecuada para intentar

optimizar una serie de objetivos de calidad que a priori son complementarios e incluso contradictorios.

El uso de este enfoque evolutivo nos brinda una herramienta para comprobar la bondad de las soluciones obtenidas al usar el enfoque Greedy. Del mismo modo nos permite explorar de manera más amplia el espacio de soluciones para localizar otras posibles buenas soluciones.

A la vista de los resultados obtenidos, podemos asegurar que nuestro modelo basado en la heurística Greedy ofrece muy buenos resultados en un periodo de tiempo muy razonable que se adecua más a las necesidades de usuarios que requieren interacción inmediata con nuestro proceso de resumen de datos.

Como característica positiva del enfoque evolutivo diremos que nos ha permitido asegurar la calidad de nuestras soluciones Greedy y gracias a su poder de exploración nos ha aportado nuevas soluciones. Como desventaja se encuentra la asociada a los recursos consumidos para hacerlo. Recursos de memoria, en tanto que debemos tener poblaciones de individuos almacenadas en memoria, y recursos temporales debido a las diferentes generaciones que se van construyendo. Por otro lado hay que tener en cuenta el tiempo que se consume al intentar hacer un ajuste óptimo de parámetros que se adecuen a cada problema. Ya hemos mencionado que no es objeto de esta memoria entrar en ese aspecto pero sí queremos dejar claro que es un proceso costoso. Por otro lado no está claro hasta qué punto es bueno ofrecerle al usuario un conjunto más grande de posibles soluciones, todas ellas con valores similares para los objetivos de calidad.

En resumen, la propuesta Greedy es ideal para situaciones en las que los usuarios necesiten soluciones *ad hoc* mientras que la propuesta evolutiva brinda una buena herramienta de exploración para ocasiones en las que la inmediatez no sea necesaria, como por ejemplo precálculo de resultados para la construcción de informes previamente definidos sobre almacenes de datos.

Como comentario final, queremos también indicar que, aunque el algoritmo genético empleado genera individuos distribuidos por todo el espacio de búsqueda, las restricciones que hemos establecido fuerzan a que las soluciones finales estén dentro del subespacio que hemos considerado para los algoritmos Greedy. Hemos establecido este criterio para poder realizar una comparación adecuada con las técnicas Greedy, y para valorar la calidad de las soluciones obtenidas con dichas técnicas. Como resultado, pudiera parecer que la eficacia de ambas técnicas es similar, y que por tanto la técnica Greedy, dada su rapidez, es claramente superior.

Sin embargo, sería posible eliminar o suavizar algunas de estas restricciones para encontrar soluciones que, siendo algo peores que las Greedy en el objetivo de cobertura, pudieran ser mucho mejores en otros objetivos y, por tanto, proporcionar

resúmenes significativamente de mayor calidad. Asimismo, el uso de nuestro enfoque basado en NSGA-II ofrece posibilidades muy interesantes a la hora de buscar resúmenes que mezclen distintos aspectos de las series de datos, algunos de los cuales comentaremos en el siguiente capítulo. Por tanto, más allá de la indudable utilidad que ha tenido el enfoque evolutivo a la hora de estudiar teóricamente el problema, presenta diversas ventajas y posibilidades prácticas que serán objeto de investigación por nuestra parte en el futuro.

5 *Generalización y aplicaciones del problema*

"Jamás en la vida encontraréis ternura mejor y más desinteresada que la de vuestra madre"
Honoré de Balzac

El presente capítulo está dedicado a presentar el uso de nuestro modelo en diferentes contextos y para diferentes conjuntos de datos. Mostraremos en este capítulo que una ventaja adicional y un valor añadido muy importante de nuestro modelo es que puede adaptarse fácilmente para resolver distintos problemas de resumen lingüístico, tanto de series de datos temporales como de otros tipos de datos. Cada una de estas posibles adaptaciones y aplicaciones constituye en sí un problema complejo cuya resolución en profundidad está más allá de los objetivos de la presente tesis. Nuestro objetivo ha sido desarrollar cada una de las adaptaciones en un nivel suficiente como para mostrar que es factible realizarlas, dejando como trabajo futuro el avanzar exhaustivamente en cada una de ellas.

En primer lugar mostraremos que, además del resumen del valor de los elementos de la serie de datos, podemos hacer resúmenes basados en otras características de la serie de datos que pueden medirse en sus distintos intervalos temporales. En particular, ilustraremos el resumen de series basado en la tendencia presentada por la serie en cuestión.

En segundo lugar, mostraremos que nuestros algoritmos pueden emplearse asimismo para resolver el problema de la comparación de series de datos, de gran utilidad en muchos campos. Describiremos aquí nuestro enfoque para la comparación lingüística de series de datos temporales, que nos ofrece posibilidades como la de comparar las ventas de dos productos distintos en el mismo intervalo de tiempo, o incluso del mismo producto en intervalos diversos, y recibir las conclusiones en formato de resumen en lenguaje natural. En secciones sucesivas presentaremos dos aproximaciones para realizar la mencionada comparación de series, con diversos métodos que pueden ser usados dependiendo de las necesidades del usuario y el contexto específico.

Finalmente, aunque el interés principal de la presente tesis se centra en el resumen de series de datos temporales, el modelo propuesto puede aplicarse a la descripción de otros tipos de datos, incluso datos complejos como las imágenes digitales. Describiremos cómo es posible adaptar nuestras técnicas de manera sencilla para realizar descripción de datos en general, e ilustraremos este potencial mediante una propuesta

preliminar de aplicación de nuestro modelo para enfrentarnos a la descripción textual de imágenes. Veremos como, empleando los algoritmos propuestos en esta memoria, es posible obtener resúmenes lingüísticos, a partir de una imagen almacenada digitalmente, que nos describan a grandes rasgos lo que en ella aparece, en base a características visuales de bajo nivel.

5.1. Resumen de la tendencia en series de datos

En el capítulo 3 presentábamos un modelo para el resumen de series de datos. Hasta ahora los resúmenes se han obtenido trabajando con el valor de la serie en cada instante de tiempo. Sin embargo, el resumen de la serie basado en el valor no es el único resumen que se puede hacer de los datos. Existen diferentes características de las series que pueden ser resumidas. En esta sección centraremos nuestra atención en el resumen de las tendencias presentes en la serie de datos. La información obtenida de este tipo de resúmenes se puede usar de forma totalmente independiente o bien en combinación con el resumen de la serie basado en el valor de la misma.

Nuestro enfoque para este problema consta de tres pasos:

1. Dada una longitud concreta a considerar para el cálculo de periodos de tiempo, obtener una serie de datos temporal, a partir de la serie original, que asigne a cada instante de tiempo una medida adecuada de la tendencia en el periodo que comienza en ese instante de tiempo.
2. Definir un marco linguístico apropiado para el problema de la descripción lingüística de la tendencia.
3. Aplicar las técnicas algorítmicas descritas en el capítulo anterior para obtener un resumen de la tendencia en la serie transformada obtenida en el primer paso, usando el marco lingüístico definido en el segundo paso.

Veamos a continuación un enfoque concreto para realizar cada uno de los pasos anteriores. Como hemos indicado, nuestro objetivo no ha sido resolver este problema de manera exhaustiva, ya que los dos primeros pasos que acabamos de detallar podrían resolverse de múltiples formas. En este apartado presentamos un enfoque concreto y los resultados obtenidos con el mismo.

5.1.1. Obtención de la serie temporal: la tendencia en cada instante de tiempo.

Como hemos indicado, nuestro objetivo en este paso es obtener una serie de datos que refleje tendencias en periodos de tiempo definidos. Como resultado obtendremos una nueva serie de datos donde a cada valor de tiempo se le asignará el valor de la tendencia de la serie calculado sobre un periodo de longitud habitualmente prefijada, y cuyo punto de inicio es el valor de tiempo mencionado. Asumiendo que los instantes de tiempo están equidistribuidos en la serie original TS a intervalos de tiempo t, podemos considerar intervalos de longitud $k * t$ que comiencen en el instante t_i y acaben en t_{i+k}. La serie resultante tendrá en general un número menor de datos que la original, $m - k$.

A la hora de medir la tendencia, hemos considerado utilizar el ángulo que forma la recta que une los puntos extremos del periodo de tiempo en una versión escalada de la serie. Dicho ángulo estará en el intervalo $(-\pi/2, \pi/2)$. La serie resultante será $STend = \{< STend_1, t_1 >, \ldots, < STend_{m-k}, t_{m-k} >\}$, con

$$STend_i = STend(t_i) = \arctan\left(\frac{TS(t_{i+k}) - TS(t_i)}{K * (t_{i+k} - t_i)}\right) = \arctan\left(\frac{v_{i+k} - v_i}{K * (t_{i+k} - t_i)}\right) \tag{5.1}$$

El factor de escala K es necesario ya que el ángulo resultante en el cálculo de la tendencia dependerá de cual sea la escala de tiempo empleada y su relación con la escala de los valores de la serie. Si consideramos que una unidad en el eje del tiempo representa un día, el resultado obtenido será muy diferente al obtenido si consideramos que esa misma unidad representa un segundo. Por tanto, lo que hacemos con la expresión es calcular una tendencia *relativa a una determinada escala temporal*. Esto último, por otra parte, no deja de ser intuitivo. Así, la tendencia de cambio medida al segundo será muy diferente a la tendencia medida en días o años. El valor de K representará cual es el incremento (resp. decremento) que debe producirse en el valor de la serie en un periodo $t * k$ para que la variación medida en ángulo sea de $\pi/4$ (resp. $-\pi/4$), que tomamos como valor de incremento proporcional al tiempo de que resulta normal al usuario. También podemos interpretarlo como medida estándar de incremento, a medio camino entre constante (incremento 0) y extremo. Por ejemplo, si utilizamos periodos medidos en días y de longitud un día, y pensamos que un cambio alrededor de 20 unidades por día supone un claro incremento, podríamos emplear $K = 20$.

5.1.2. Un marco lingüístico para la tendencia.

El siguiente paso es determinar el marco lingüístico adecuado. Generalmente, y a no ser que cambien las necesidades, la jerarquía temporal seguirá siendo la misma que cuando hacíamos resumen por valor. En cambio, la partición de la variable bajo estudio sí que sufrirá un cambio importante. Necesitaremos un conjunto de etiquetas que en lugar de describir lingüísticamente el valor (*alto*, *muy bajo*, etc.), describa la tendencia. Una posibilidad es utilizar las etiquetas que se muestran en la tabla 5.1. En dicha tabla, las etiquetas están expresadas de forma que cada unidad representa un ángulo de $\pi/32$, es decir, que 16 representa $\pi/2$, 8 representa $\pi/4$, etc. Hay que destacar que este mismo conjunto de etiquetas puede utilizarse con mínimos cambios cualquiera que sea el rango de valores de la variable medida en la serie temporal, aunque por supuesto pueden definirse conjuntos alternativos de etiquetas según las necesidades.

Las etiquetas nos permiten describir lingüísticamente la nueva serie, donde apa-

Etiqueta	Definición
altamente decreciente	(-16, -16, -12, -10)
decreciente	(-12, -10, -6, -4)
suavemente decreciente	(-6, -4, -2, -1)
casi constante	(-2, -1, 1, 2)
suavemente creciente	(1, 2, 4, 6)
creciente	(4, 6, 10, 12)
altamente creciente	(10, 12, 16, 16)

Tabla 5.1: Una posible partición del dominio de la medida de tendencia de la Ecuación (5.1). Cada unidad corresponde a un ángulo de $\pi/32$.

recerán valores negativos y positivos. Los valores negativos representan que la serie original es decreciente en ese periodo (valor del instante final menor que valor de instante inicial) mientras que los positivos representan que es creciente. Valores muy cercanos al cero nos informan a su vez de que las tendencias son más o menos constantes.

Podemos destacar que en este caso no se ha usado una partición de etiquetas de soporte semejante sino que hemos optado por una etiqueta central más ajustada, y tres periodos más amplios a cada lado, uno de ellos representando el concepto creciente (resp. decreciente) centrado en el valor $\pi/4$ que hemos destacado anteriormente.

El siguiente paso es aplicar los algoritmos que hemos propuesto. Para ilustrar este paso, vamos a aplicar la adaptación descrita a un ejemplo concreto en el siguiente apartado.

5.1.3. Tendencias en el ejemplo centro de salud C_B

De nuevo, y con objeto de mostrar mediante un ejemplo lo expuesto anteriormente, rescatamos el cubo con los datos relativos a los centros de salud. En la figura 4.15 dentro del capítulo 4, ya se introdujo la serie de datos que describía "la afluencia masculina a un centro C_B durante un año completo". Volveremos a trabajar con dicha serie, pero esta vez para realizar un resumen de las tendencias que en ella aparecen.

En este ejemplo, la escala de tiempo está expresada en dias. Hemos considerado periodos de tamaño 1, y un factor de escala de K=*8*. En la Figura 5.1 podemos ver la serie de datos que representa "la variación en la afluencia masculina a un centro C_B durante un año completo". Además de la serie se encuentran representadas las particiones en ambos ejes que representan parte del marco lingüístico considerado.

Como ya se ha comentado, la jerarquía de particiones que describe la dimensión temporal no sufre ningún cambio ya que la dimensión temporal continúa siendo la misma y por lo tanto la forma de particionar el tiempo también (siempre y cuando no

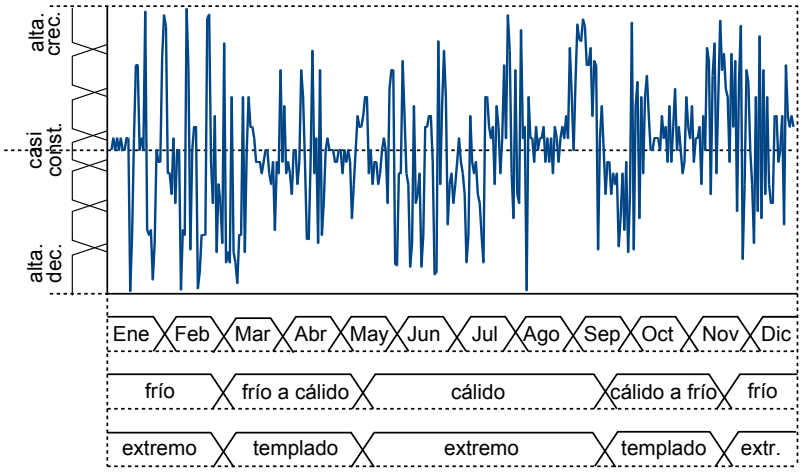


Figura 5.1: Variación en la afluencia masculina al centro de salud C_B durante un año.

cambien los requerimientos del problema). Volveremos pues a tomar las particiones representadas en la Tabla 4.2 para trabajar en este ejemplo. En lo referente a la partición de etiquetas lingüísticas que formarán parte de la componente A en las sentencias cuantificadas, utilizaremos las etiquetas definidas en la tabla 5.1.

En cuanto al subconjunto coherente de cuantificadores usaremos el conjunto definido para el problema de resumen del Capítulo 4 donde trabajábamos con los datos relativos a los valores tomados por el IBEX-35 en un periodo de tiempo. Dichos cuantificadores se encuentran expuestos en la Tabla 4.33 y ya fueron usados también en el capítulo 4.

Para el resto de los parámetros que completan el contexto lingüístico: el umbral de cumplimiento de las sentencias cuantificadas, τ, se ha igualado a 0.7. Respecto a los límites tenemos que $Qlim_i = 3$ y $Glim_i = 3$ para todos los niveles i; por lo tanto se podrán usar tres cuantificadores comenzando desde el más estricto y se permitirán parejas de etiquetas.

Una vez calculada la serie que refleja la tendencia y definido el marco lingüístico, podemos aplicar nuestro modelo para obtener un resumen de las tendencias de la serie que, una vez postprocesada, quedaría como sigue (con primera estrategia Greedy):

> "Al menos el 70 % de los días con clima frio y en Agosto, la variación es altamente creciente o suavemente creciente. Al menos el 70 % de los días en Abril y Mayo, la variación es casi constante o suavemente creciente. Al menos el 70 % de los días en Septiembre, la variación es altamente decreciente, suavemente decreciente, o suavemente creciente. Al menos el 70 % de los días en Octubre y Noviembre, la variación es altamente decreciente, casi constante, o altamente creciente. El resto del año, la variación es altamente variable."

5.1.4. Discusión y trabajo futuro

Como hemos visto, es posible aplicar las técnicas de resumen lingüístico desarrolladas en los capítulos anteriores para otras características de las series de datos, como es el caso de la tendencia. Otras características como estacionalidad, etc. pueden ser potencialmente analizadas empleando nuestros modelos de resumen, siendo también muy interesante la posibilidad de realizar resúmenes que consideren diversas características al mismo tiempo, bien presentando todas ellas, bien destacando para cada periodo de tiempo solo las más significativas o interesantes según la aplicación.

Como podemos ver, el problema de medir la tendencia de la serie de datos, así como las otras posibilidades que hemos esbozado en el párrafo anterior, constituyen una línea donde queda una gran cantidad de trabajo por hacer, que será objeto de nuestro interés en el futuro.

La calidad de los resúmenes en el caso de la tendencia se ve afectada por nuevos parámetros, como la longitud de los periodos de tiempo empleados y el factor de escala K. Será necesario considerar por tanto técnicas de exploración que incorporen distintas longitudes de periodos de tiempo y los correspondientes valores de escala.

Otro aspecto interesante que consideraremos es incorporar al resumen obtenido para periodos breves de tiempo la información correspondiente al intervalo de la jerarquía utilizado en la sentencia, es decir, proporcionar sentencias del tipo "La mayoría de los días en clima extremo, la variación del flujo de pacientes es altamente creciente o altamente decreciente, *siendo globalmente casi constante en todo el periodo*". Esta información requiere la definición de una medida de variación que tenga en cuenta los grados de pertenencia de los distintos instantes de tiempo al intervalo temporal difuso.

Por otra parte, la medida de la tendencia que hemos calculado se basa en un criterio absoluto, ya que el factor K corresponde al número absoluto de unidades que debe aumentar el valor de la serie. Sin embargo, en muchas ocasiones se considera el aumento o disminución porcentual a la hora de hablar de incrementos o decrementos moderados o grandes. Esta es otra alternativa interesante a explorar. Respecto a la diferencia entre valoraciones absolutas y relativas de la variación, hablaremos en el siguiente apartado, pero dentro de un contexto distinto, como lo es el problema de la comparación de series de datos temporales. Las técnicas que expondremos en el siguiente capítulo son potencialmente útiles para afrontar la resolución del problema de la medida de tendencias con factores de escala que recojan esa semántica de variaciones porcentuales.

5.2. Comparación de series de datos

La comparación de series de datos temporales es muy importante hoy en día. Podemos usar la comparación de series temporales en *tecnología* cuando estudiamos el comportamiento de dos materiales a lo largo del tiempo, en *ciencias de la salud* cuando comparamos la presión sanguínea o la temperatura de dos pacientes a lo largo de un periodo de tiempo, o en *ciencias ambientales* cuando comparamos la concentración de dos tipos de polen en la atmósfera.

Otro campo en el que la comparación de series de tiempo es muy importante, si no vital, es el *ámbito empresarial y económico*. En un mundo en el que el consumo lo gobierna todo, es esencial para las empresas obtener conocimiento acerca de los productos que venden, así como de los que venden sus competidoras, y tener la posibilidad de comparar las series que los representan. De esta forma una empresa podría comparar la venta de un determinado producto X en dos periodos de tiempo distintos, o comparar las ventas de los productos X e Y en un mismo periodo. Se podrían también comparar la evolución de las bolsas de diversos países en un periodo conflictivo o un periodo de calma, etcétera.

La comparación lingüística de series de datos económicos ha sido el tema central de trabajos como [73,74]. En ellos se trata de describir la diferencia entre diferentes valores obtenidos de fondos de inversión e índices de la bolsa de Varsovia - WIG y WIG20 (Warsaw Stock Exchange). La descripción de la comparación se realiza basándose en las tendencias presentes en los distintos segmentos de las series temporales, pero podría realizarse también en términos del valor de las series, u otras características relevantes.

En general, vamos a considerar dos series, TS_1 y TS_2 definidas sobre el mismo conjunto de instantes de tiempo. $TS_i(t_j)$ representará el valor de la variable V de la serie TS_i en el instante t_j.

De manera similar al caso de las tendencias, nuestro enfoque para este problema consta de tres pasos:

1. Obtener una serie de datos temporal, a partir de las series originales a comparar, que asigne a cada instante de tiempo una medida adecuada de la diferencia entre ambas.

2. Definir un marco lingüístico apropiado para el problema de la descripción lingüística de la medida definida en el paso anterior.

3. Aplicar las técnicas algorítmicas descritas en el capítulo anterior para obtener un resumen de la serie obtenida en el primer paso, usando el marco lingüístico definido en el segundo paso.

Una vez más, queremos destacar que nuestro objetivo no ha sido resolver este problema de manera exhaustiva. Existen muchas formas posibles de establecer la semejanza entre series temporales, en términos del valor, de la tendencia, o de la combinación de éstas y/u otras características relevantes de las series. También es posible considerar medidas simétricas de semejanza, o medidas no simétricas de variación de una serie con respecto a otra. En los siguientes apartados mostraremos algunas posibilidades de afrontar este problema, así como algunos resultados ilustrativos de su aplicación. Concretamente se mostrarán dos enfoques mediante los cuales afrontar el proceso de descripción de la comparación. En el primero de ellos se realiza una comparación basada en el valor de la serie a lo largo del tiempo, mientras que el segundo se centra en los cambios locales de la misma. Para cada caso mostraremos asimismo el marco lingüístico que hemos definido, y algunos ejemplos.

5.2.1. Comparación basada en valor

La comparación basada en valor se basa en la diferencia entre los valores de la variable entre ambas series temporales en un mismo instante de tiempo. Dadas las series temporales TS_1 y TS_2, dicha diferencia puede representarse como una nueva serie temporal. En el siguiente apartado presentaremos tres alternativas diferentes mediante las que afrontar el cálculo de dicha serie temporal haciendo uso de diferentes matices semánticos: una de ellas absoluta y otras dos relativas.

Estrategias de obtención de la serie comparación basada en valor

La primera alternativa define la nueva serie ΔTS como la diferencia, en términos absolutos, entre las dos series originales TS_1 y TS_2. Formalmente,

Definición 5.1 (Serie diferencia absoluta) *Sean TS_1 y TS_2 dos series temporales definidas sobre la misma variable V en un cierto periodo de tiempo. La serie diferencia absoluta $\Delta TS_{abs,TS_1,TS_2}$ se define como*

$$\Delta TS_{abs,TS_1,TS_2}(t_i) = TS_1(t_i) - TS_2(t_i) \tag{5.2}$$

para todo t_i en el dominio temporal.

Como alternativa a este primer método, el cálculo de la serie diferencia puede ser realizado también en términos relativos. En este sentido, dos nuevos métodos de cálculo de ΔTS pueden ser definidos.

Definición 5.2 (Serie diferencia relativa global) *Sean TS_1 y TS_2 dos series temporales definidas sobre la misma variable V en un cierto periodo de tiempo. La serie*

diferencia relativa global $\Delta TS_{global,TS_1,TS_2}$ *se define, para todo* t_i *del dominio temporal, como,*

$$\Delta TS_{global,TS_1,TS_2}(t_i) = \begin{cases} 0, & si\ TS_1(t_i) - TS_2(t_i) = 0 \\ \dfrac{TS_1(t_i) - TS_2(t_i)}{M - m}, & en\ otro\ caso \end{cases} \tag{5.3}$$

donde M *es el máximo global de* TS_1 *y* TS_2*, y* m *es el mínimo global de* TS_1 *y* TS_2*.*

Definición 5.3 (Serie diferencia relativa local) *Sean* TS_1 *y* TS_2 *dos series temporales definidas sobre la misma variable* V *en un cierto periodo de tiempo. La serie diferencia relativa local* $\Delta TS_{local,TS_1,TS_2}$ *se define, para todo* t_i *en el dominio temporal, como,*

$$\Delta TS_{local,TS_1,TS_2}(t_i) = \begin{cases} 0, & si\ TS_1(t_i) - TS_2(t_i) = 0 \\ \dfrac{TS_1(t_i) - TS_2(t_i)}{max(TS_1(t_i), TS_2(t_i)) - m}, & en\ otro\ caso \end{cases} \tag{5.4}$$

donde m *es el mínimo global de* TS_1 *y* TS_2*.*

Estas dos últimas definiciones afrontan el problema en términos relativos, pero con dos aproximaciones semánticamente diferentes:

- $\Delta TS_{global,TS_1,TS_2}(t_i)$ es la diferencia, en términos relativos, entre las dos series originales en el punto t_i, de acuerdo a la escala de valores de las series originales (esto es, la diferencia entre el el máximo y el mínimo de las dos series).
- $\Delta TS_{local,TS_1,TS_2}(t_i)$ es la diferencia, también en términos relativos, entre las dos series originales en el punto t_i, pero ahora de acuerdo a la escala de valores en un punto dado en las dos series originales (esto es, la diferencia entre el máximo valor en un punto dado y el mínimo global).

A partir de este momento, y debido a razones de claridad y simplicidad, nos referiremos a $\Delta TS_{abs,TS_1,TS_2}$, $\Delta TS_{global,TS_1,TS_2}$, y $\Delta TS_{local,TS_1,TS_2}$ como ΔTS_{abs}, ΔTS_{global}, y ΔTS_{local}, respectivamente.

De cara a describir las diferencias entre las tres alternativas planteadas, en la Figura 5.2 se representa el comportamiento de una variable dada V a lo largo del tiempo T en dos series temporales distintas.

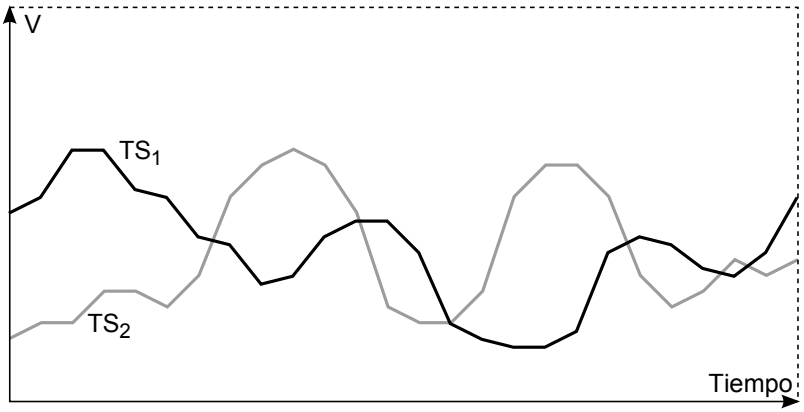


Figura 5.2: Series de datos temporales TS_1 y TS_2.

La elección entre una estrategia u otra de las tres propuestas dependerá de las necesidades específicas del usuario y el problema en una situación particular. ΔTS_{abs} es la única elección posible si el usuario se encuentra interesado en el análisis de la diferencia entre las series en términos absolutos. La Figura 5.3 muestra un ejemplo de uso de esta primera alternativa. Como se puede ver, la nueva serie se mueve en el mismo rango de valores que las series originales.

Sin embargo, si estamos interesados en el análisis de series en términos relativos, deberíamos considerar el uso de alguna de las dos alternativas propuestas para esta situación. La Figura 5.4 muestra las series ΔTS_{global} y ΔTS_{local} obtenidas para el mismo ejemplo.

La Figura 5.4 nos ilustra con un ejemplo la diferencia existente entre ambas estrategias: mientras que en ΔTS_{global} la misma diferencia entre las series originales siempre produce el mismo valor *relativo* en la nueva serie, en ΔTS_{local} cuanto más bajos son los valores originales, mayor es la relevancia de la diferencia *relativa*. Ambas situaciones pueden verse en las parejas de puntos a y b, y c y d, respectivamente. En este sentido, ΔTS_{abs} se comporta como ΔTS_{global} pero en diferente escala (ver puntos a y b en la Figura 5.3).

Marco lingüístico para la comparación basada en valor

El marco lingüístico a emplear dependerá de la estrategia que se haya escogido para obtener la serie diferencia:

- En el caso de ΔTS_{abs}, el dominio de la serie temporal es $[d_1, d_2]$ donde d_1 y d_2 son el mínimo y el máximo de las diferencias entre ambas series en cada punto, respectivamente. Para poder apreciar de manera más clara la diferencia,

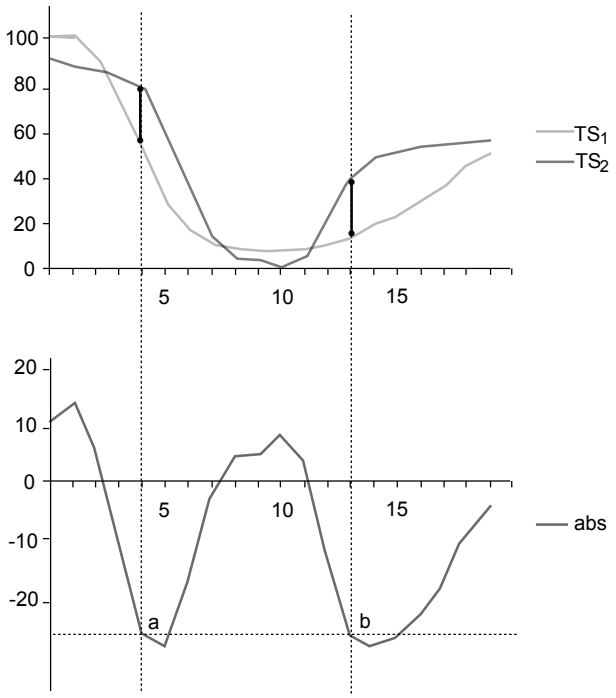


Figura 5.3: Series originales y ΔTS_{abs}.

y para definir un conjunto simétrico de etiquetas lingüísticas que la describan, consideraremos habitualmente como dominio un intervalo más amplio dado por $[-max, max]$, siendo $max \geq$ máx$\{|d_1|, |d_2|\}$. La Figura 5.5 muestra un conjunto de etiquetas de este tipo.

- Por contra, en el caso de ΔTS_{global} y ΔTS_{local}, el dominio subyacente es siempre el mismo, el intervalo $[-1, +1]$, independientemente de las series iniciales con las que estemos trabajando. La universalidad del dominio subyacente hace posible poder comparar de forma sencilla los resúmenes obtenidos para diferentes problemas. La Figura 5.6 muestra un posible conjunto de etiquetas para este intervalo.

Debemos hacer una puntualización con respecto a las etiquetas presentadas en la Figura 5.6. A pesar de que ΔTS_{global} y ΔTS_{local} puedan compartir la misma partición, la interpretación de cada etiqueta difiere. Dicha interpretación depende fuertemente de la semántica implícita en cada uno de los métodos relativos. Mientras con ΔTS_{global} una cierta etiqueta siempre *sugerirá* el mismo rango absoluto en la diferencia, con ΔTS_{local} el rango absoluto *sugerido* dependerá del punto temporal.

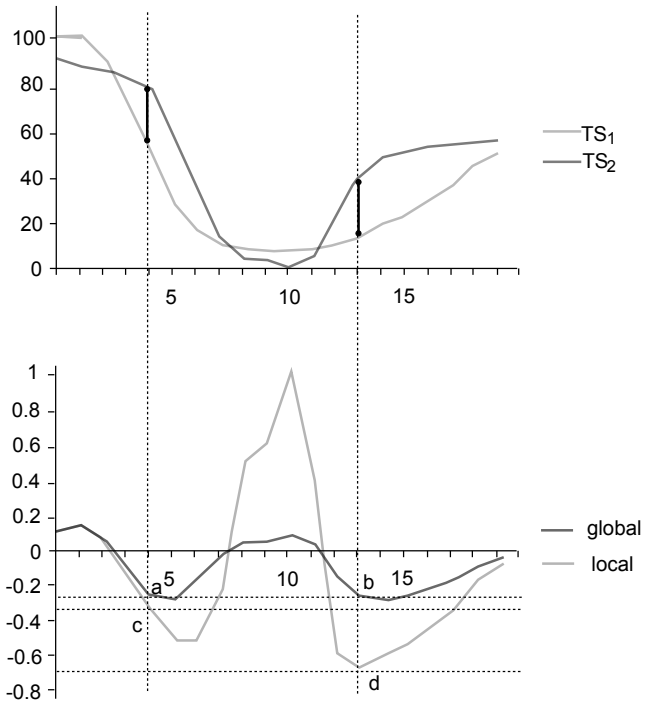


Figura 5.4: Series originales, ΔTS_{global} y ΔTS_{local}.

Ejemplo

Recuperemos de nuevo el cubo con los datos relativos a distintos centros de salud, y dentro del mismo a las series ya utilizadas para describir la afluencia masculina en el centro C_A y el centro C_B durante un año completo (Figuras 4.4 y 4.15 respectivamente). En la Figura 5.7 podemos ver ambas series y el marco lingüístico que ya se ha usado con anterioridad.

Como primer paso para la comparación de estas series se debe obtener una nueva serie que contenga información sobre la diferencia entre ellas. A continuación y dependiendo de la técnica elegida para hacerlo se debe construir la partición que describa la variable (en este caso la diferencia y no el valor).

En la Tabla 5.2 se muestran particiones tanto para usar con el enfoque absoluto como con los relativos. Como se puede observar las etiquetas para el enfoque absoluto están definidas en un rango de $[-500, 500]$, mientras que en el caso relativo lo hacen en $[-1, 1]$.

El resto de parámetros lingüísticos se han inicializado de la siguiente manera: el

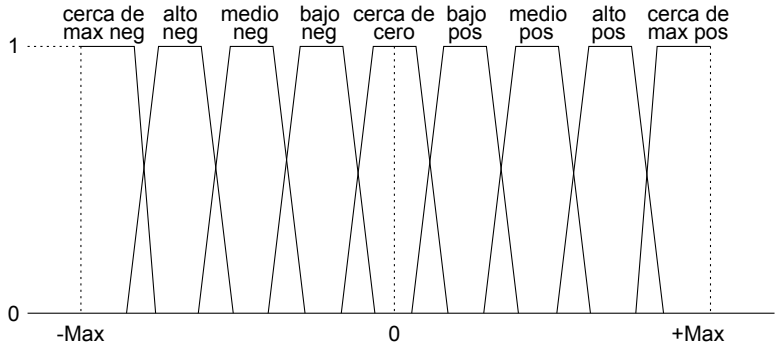


Figura 5.5: Ejemplo de dominio lingüístico para ΔTS_{abs}.

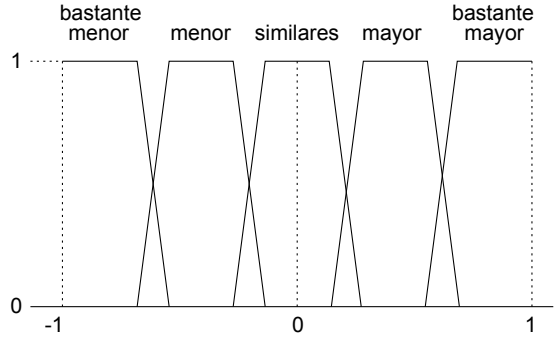


Figura 5.6: Ejemplo de dominio lingüístico para ΔTS_{global} y ΔTS_{local}.

umbral τ toma valor 0.7, y los límites $Qlim_i = Glim_i = 2$ para todos los niveles i de la jerarquía. El subconjunto de cuantificadores es el mostrado en la Tabla 4.3 presentada en el anterior capítulo, y que también usamos para resumir uno de los conjuntos de datos del ejemplo. Finalmente, se usará el Algoritmo 1 de los dos Greedy propuestos.

Las Figuras 5.8, 5.9 y 5.10 muestran las series diferencia para este ejemplo en términos absolutos, relativos globales y relativos locales respectivamente. Podemos apreciar que, como era de esperar, la primera y la segunda son idénticas pero en rangos diferentes, estando la segunda de ellas escalada al rango [-1, 1]. Por contra, la segunda y la tercera, aún compartiendo rango y por tanto partición de etiquetas, muestran diferencias en periodos donde los valores de la serie diferencia son menores.

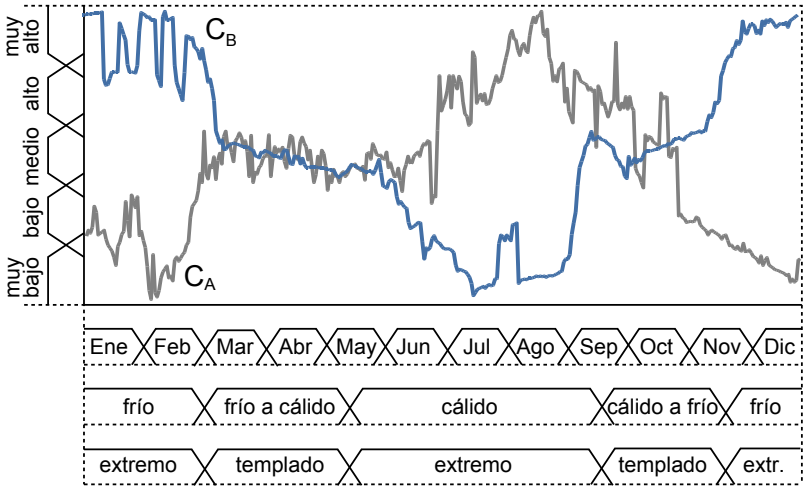


Figura 5.7: Afluencia de pacientes masculinos a los centros C_A y C_B durante un año.

Figura 5.8: Diferencia absoluta entre C_B y C_A durante un año.

Etiqueta	Definición
negativo muy alta	(-500, -500, -410, -390)
negativo alta	(-410, -390, -310, -290)
negativo media	(-310, -290, -210, -190)
negativo baja	(-210, -190, -110, -90)
negativo muy baja	(-110, -90, 0, 0)
positivo muy baja	(0, 0, 90, 110)
positivo baja	(90, 110, 190, 210)
positivo media	(190, 210, 290, 310)
positivo alta	(290, 310, 390, 410)
positivo muy alta	(390, 410, 500, 500)

a) Diferencia absoluta

Etiqueta	Definición
mucho mayor	(-1, -1, -0.8, -0.6)
mayor	(-0.8, -0.6, -0.3, -0.1)
similar	(-0.3, 0, 0, 0.3)
menor	(0.1, 0.3, 0.6, 0.8)
mucho menor	(0.6, 0.8, 1, 1)

b) Diferencia relativa

Tabla 5.2: Partición del dominio de la variable en el caso de comparación basada en valor de los centros C_A y C_B.

Figura 5.9: Diferencia relativa global entre C_B y C_A durante un año.

Figura 5.10: Diferencia relativa local entre C_B y C_A durante un año.

El resultado final de la comparación de las series C_B con respecto a C_A en términos de diferencia absoluta (Figura 5.8), usando los parámetros mostrados es el siguiente:

"Aproximadamente más del 70 % de los días en clima cálido, la diferencia en el flujo de pacientes es negativa muy baja o positiva muy baja. (0.81)
La mayoría de los días en Enero, la diferencia en el flujo de pacientes es positiva media o positiva alta (0.99)
En Febrero, la diferencia en el flujo de pacientes presenta variabilidad
Aproximadamente más del 70 % de los días en Noviembre, la diferencia en el flujo de pacientes es positiva baja o positiva media (0.99)
Aproximadamente más del 70 % de los días en Diciembre, la diferencia en el flujo de pacientes es positiva alta o positiva muy alta (0.78)
La mayoría de los días en Mayo, la diferencia en el flujo de pacientes es negativa muy baja o positiva muy baja (1)
La mayoría de los días en Junio, la diferencia en el flujo de pacientes es negativa muy baja o positiva muy baja (1)
La mayoría de los días en Julio, la diferencia en el flujo de pacientes es negativa alta o negativa media (0.87)
Aproximadamente más del 70 % de los días en Agosto, la diferencia en el flujo de pacientes es negativo muy alto o negativo alto (0.87)
En Septiembre, la diferencia en el flujo de pacientes presenta variabilidad"

Veamos ahora los resultados obtenidos al usar los métodos relativos. Comencemos por obtener el resumen de la diferencia en términos relativos absolutos de la serie C_B con respecto a la serie C_A (Figura 5.9):

"La mayoría de los días en clima frío, la serie C_B es mucho mayor o mayor que C_A (0.73)
La mayoría de los días en Mayo, la serie C_B es similar a C_A (0.77)
Aproximadamente más del 70 % de los días en clima templado, la serie C_B es mayor o similar que C_A (0.77)
Aproximadamente más del 70 % de los días en Agosto, la serie C_B es menor o mucho menor que C_A (0.72)
En Septiembre, la diferencia en el flujo de pacientes presenta variabilidad
En Junio, la diferencia en el flujo de pacientes presenta variabilidad
En Julio, la diferencia en el flujo de pacientes presenta variabilidad"

Mientras que si calculamos la diferencia en términos relativos locales (Figura 5.10) obtenemos:

"La mayoría de los días en clima frío, la serie C_B es mucho mayor o mayor que C_A (0.75)
La mayoría de los días en Noviembre, la serie C_B es mucho mayor o mayor que C_A (0.74)
Aproximadamente más del 70 % de los días en clima cálido, la serie C_B es menor o mucho menor que C_A (0.72)
Aproximadamente más del 70 % de los días en clima frío a cálido, la serie C_B es mayor o similar que C_A (0.70)
Aproximadamente más del 70 % de los días en Octubre, la serie C_B es similar o menor que C_A (0.72)
En Septiembre, la diferencia en el flujo de pacientes presenta variabilidad."

Con el fin de apreciar mejor las diferencias en cuanto a resumen entre los tres métodos, recuperaremos las gráficas con sombreados en las intersecciones entre las etiquetas que se han utilizado en cada una de las sentencias cuantificadas de los resúmenes.

La Figura 5.11 muestra la representación gráfica del resumen encontrado para la serie diferencia absoluta. En cambio, las Figuras 5.12 y 5.13 representan el resumen obtenido al resumir las series diferencia relativa global y local, respectivamente.

Mediante estas figuras ilustramos algunas ideas que se han introducido con anterioridad:

- Existe una diferencia clara entre las descripciones obtenidas mediante el uso de los enfoques absoluto y relativo. Esto se debe a la gran influencia que tienen sobre el resumen las distintas particiones de etiquetas lingüísticas usadas por los diferentes enfoques.

- Como se puede observar en las Figuras 5.9 y 5.10 la diferencia relativa local obtiene resultados más acentuados que la global en periodos como por ejemplo Junio, Julio y Noviembre. Como consecuencia, los enfoques relativos obtienen diferentes resúmenes, ver Figuras 5.12 y 5.13.

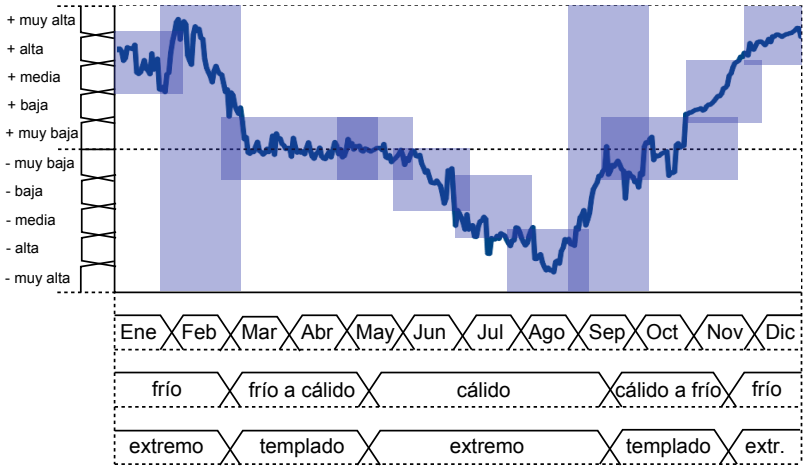


Figura 5.11: Resumen de la diferencia absoluta entre C_B y C_A durante un año.

5.2.2. Comparación basada en tendencias

En este apartado planteamos una comparación lingüística entre las series mediante el uso de cambios locales en grado y signo.

En los siguientes apartados estudiaremos la definición de las series que definen las diferencias en grado y signo, así como el marco lingüístico apropiado, y algunos ejemplos. Como en casos anteriores, existen muchas alternativas posibles para definir y utilizar estos aspectos, y lo que presentamos es una forma específica para ilustrar las posibilidades que ofrece nuestro modelo.

Definición de la serie temporal: dinámicas de cambio

Para la comparación de series temporales en función de la tendencia partiremos de las series temporales que definen la tendencia de cada serie, tal y como se calcularon en el apartado 5.1.1, Ecuación (5.1). A la hora de comparar dichas tendencias consideraremos dos aspectos diferenciados, cada uno de los cuales puede representarse mediante una serie: signo y magnitud de la variación. En el caso del signo, la nueva serie se calculará como $STend_{TS_1}(t_i) * STend_{TS_2}(t_i)$ en cada punto t_i, mientras que en el caso de la magnitud la expresión que utilizaremos será $||STend_{TS_1}(t_i)| - |STend_{TS_2}(t_i)||$. Es importante destacar en esta última que, para que la diferencia pueda ser significativa, es importante considerar un mismo factor de escalado para ambas series.

Marco lingüístico para las dinámicas de cambio

En este caso, para la descripción lingüística de la dinámica de cambio no vamos a utilizar una sola partición de un dominio numérico, sino las siguientes etiquetas

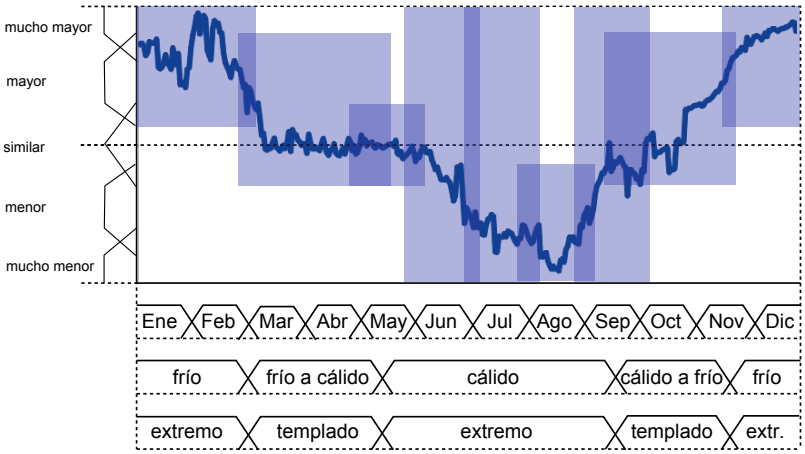


Figura 5.12: Resumen de la diferencia relativa global entre C_B y C_A durante un año.

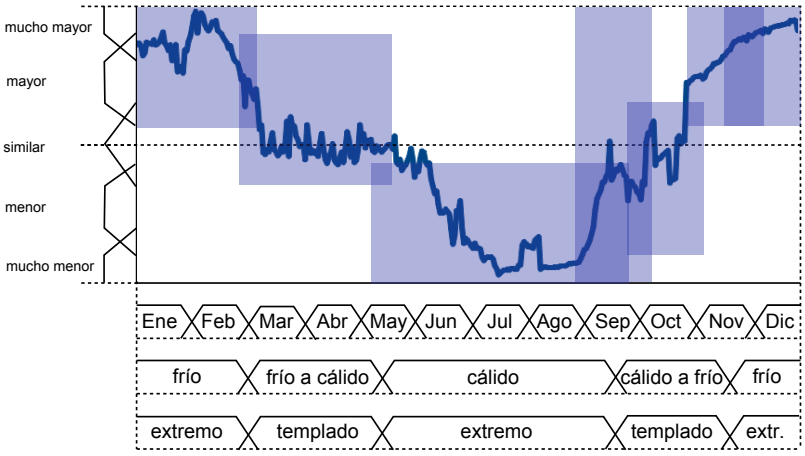


Figura 5.13: Resumen de la diferencia relativa local entre C_B y C_A durante un año.

lingüísticas:

Definición 5.4 (Etiqueta mismo signo) *Sean TS_1 y TS_2 dos series temporales definidas sobre la misma variable V en un cierto periodo de tiempo $t_1, \ldots, t_m$. El grado de pertenencia a la etiqueta "mismo signo", SS_{TS_1,TS_2}, se define para todo t_i como,*

$$SS_{TS_1,TS_2}(t_i) = \begin{cases} 1, & si\ (STend_{TS_1}(t_i) * STend_{TS_2}(t_i)) \geq 0 \\ 0, & en\ otro\ caso \end{cases} \tag{5.5}$$

De la misma forma, podemos definir la etiqueta "signo diferente" como $DS_{TS_1,TS_2}(t_i) = 1 - SS_{TS_1,TS_2}(t_i)$. Estas dos etiquetas son por definición conceptos crisp.

Si sólo nos fijamos en la magnitud del cambio, podemos definir dos etiquetas adicionales:

Definición 5.5 (Variación similar) *Sean TS_1 and TS_2 dos series temporales definidas sobre la misma variable V en un cierto periodo de tiempo $t_1, \ldots, t_m$. El grado de cumplimiento de la etiqueta "variación similar", $SV_{TS_1,TS_2}(t_i)$ en el punto t_i entre las series TS_1 y TS_2 se define como,*

$$SV_{TS_1,TS_2}(t_i) = 1 - \frac{||STend_{TS_1}(t_i)| - |STend_{TS_2}(t_i)||}{\pi/2} \tag{5.6}$$

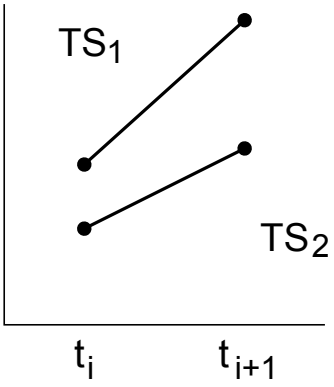
A través del cálculo del grado de variación similar obtendremos la similitud observada entre los dos ángulos responsables del cambio local en un punto t_i dado, en ambas series, sin tener en cuenta el signo. De la misma forma, podemos definir el grado de la etiqueta "variación diferente" como $DV_{TS_1,TS_2}(t_i) = 1 - SV_{TS_1,TS_2}(t_i)$.

En la Figura 5.14 se muestra una serie de posibilidades que podremos tener en un punto dado respecto al signo y la variación del cambio local. Aunque en los ejemplos la variación parece crisp por cuestiones de claridad y simplicidad, debemos decir que las definiciones anteriores son difusas.

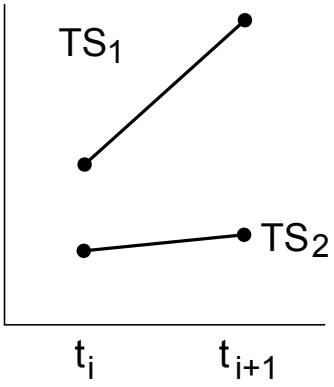
Ejemplo

En esta sección volveremos a comparar las series C_A y C_B pero en esta ocasión a través de los cambios locales producidos con respecto al signo y a la magnitud de la variación.

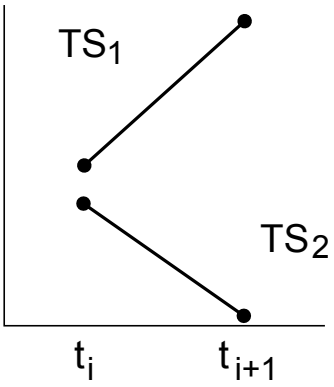
Hemos usado puntos consecutivos para el análisis de las dinámicas de cambio en las series, y hemos usado un valor $K = k = 1$ para construir $STend$. Para hacer esto,



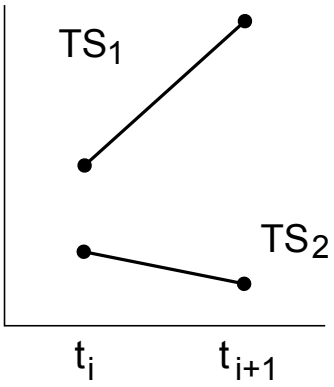
a) Mismo signo, variación similar



b) Mismo signo, variación diferente



a) Diferente signo, variación similar



b) Diferente signo, variación diferente

Figura 5.14: Ejemplos de cambios locales respecto al signo y la variación.

medimos el cambio local para cada punto de tiempo con respecto al siguiente en la línea temporal. A esta versión concreta de *STend* la hemos denominado *cambio local*.

Definición 5.6 (Cambio local) *Sea* $TS = \{< t_1, v_1 >, ..., < t_m, v_m >\}$ *la serie temporal y* t_i *con* $i \in [1, m-1]$ *un punto determinado de la serie. El cambio local* $CL_{TS}(t_i)$ *es*

$$CL_{TS}(t_i) = \arctan\left(\frac{TS(t_{i+1}) - TS(t_i)}{t_{i+1} - t_i}\right) = \arctan\left(\frac{v_{i+1} - v_i}{t_{i+1} - t_i}\right)$$

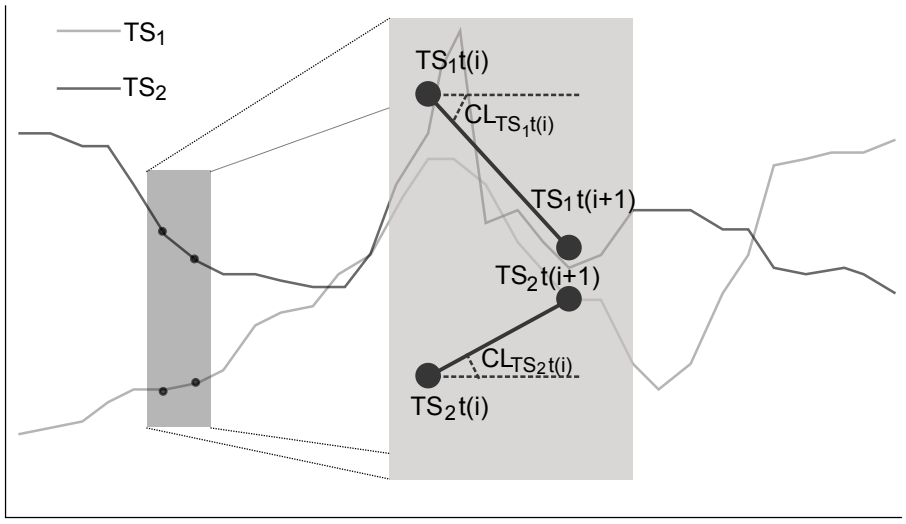


Figura 5.15: Cambio local

En la Figura 5.15 se puede ver de forma gráfica e intuitiva la semántica que hay detrás de las definiciones referentes al cambio local en un punto dado t_i.

Hemos utilizado el marco lingüístico formado por las etiquetas descritas en la sección anterior, junto con las etiquetas de tiempo que hemos utilizado en anteriores ejemplos, y el conjunto de cuantificadores definido en la Tabla 5.3.

Cuantificador	Definición
La mayoría	(0, 0.7, 0.9, 1)
Aproximadamente el 80 %	(0, 0.6, 0.8, 1)
Aproximadamente el 70 %	(0, 0.5, 0.7, 1)
Aproximadamente el 60 %	(0, 0.4, 0.6, 1)

Tabla 5.3: Cuantificadores para la comparación de series basadas en cambios locales.

Con respecto a los parámetros lingüísticos: el umbral $\tau = 0.7$; $Qlim_i$ toma valores $Qlim_1 = Qlim_2 = 2$ y $Qlim_3 = 3$, de forma que cuanto más se profundiza en la jerarquía temporal más permisivo con el uso de cuantificadores menos restrictivos se es.

El resumen obtenido para la comparación de las series es el siguiente:

"Aproximadamente el 60 % de los días en Enero, ambas series presentan cambios locales con el mismo signo (0.72)
La mayoría de los días en Diciembre, ambas series presentan cambios locales con variación similar pero diferente signo (0.98)
Aproximadamente el 60 % de los días en Mayo, ambas series presentan cambios locales con el mismo signo (1)
Aproximadamente el 70 % de los días en Junio, ambas series presentan cambios locales con variación similar y el mismo signo (0.72)
Aproximadamente el 60 % de los días en Abril, ambas series presentan cambios locales con el mismo signo (1). Para el resto de periodos la serie presenta variabilidad."

Si centramos nuestra atención en la cuarta sentencia cuantificada, la cual describe los cambios locales en Junio, podríamos llegar a pensar que existe una contradicción entre la gráfica que representaba ambas series (ver 5.7) y el resumen final. El hecho es que nuestro enfoque no trata de describir cambios globales sino locales, de modo que no se describe la tendencia de forma general sino localizada a puntos de tiempo consecutivos. Al contrario de lo que ocurre con la tendencia global, este tipo de información no es fácilmente detectable a través de representaciones gráficas, a pesar de ser una característica a tener en cuenta cuando se trata de realizar comparación entre series.

Lo que comentamos se debe a que cambios con diferente signo y valor alto entre algunos pares de puntos consecutivos puede llegar a compensar el efecto de un gran número de parejas que presentan cambios con el mismo signo pero pequeña magnitud de cambio, de forma que se crea una tendencia global diferente cuando la mayoría de las ocasiones la serie varía en la misma manera. Debemos decir que la tendencia global ofrece información muy relevante en el ámbito de las comparaciones, pero será objeto de nuestra investigación en el futuro.

5.2.3. Discusión y trabajo futuro

Hemos estudiado la comparación de series temporales en términos de valor y tendencia. En ambos casos, la comparación se basa en el resumen de una comparación a nivel muy local de las series, punto a punto. La principal ventaja de esta comparación es que permite destacar detalles que quedan ocultos al ser humano en muchas ocasiones, ya que nosotros tendemos a realizar resúmenes a nivel global o considerando entornos menos localizados, como intervalos temporales de cierta amplitud.

De nuevo nos encontramos con un problema muy extenso y con grandes posibilidades de trabajo futuro. Una de estas líneas futuras es precisamente la que acabamos de comentar, es decir, considerar comparación en términos de características de periodos más amplios de la serie. También hay mucho trabajo por hacer en cuanto a las características a considerar en la comparación, donde podemos incorporar otras características ya mencionadas en este capítulo como la estacionalidad o cualquier otra que podamos emplear. Además, la comparación basada en una combinación de

diversas características es una posibilidad a la vez interesante y donde hay mucho trabajo por realizar.

También es relevante el problema del resumen de conjuntos de series temporales de tamaño mayor que 2, bien tratando de describir las características que más se repiten entre todas ellas, bien calculando diferencias entre cada par de series y resumiendo a continuación dichas características, o agregando las mismas y resumiendo la agregación, etc. El número de posibilidades para nuestras técnicas de resumen es, como queremos dejar de manifiesto, enorme en el ámbito de las series de datos temporales. Pero, como veremos en la siguiente sección, incluso el ámbito de las series temporales, donde se centra el interés de la presente tesis, es solo la punta del iceberg. El potencial de nuestras técnicas a la hora de resumir puede extenderse a conjuntos de datos más generales e incluso muy complejos. En la siguiente sección, y en el mismo espíritu que motiva el resto del presente capítulo, discutiremos brevemente sobre este aspecto y esbozaremos, a efectos ilustrativos y para demostrar el potencial mencionado, una aplicación de nuestros algoritmos en un ámbito tan complejo como la descripción de imágenes.

5.3. Descripción lingüística de imágenes

En los apartados anteriores hemos visto cómo es posible aplicar nuestras técnicas a la descripción lingüística de diversos aspectos relativos a series de datos temporales, que son los tipos de datos objeto de nuestro interés en esta tesis. Sin embargo, las técnicas que hemos propuesto pueden aplicarse para obtener descripciones lingüísticas de cualquier conjunto de datos que cumpla unas condiciones mínimas:

- En primer lugar, es necesario poder estructurar los datos disponibles mediante una partición jerárquica difusa. Esta condición es realmente muy poco restrictiva. En ocasiones, dicha partición ha sido proporcionada como parte de los datos, como en el caso de las bases de datos multi-dimensionales que incorporan una dimensión tiempo. Cualquier otra dimensión organizada de forma jerárquica, incluso si no existe un orden subyacente, es susceptible de ser utilizada (aunque la falta de un orden puede limitar el tipo de resúmenes que se pueden realizar, no siendo posible por ejemplo analizar tendencias). Asimismo, cuando no se proporciona directamente la partición jerárquica, ésta puede obtenerse mediante el uso de cualquiera de las muchas técnicas de *clustering* jerárquico difuso existentes. Más aún, en la práctica es posible obtener muchas particiones de este tipo utilizando distintas combinaciones de atributos del conjunto de datos y distintas medidas de distancia o semejanza.

- En segundo lugar, es necesario disponer de un conjunto de etiquetas lingüísticas que describan características del conjunto de datos y que, junto con el uso de

otros elementos independientes de los datos, como los cuantificadores, completen el marco lingüístico.

Como vemos, en la práctica nuestro modelo nos permitirá obtener descripciones lingüísticas para la gran mayoría de los conjuntos de datos que se manejan en la actualidad. Para ilustrar estas ideas, mostramos en esta sección una aplicación de nuestro modelo a la descripción lingüística de imágenes en base a etiquetas lingüísticas de conceptos visuales básicos como color, relación espacial y localizaciones espaciales. Esta aplicación ha surgido a través de una colaboración con un grupo de investigadores de la Universidad de Granada que trabajan en el ámbito de la descripción lingüística de imágenes, y que han proporcionado el marco lingüístico necesario.

5.3.1. El marco lingüístico

Las imágenes pueden almacenarse en un ordenador empleando diversas representaciones. Dejando de lado aspectos relativos a la compresión, que buscan eficiencia en el almacenamiento, podemos ver una imagen como un conjunto de puntos de color denominados *píxeles*, organizados en una estructura matricial o reticular que define relaciones espaciales entre los mismos. La representación del color de un pixel se realiza en base a un *espacio de color*, donde cada color se representa mediante una tripleta de valores reales, cada uno de ellos dentro de unos dominios variables según el espacio de color. Es habitual asimismo ver las imágenes como un grafo en el cual los píxeles son los nodos y existe un arco entre dos píxeles cuando éstos son *vecinos*, es decir, son adyacentes en el retículo o matriz considerado.

A la hora de aplicar nuestras técnicas, los píxeles jugarán el papel correspondiente a los instantes de tiempo en las series temporales, es decir, identificar los ítems básicos cuyas características describiremos mediante sentencias cuantificadas, utilizando un marco lingüístico apropiado. Dicho marco lingüístico estará compuesto por una segmentación jerárquica difusa de la imagen, que jugará el papel correspondiente a nuestra jerarquía de tiempo, y una partición difusa del espacio de color en base al concepto de *color difuso*. Asimismo utilizaremos una partición difusa de las localizaciones de la matriz de píxeles para asignar a los mismos no sólo etiquetas de color, sino también de localización.

Segmentación jerárquica

El uso de técnicas de segmentación jerárquica es muy común cuando se quiere obtener el conjunto de regiones relevantes de una imagen. La segmentación difusa obtiene como resultado una colección de regiones difusas (subconjuntos difusos de píxeles conectados que poseen características similares) que forman una partición difusa de los píxeles de la imagen [23, 98, 101, 105, 122, 124, 148].

Existen diversos enfoques para la segmentación jerárquica de imágenes [57, 68, 92, 154]. El uso de la segmentación jerárquica en procesamiento de imágenes y visión por computador es muy importante en aplicaciones como compresión de imágenes [140, 153], descripción de escenas y parseo de imágenes [159], descubrimiento de conocimiento en imágenes [155] y recolección de datos mediantes sensores[1] [95], entre otros [60, 97].

En [19, 125] se describe un enfoque para la segmentación de una imagen mediante una jerarquía difusa en base a una segmentación difusa. Esta técnica puede ser aplicada a segmentaciones difusas obtenidas usando cualquier método pero nosotros la aplicaremos sobre aquella que obtengamos al aplicar la técnica descrita en [124].

La generalización del modelo que presentamos aquí es independiente de cómo se haya obtenido la jerarquía difusa siempre y cuando:

- La segmentación de una imagen se encuentre organizada en n niveles $L = L_1, \ldots, L_n$.
- Cada nivel L_i tenga asociada una segmentación difusa de la imagen en p_i regiones $\{D_{i,1}, \ldots, D_{i,p_i}\}$.
- La función de pertenencia para las regiones difusas esté normalizada.

Se asume que cada nivel contiene una partición difusa de los píxeles de la imagen, donde $\{X_1, ..., X_r\}$ se considera una partición en X si y solo si:

1. $\bigcup_{i \in \{1,\ldots,r\}} Support(X_i) = X$.
2. $\forall i, j \in \{1, \ldots, r\}, i \neq j, Core(X_i) \cap Core(X_j) = \emptyset$.
3. $\forall i \in \{1, \ldots, r\}\ \exists x \in X$ tal que $X_i(x) = 1$, esto es, que haya al menos un objeto completamente representativo de X_i.

La Condición 3 se refiere a que los conjuntos difusos de la partición estén normalizados. Como se ve, las restricciones aplicadas para considerar a un conjunto de niveles como una jerarquía son las mismas que las vistas para el modelo original (Sección 3.3).

Localizaciones absolutas

Hemos considerado para la descripción lingüística un modelo de localización absoluta de las regiones de una imagen [100]. Dichas localizaciones absolutas pueden ser interpretadas como relativas con respecto a los límites de la imagen.

[1] remote sensing

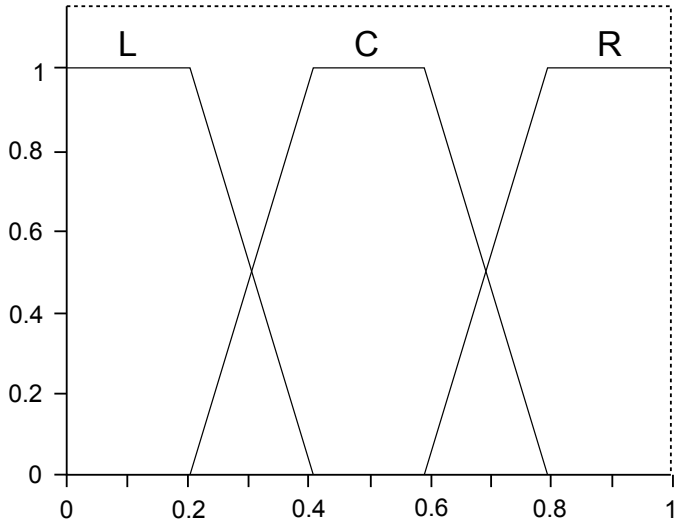


Figura 5.16: Posición horizontal difusa. L: izquierda; C: centro; R: derecha.

Para nuestra propuesta se empleará una partición difusa de la imagen con la que definir la localización absoluta de una región. Las Figuras 5.16 y 5.17, muestran la partición definida propuesta en el dominio de porcentajes con respecto a la longitud vertical y horizontal de la imagen, respectivamente. El producto cartesiano de ambas particiones usando el mínimo nos da una partición difusa del área de la imagen tal y como se muestra en la Figura 5.18. Dicha partición puede ser refinada mediante el uso de más etiquetas en ambas longitudes en caso de que fuese necesario.

Determinaremos el grado en el que una partición D está en una localización absoluta difusa A mediante la evaluación de una sentencia cuantificada de la forma *Q de D son A* utilizando el método GD, introducido al presentar el modelo original en la Sección 3.3.

Las localizaciones absolutas mostradas en 5.18, o alguna otra alternativa de granularidad más fina, pueden ser enriquecidas a través de un agrupamiento jerárquico de localizaciones. La idea es que regiones extensas tal vez no puedan ser incluidas en localizaciones por su tamaño, de modo que se deberán considerar localizaciones mayores y menos precisas si fuera necesario. Con esta idea se podría obtener una ontología completa de localizaciones en la cual, por ejemplo, la unión de las etiquetas DL, DC y DR se llamará simplemente *"Bajo"* y la unión de todas las etiquetas excepto MC, sería llamada *"Perímetro"*. Para obtener la localización de la ontología que mejor describa la figura, debemos considerar el máximo grado de cumplimiento de la sentencia cuantificada correspondiente y buscar sentencias con las localizaciones más precisas. De este modo, evitaremos las localizaciones extensas en la medida de lo

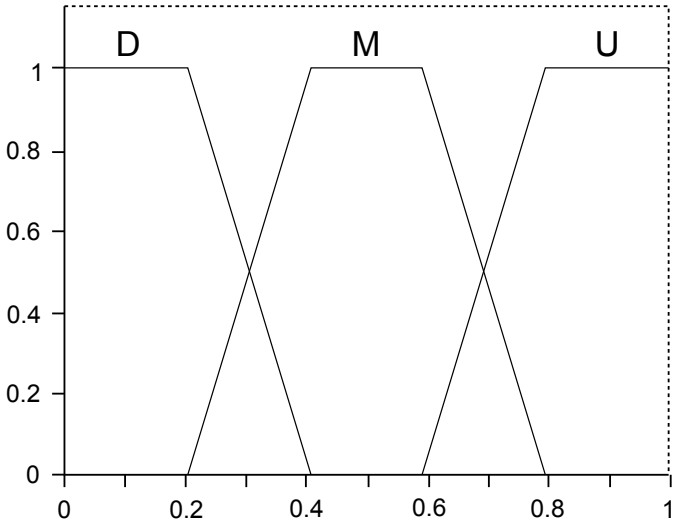


Figura 5.17: Posción vertical difusa. D: abajo; M: en medio; U: arriba.

posible.

Caracterización lingüística del color en las regiones

Como ya hemos indicado, en ciencias de la computación, el color se representa normalmente por una tripleta de valores reales. Estos valores pueden ser diferentes, teniendo diferentes dominios y semántica para dichos valores reales. Cada uno de estos sistemas se llama *espacio de color*. Un buen ejemplo, muy conocido y extendido, es el espacio de color RGB[2], donde la semántica de los tres valores que definen el color (enteros en [0, 255]) es la cantidad de rojo, verde y azul necesarios para reproducir el color.

Los humanos podemos diferenciar y trabajar con un número relativamente pequeño de colores (hasta 300) en comparación con la cantidad de ellos que pueden ser expresados a través de los espacios de color (algunos millones). Hacemos uso de dichos colores a través de términos lingüísticos que los representan. Por ejemplo, nosotros, de forma natural, no empleamos una tripleta de la valores numéricos $[255, 0, 0]$ cuando hablamos, sino que usamos el término lingüístico *rojo*. Además, no existe una relación unívoca entre un término lingüístico y el color del espacio de color, sino que cada término lingüístico representa un subconjunto de representaciones. Desafortunadamente, las fronteras de dichas representaciones son difusas por ser altamente subjetivas, dependiendo del dominio de aplicación y de aspectos culturales.

[2]acrónimo en inglés para Red, Green, Blue

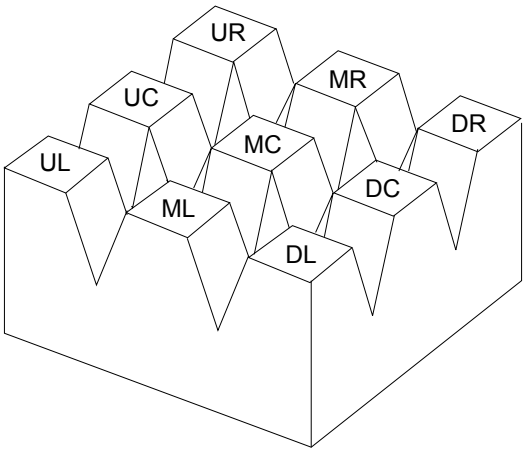


Figura 5.18: Localizaciones difusas absolutras como combinación de las longitudes horizontal y vertical.

La falta de correspondencia clara entre espacios de color y términos lingüísticos es un claro ejemplo de lo que se suele denominar mediante el término inglés *semantic gap*, y constituye un importante problema para las aplicaciones que pretenden afrontar de manera solvente la generación de lenguaje natural. En este modelo, y con el fin de manejar la imprecisión en la descripción de colores, tomaremos las ideas presentadas en [149],

Definición 5.7 *Un color difuso $\widetilde{C}$ es una etiqueta lingüística cuya semántica se representa en un espacio de color XYZ por un conjunto difuso normalizado de $D_X \times D_Y \times D_Z$.*

Definición 5.8 *Un espacio de color difuso $\widetilde{XYZ}$ es un cojunto de colores difusos XYZ que define una partición difusa de $D_X \times D_Y \times D_Z$.*

En esta última definición, la noción de partición difusa usada es la que se introdujo previamente en la Sección 5.3.1.

En el mismo trabajo se presenta una propuesta para la construcción de un espacio de color difuso adaptado, tomando como punto de partida una representación crisp de los colores $R = \{\mathbf{r}_1, \ldots, \mathbf{r}_m\}$ totalmente representativa de los colores difusos que se desean obtener. Para cada $\mathbf{r}_i$ obtendremos un color difuso atómico $\widetilde{C}_i$ basada en

la partición del espacio de color RGB usando la distancia euclídea d para obtener las funciones de pertenencia. Los colores que se obtengan cumplirán las siguientes propiedades:

- Los conjuntos difusos obtenidos son normalizados y convexos.
- El conjunto de conjuntos difusos obtenidos forman un espacio de color difuso ya que conforman una partición difusa en el sentido que se indica en la Sección 5.3.1.
- $\widetilde{C}_i(\mathbf{c}) > 0{,}5$ si y solo si $d(\mathbf{r}_i, \mathbf{c}) < d(\mathbf{r}_j, \mathbf{c})\ \forall i \neq j$.
- Si un color difuso $\mathbf{c}$ es equidistante de dos representantes $\mathbf{r}_i$ y $\mathbf{r}_j$ entonces $\widetilde{C}_i(\mathbf{c}) = \widetilde{C}_j(\mathbf{c}) = 0{,}5$

Usando dicha metodología, se desarrollan una serie de espacios de color difusos usando nombres de color del sistema ISCC-NBS [85,86]. Dicho sistema está basado en el trabajo de Berlin y Kay [10] acerca del nombramiento de colores y ha sido probado con humanos en tareas de descripción, de modo que es adecuado para aplicarlo en nuestro modelo. ISCC-NBS provee varios conjuntos de colores en forma de pares (término lingüístico, color crisp) y en [149] se han definido tres espacios de color difusos con diferentes grados de granularidad:

- **Conjunto básico**: 13 nombres de color que corresponden a los diez términos de color básicos (rosa, amarillo, rojo, naranja, marrón, oliva, verde, azul, violeta y morado) y tres acromáticos (blanco, gris y negro).
- **Conjunto extendido**: 31 nombres de color que corresponden a aquellos del conjunto básico y algunas combinaciones de ellos (por ejemplo, naranja amarronado o naranja rojizo entre otros).
- **Conjunto completo**: 267 nombres de color que se obtienen a partir del conjunto extendido añadiendo cinco modificadores de tono (muy luminoso, luminoso, medio, oscuro, muy oscuro) y cuatro adjetivos de saturación (grisáceo, moderado, fuerte y vívido). Además, tres términos adicionales que sustituyen ciertas combinaciones de luminosidad y saturación (pálido para grisáceso luminoso, brillante para luminoso fuerte y profundo para oscuro fuerte). Estos nombres de color se representan usando el *Lenguaje Universal del color (Universal Color Language)*, nivel 3 en el sistema ISCC-NBS.

Puntualicemos que la elección de un espacio de color u otro tendrá una influencia notable en la descripción lingüística final. Por ejemplo, entre los tres espacios presentados anteriormente, el básico permite la obtención de una descripción más breve

ya que hay menos colores difusos que contienen más colores crisp en ellos, de modo que será más sencillo encontrar regiones que contengan, en su mayoría, píxeles correspondientes a un color difuso. Por el contrario, el espacio completo nos aportará la posibilidad de contar con descripciones con más niveles de detalle y colores mas precisos. El espacio extendido constituye un compromiso entre los anteriores. Recordemos que los presentados son sólo tres ejemplos y que existen muchos más, y sobre todo, que el modelo de resumen presentado es independiente de la elección que se realice, aunque estará altamente influenciado por ella.

Relaciones espaciales

De cara a enriquecer la descripción de la imagen, utilizaremos relaciones espaciales entre regiones. La distribución espacial de objetos proporciona información clave cuando se trata de la descripción de imágenes. Usando el resultado del proceso de segmentación difusa, una imagen puede ser interpretada como un grafo cuyas regiones son los vértices, y dos regiones estén conectadas cuando limitan entre sí. Una mejora de esta representación se obtiene al etiquetar los arcos del grafo con la relación espacial entre las regiones, de este modo obtendríamos finalmente un grafo etiquetado. Antes de poder obtener un grafo etiquetado a partir de una imagen deberemos:

- Determinar las posibles relaciones espaciales entre regiones y definir los términos lingüísticos usados para denominar cada una de ellas.

- Proporcionar un procedimiento para determinar la relación espacial que es adecuada para un par de regiones dado.

Existen muchos enfoques que se enfrentan al primer punto modelando los términos que denominan la relación como ontologías. El modelo RCC-8 propuesto por [26] se usa para la ontología en [100,152] y la versión difusa (fuzzy RCC-8) propuesta por [142] es usada en [152].

Para obtener las relaciones espaciales nosotros usaremos el modelo RCC-8 difuso tal y como se usa en [152]. En la Tabla 5.4 se muestran las relaciones espaciales, así como su descripción lógica difusa tal y como aparecen en el trabajo mencionado.

Debemos hacer hincapié en que nuestro modelo es independiente del método utilizado para obtener las relaciones espaciales usadas y el modo en el que están relacionadas con la imagen. A partir de este momento, asumiremos que tenemos disponible un conjunto de relaciones espaciales $\mathcal{SR} = \{R_1, \ldots, R_k\}$.

Un concepto muy importante asociado a la idea de relaciones espaciales es el siguiente: supongamos una serie de relaciones espaciales difusas $\mathcal{SR} = \{R_1, \ldots, R_k\}$.

Nombre	Relación	Definición RCC
Disconnected	DC	$\neg C(a,b)$
Part	P	$\forall c.(C(c,a) \rightarrow C(c,b))$
Proper Part	PP	$P(a,b) \wedge \neg P(b,a)$
Equals	EQ	$P(a,b) \wedge P(b,a)$
Overlaps	O	$\exists c.(P(c,a) \wedge P(c,b))$
Discrete	DR	$\neg O(a,b)$
Partially Overlaps	PO	$O(a,b) \wedge \neg P(a,b) \wedge \neg P(b,a)$
Externally connected	EC	$C(a,b) \wedge \neg O(a,b)$
Non Tangential Part	NTP	$\forall c.(C(c,a) \rightarrow O(c,b))$
Tangential PP	TPP	$PP(a,b) \wedge \neg NTP(a,b)$
Non-Tangential PP	NTPP	$PP(a,b) \wedge NTP(a,b)$

Tabla 5.4: Relaciones espaciales difusas RCC-8

Para una cierta segmentación de la imagen donde $D_i = \{D_{i,j}$ tal que $, j \in \{1, \ldots, p_i\}\}$ y

$$D = \bigcup_{i \in \{1,\ldots,n\}} D_i$$

es decir, D es el conjunto de todas las regiones difusas que aparecen los distintos niveles de la jerarquía, llamaremos *grafo de regiones* de la imagen al grafo dirigido $G = (D, E)$ donde los vértices son regiones en D y existe un arco dirigido entre dos regiones diferentes $D_{i,j}$ y $D_{k,l}$ en D. El etiquetado del grafo consiste en la asignación de enteros en $\{1, \ldots, k\}$ a los arcos, de modo que el arco que una las regiones $D_{i,j}$ y $D_{k,l}$ se etiquete con un entero z si y solo si, la relación espacial difusa que mejor define la relación espacial entre ambas regiones sea R_z.

5.3.2. Aplicación del modelo a la descripción de imágenes

Una vez que hemos visto cómo obtener un grafo etiquetado para representar una imagen dada, veremos aquí el enfoque que seguiremos para conseguir una descripción lingüística en base al color y la localización de las regiones. Este enfoque puede extenderse a otros términos lingüísticos, uno de ellos, por ejemplo, la textura. El enfoque se divide en dos fases:

1. Obtener el resumen de la información sobre el color en las imágenes asignando el término de color más representativo a cada región de un determinado subconjunto. La representatividad se mide en términos de la cantidad de píxeles que concuerdan con el color y un mínimo umbral con un cierto valor aportado por el usuario a través de un cuantificador lingüístico y un mínimo grado de

cumplimiento. Esto garantizará que el resumen sea *exacto*. El subconjunto viene determinado como una partición de la imagen con una colección mínima de regiones difusas, de modo que se intentará asegurar la *brevedad* del resumen al mismo tiempo que se cubra toda la imagen.

2. Se construirá una colección de sentencias en lenguaje natural a partir de las sentencias obtenidas en el paso anterior y la información espacial en forma de localizaciones absolutas representadas en el grado etiquetado. Adicionalmente, y para enriquecer el resumen con una fase de post-proceso, utilizaremos relaciones espaciales entre regiones para enlazar unas sentencias con otras.

Resumen de la imagen usando colores difusos y localizaciones absolutas

En esta Sección veremos cómo usar el modelo presentado en el Capítulo 3 e implementado en el Capítulo 4 para hacer resumen de una imagen en base a colores difusos y localizaciones absolutas. Tengamos en cuenta que tanto los colores difusos como las regiones difusas son subconjuntos difusos de los píxeles de una imagen. Como en ocasiones anteriores nuestro resumen estará formado por una colección de sentencias cuantificadas de la forma Q de $D_{i,j}$ son A, pero ahora,

- $D_{i,j}$ es una etiqueta j miembro de un determinado nivel i de la jerarquía asociada a la segmentación difusa de la imagen.
- A es un color difuso del espacio de color difuso escogido por el usuario.

De la misma forma el usuario deberá proveer al algoritmo con un subconjunto de una familia coherente de cuantificadores, un umbral τ de mínimo grado de cumplimiento y una pareja de límites. La evaluación de las sentencias se realizará mediante el método GD.

Generación del resumen final

Una vez que tenemos disponible el resumen de la imagen basada en color y localizaciones, usaremos dicha información para crear la descripción lingüística final. Sea D' el conjunto de regiones difusas que se ha utilizado en las sentencias del resumen generado (una sentencia por región). El proceso genérico se detalla en Algoritmo 6.

Como se puede ver en este pseudo-código sobre cómo afrontar el proceso del resumen encontramos temas abiertos. El primero de todos, ¿cómo establecer el orden de las regiones difusas de la descripción?. Algunos autores, como por ejemplo [110], proponen usar medidas de preferencia que reflejen los intereses del usuario en cuestión. Nosotros consideramos diversas posibilidades de ordenación basadas en términos de posición, tamaño, color o combinaciones entre ellas. En nuestro modelo, asumimos que

Algoritmo 6 : algoritmo para obtener la descripción lingüística final.

1: Ordenar las regiones difusas D' siguiendo un order total. Suponiendo $D' = \{D^1, \ldots, D^u\}$ con $D^i \prec D^{i+1}\ \forall i$.
2: Añadir la sentencia lingüística de D^1 más la ubicación absoluta de D^1.
3: $i \leftarrow 2$
4: **mientras** $(i \leq u)$ **hacer**
5: Enlazar sentencias utilizando la relación espacial entre D^{i-1} y D^i
6: Añadir la sentencia lingüística para D^i más la ubicación absoluta de D^i
7: $i \leftarrow i + 1$
8: **fin mientras**

las regiones difusas se encuentran ordenadas por tamaño. En segundo lugar, la descripción lingüística generada puede incorporar diferentes elementos. Nosotros empleamos términos lingüísticos relativos al color y la localización absoluta. Para terminar, la relación espacial entre regiones se emplea para relacionar las diferentes descripciones de las regiones en algo parecido a una navegación por la escena.

Ejemplo

La presente sección está dedicada a ilustrar con ayuda de un ejemplo sencillo la propuesta de aplicación de nuestro modelo de resumen lingüístico para la descripción de imágenes. La Figura 5.19 muestra la imagen que se usará en el ejemplo. Los cuadrados azules que se aprecian en la imagen indican las semillas empleadas en el proceso algorítmico de segmentación difusa, basada en crecimiento de regiones. Como se puede ver algunas de las semillas se han colocado sobre la misma región lo que puede provocar una sobre-segmentación de la imagen. Siguiendo el algoritmo de segmentación propuesto en [124] obtendremos dos regiones superpuestas. Obviamente, estaremos interesados en la descripción de una de ellas. Como veremos nuestro modelo resolverá este inconveniente.

A partir del algoritmo de segmentación se obtendrá una segmentación jerárquica difusa usando [19, 125]. La segmentación jerárquica se muestra en la Figura 5.20. La jerarquía está compuesta por ocho niveles $L_1, \ldots, L_8$ descritos en ocho columnas. Cada columna contiene la función de pertenencia de cada región en el nivel correspondiente. La función de pertenencia se representa en forma de imagen, donde el color blanco representa pertenencia con grado uno y el color negro, pertenencia con grado 0. Los colores grises se corresponden con grados intermedios. La inclusión de regiones difusas en un nivel con respecto a las regiones en otros niveles también queda marcada en la figura. La unión de regiones difusas se obtiene usando el máximo.

El primer nivel L_1 de la jerarquía corresponde con la segmentación difusa obtenida

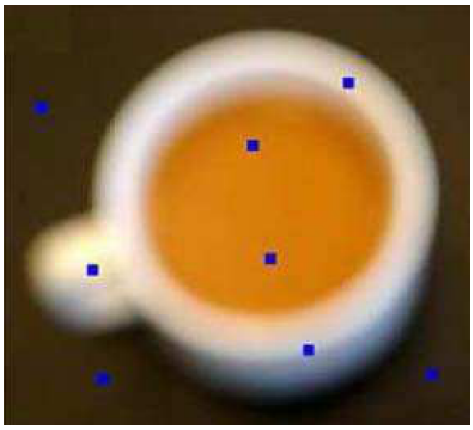


Figura 5.19: Imagen de ejemplo.

por el algoritmo propuesto en [124] y consta de ocho regiones difusas que se corresponden con las ocho semillas iniciales representadas en la figura 5.19. La numeración de las regiones se realizará de arriba a abajo en cada nivel. Por ejemplo, en el nivel L_2 la región difusa $D_{2,1}$ se corresponde con la unión de las regiones difusas $D_{1,1}$ y $D_{1,2}$ del nivel L_1.

Para este ejemplo se ha considerado un solo cuantificador trapezoidal representado por $Q = (0,0{,}7, 0{,}9, 1)$ que se ha llamado *La mayoría* y un umbral $\tau = 0{,}7$. No se ha considerado el agrupamiento de colores. Con esta configuración los parámetros $Qlim$ y $Glim$ no tienen razón de ser y se quedan con el valor por defecto 1. Se ha usado el conjunto básico de colores del sistema ISCC-NBS, las localizaciones absolutas representadas en la Figura 5.18 y las relaciones espaciales expuestas en la Tabla 5.4.

A continuación mostraremos en detalle el proceso de descripción de color aplicando el Algoritmo 1. En primer lugar se inicializa la cola *ParaResumir* con las regiones difusas del último nivel, en este caso $\{D_{8,1}\}$. Se saca el primer elemento, y único, y se analiza. El algoritmo no encuentra un buen resumen que la región difusa (conducta esperada al representar la imagen completa, que no es un color homogéneo) de modo que se procede a añadir a la cola los hijos de dicha región $ch(D_{8,1}) = \{D_{7,1}, D_{7,2}\}$. Se analiza el siguiente elemento de la cola, es decir, $\{D_{7,1}\}$. En este caso el algoritmo ha tenido éxito al intentar encontrar un color difuso que, insertado en la sentencia cuantificada, tenga un grado de cumplimiento mayor o igual al umbral τ. La sentencia que se añade al resumen (Resumen) es *La mayoría de píxeles en la región* $\{D_{7,1}\}$ *son de color naranja.* Se continúa con el análisis de la siguiente región en la cola $\{D_{7,2}\}$. Esta vez el algoritmo no encuentra ninguna sentencia que logre describir la región, de modo que añade a la cola los hijos de $\{D_{7,2}\}$, $ch(D_{7,2}) = \{D_{6,2}, D_{6,3}\}$. El análisis de la región $\{D_{6,2}\}$ concluye con una nueva sentencia que se añade al resumen, siendo

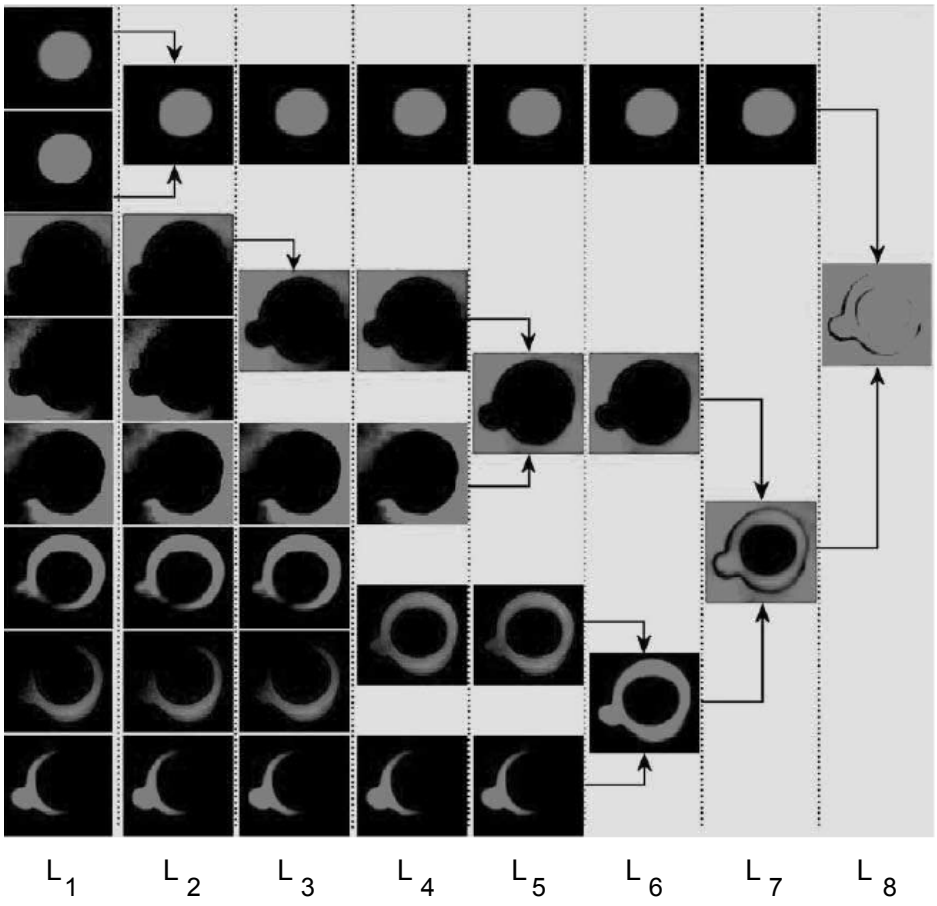


Figura 5.20: Segmentación jerárquica de la imagen de la Figura 5.19.

esta *La mayoría de píxeles en la región* $\{D_{6,2}\}$ *son de color oliva.* Mientras que para la región $\{D_{6,3}\}$ se obtiene *La mayoría de píxeles en la región* $\{D_{6,3}\}$ *son de color blanco.* Llegados a este punto la cola está vacía (*ParaResumir*$= \emptyset$), de modo que el proceso finaliza con la descripción de la totalidad de la imagen con sólo tres sentencias. En este ejemplo tenemos que $D' = \{D_{7,1}, D_{6,2}, D_{6,3}\}$.

Con respecto a las localizaciones absolutas de las regiones, podemos decir que son regiones bastante grandes en comparación con la totalidad del tamaño de la imagen. Empleando un umbral de cumplimiento relativamente bajo se obtiene que la localización para las regiones $D_{7,1}$ y $D_{6,3}$ es MC (*Medio-Centro*), aunque para la región $D_{6,3}$ el grado de cumplimiento era casi el mismo para *Perímetro*; mientras que para $D_{6,2}$ la posición es, esta vez sí, *Perímetro.*

Para la generación de la descripción lingüística tomamos las regiones ordenadas de la siguiente forma, $D_{6,2}, D_{6,3}, D_{7,1}$. El resumen final se genera describiendo cada una de estas regiones, en orden, en términos de su posición y color. Para la primera región se presentan las posición absoluta y el color. A continuación, la posición relativa entre ésta y la siguiente región, y el color de la misma; y así hasta que esté descrita la última de las regiones. El resultado es,

> *Existe una región en el perímetro de la imagen donde la mayoría de los píxeles son de color oliva; dicha región presenta una relación no tangencial con la región situada en el medio centro de la imagen y donde la mayoría de los píxeles son de color blanco; dicha región presenta una relación no tangencial con otra región situada en el medio centro de la imagen donde la mayoría de los píxeles son de color naranja.*

5.4. Conclusiones

Durante el transcurso de este capítulo se han presentado diversas adaptaciones sencillas de nuestras técnicas para la descripción lingüística de datos, tanto en problemas relacionados con la descripción de series de datos temporales, como en otros tipos de datos. Nuestra conclusión principal es que las técnicas propuestas presentan, como valor añadido, la posibilidad de adaptarse de manera relativamente simple a una extensa casuística de problemas de descripción de datos.

Cada uno de los problemas analizados en este capítulo ha sido objeto de un estudio relativamente superficial, en el sentido de que hemos mostrado una aplicación, con una secuencia de decisiones bastante concreta (aunque en ocasiones hemos proporcionado una serie de alternativas), que demuestre que es posible la aplicación de nuestras técnicas al problema.

En primer lugar encontramos una aplicación sencilla del modelo para la descripción de series de datos temporales pero esta vez sin tener en cuenta el valor en sí, sino las variaciones de valor en periodos determinados. Hemos dado un paso más al, en segundo lugar, aplicar nuestro modelo a la comparación de series de datos. La descripción de la comparación entre series se ha realizado en base a la descripción de la serie diferencia entre dos series originales que se desee comparar. Se han presentado varias formas de calcular la mencionada serie diferencia entre las cuales el usuario puede elegir la que más se adapte a sus necesidades en un momento determinado.

Por un lado se han presentado una serie de técnicas que nos permiten la comparación de series basándonos en el valor de las mismas en cada instante de tiempo. En este sentido se han desarrollado enfoques tanto de carácter absoluto como relativo. Por otro lado, hemos ampliado la comparación de series incorporando, también aquí,

otras características de las series como son las representadas por los cambios locales en cuento al signo y la variación en magnitud.

En tercer y último lugar, y para subrayar la versatilidad del modelo, hemos discutido sobre la aplicación del modelo sobre otros conjuntos de datos. En la parte final del capítulo se presenta un propuesta en la que se trabaja en la descripción de conjuntos de datos que representan imágenes en formato digital. Para este fin se ha segmentado la imagen en diferentes zonas de las que nos interesan tanto sus características (en este caso es color) como su ubicación en la imagen y su relación con otras áreas. Para la descripción se han utilizado tanto técnicas de segmentación jerárquica y difusa de la imagen como un espacio de color difuso que nos permita trabajar con colores de forma sencilla pero también cercana al ser humano.

En todos los problemas estudiados, y sin perjuicio de nuestras conclusiones sobre el valor añadido potencial del modelo, quedan abiertas una gran cantidad de líneas de estudio y trabajo futuro.

6 *Linguistic F-Cube Factory*

"La naturaleza nunca hace nada sin motivo"
Aristóteles

Como hemos comentado en más de una ocasión en esta memoria, las compañías y organizaciones generan y consumen ingentes cantidades de datos durante el desarrollo de sus actividades. Sin embargo, la posesión de numerosos datos no es directamente equivalente a poseer mucha información. Ser capaces de, a partir de un conjunto de datos, obtener información que, además de previamente desconocida y oculta, sea útil y relevante es una tarea de gran importancia.

Las diferentes herramientas desarrolladas dentro del marco de la Inteligencia de Negocio (Business Intelligence) tienen como objetivo facilitar los procesos de obtención de dicha información en las empresas. De esta forma ofrecen a los directores de las compañías la posibilidad de conocer y entender mejor el desarrollo de las actividades que llevan a cabo. Un mejor entendimiento lleva a una mejora en la toma de decisiones comerciales relevantes.

En este sentido, los sistemas de apoyo a la toma de decisiones son los encargados de ofrecer a usuarios que se encuentran en posiciones de dirección de la empresa las herramientas necesarias para facilitar el desempeño de su trabajo.

En el ámbito de los sistemas de apoyo para la toma de decisiones, el modelo de datos multi-dimensional juega un papel cada vez más protagonista a la hora de organizar los datos. Por eso, respondiendo a nuestro objetivo de llevar nuestro modelo de resumen a un campo de aplicación real a través del uso de una herramienta amigable y sencilla de utilizar, presentamos el sistema Linguistic F-Cube Factory, una herramienta de análisis de cubos OLAP que incorpora habilidades de resumen lingüísticas basadas en el uso de nuestras propuestas.

En este capítulo veremos, en primer lugar, la importancia de contar con herramientas que incorporen capacidad de análisis en modelos multi-dimensionales difusos a través del uso del lenguaje natural, y más concretamente del resumen lingüístico. Luego nos centraremos en presentar una herramienta existente que nos permite realizar de forma sencilla y rápida la gestión de una base de datos con modelo multi-dimensional: F-Cube Factory; y, a continuación, presentaremos la ampliación de dicha herramienta mediante la adición de capacidades de resumen lingüístico, dando lugar al denominado: *Linguistic F-Cube Factory.*

6.1. Motivación

La gestión de grandes conjuntos de datos en el ámbito de la toma de decisiones se suele hacer mediante el apoyo que nos ofrecen las bases de datos multi-dimensionales. Mediante este modelo, los almacenes de datos o *data warehouses* se encargan de alojar grandes cantidades de datos haciendo uso de una estructura basada en la presencia de varias dimensiones.

El modelo multi-dimensional se basa en el uso de cubos de datos. Cada cubo de datos contiene datos relativos a un hecho dado cuyo contexto se encuentra descrito a través de múltiples dimensiones con estructura jerárquica. Las herramientas OLAP (OnLine Analytical Processing) nos ofrecen la capacidad de consultar dichos almacenes de datos para obtener la información deseada, que es tan importante en el ámbito de la Inteligencia del Negocio.

Como se ha anticipado, los receptores de la información obtenida a través de estas consultas son personas y, para favorecer el entendimiento de la misma, es de interés poder presentar los resultados haciendo uso del lenguaje natural.

Con el objetivo de introducir capacidades lingüísticas en los cubos de datos, en [106] se presenta un modelo de datos multi-dimensional denominado difuso o lingüístico. Para poder convertir los datos en texto se recurre al uso de conjuntos difusos y etiquetas lingüísticas durante la definición de las diferentes dimensiones del modelo.

Con la idea de mejorar la toma de decisiones de los usuarios, pensamos que es deseable que, además de permitir la introducción del lenguaje natural en la definición de las dimensiones, se introduzca en la descripción de los hechos. En este punto hemos fijado nuestro objetivo; hemos centrado nuestros esfuerzos en conseguir que a través del uso de OLAP difuso sobre cubos multi-dimensionales difusos se puedan obtener resúmenes lingüísticos de los datos almacenados.

En este contexto nos hemos decidido a ampliar la mencionada herramienta de gestión de cubos de datos, F-Cube Factory, con la introducción de nuestro modelo de resumen lingüístico, algo que es posible en una herramienta interactiva gracias a la aproximación Greedy para la generación automática de los resúmenes presentada anteriormente en esta memoria.

A continuación introduciremos algunos de los aspectos principales de la herramienta F-Cube Factory para, a continuación, pasar a presentar la nueva herramienta desarrollada: Linguistic F-Cube Factory.

6.2. F-Cube Factory

La presente sección se encuentra dedicada a presentar de forma concisa algunos de los aspectos más importantes con respecto a la plataforma F-Cube Factory. En primer lugar mostraremos una breve introducción en forma de presentación general, para luego centrarnos en las características que nos han hecho considerar su uso.

F-Cube Factory es un sistema que implementa un modelo multi-dimensional de almacenamiento de datos [36,106]. Mediante esta herramienta, usuarios sin conocimiento experto tienen la oportunidad de trabajar con cubos de datos, tanto convencionales como difusos, de forma sencilla. Se ponen a disposición del usuario funcionalidades como la creación de cubos de datos, la posterior modificación o eliminación de los mismos, así como una serie de operaciones que permiten la consulta de los datos almacenados.

El sistema puede trabajar con distintos modelos para la gestión de los cubos de datos:

- Modelo ROLAP u OLAP relacional: el sistema puede gestionar cubos de datos mediante el uso de una base de datos relacional que permita el almacenamiento de datos, así como la obtención de datos para la construcción de nuevos cubos.

- Modelo MOLAP (Multidimensional OLAP) crisp: en este modelo los cubos de datos se almacenan usando una estructura puramente multi-dimensional.

- Modelo MOLAP difuso/lingüístico: en este modelo se implementan el modelo multi-dimensional difuso y lingüístico presentado en [106].

Podemos encuadrar la herramienta dentro del campo de la minería de datos o extracción del conocimiento en bases de datos ya que posibilita la recuperación de información novedosa que se encuentra oculta entre la gran cantidad de datos guardados en los almacenes de datos digitales.

El sistema F-Cube Factory sigue una arquitectura cliente-servidor. El sistema al completo ha sido desarrollado utilizando el lenguaje de programación Java. En el desarrollo de la nueva versión lingüística, se ha optado por mantener dicho lenguaje y la estructura cliente-servidor. De este modo, la aplicación producida sigue siendo independiente de la máquina o el sistema operativo sobre el que se vaya a ejecutar, además de contar con plataformas de desarrollo totalmente gratuitas.

La parte del servidor es la encargada de soportar la componente más pesada del sistema y para ello está compuesta por dos módulos principales sobre los que se asocian otros módulos adicionales.

El primero de los módulos principales es el que se encarga de los *cubos de datos*. Dichos cubos pueden ser construidos siguiendo los tres enfoques (ROLAP, MOLAP y MOLAP difuso) comentados con anterioridad y el acceso a los mismos puede hacerse de forma homogénea y transparente. Relacionados con este módulo están los que se encargan de la gestión de los cubos de datos y la conexión con la base de datos. Una de las funcionalidades del módulo es el soporte a las consultas. Se incluye además la posibilidad de trabajar con vistas de usuario.

El otro módulo principal se ocupa de las funciones de agregación usadas en las consultas. Dicho módulo interactúa con el anterior cuando queremos cambiar el nivel de detalle de un cubo de datos. Existen dos tipos de agregaciones implementadas: las usuales para cubos de datos convencionales o crisp y las difusas que se aplican sobre cubos de datos difusos. Este módulo de agregación, como veremos, ha sido completado con adaptaciones adecuadas de los algoritmos Greedy presentados en capítulos anteriores y que permiten incorporar las nuevas habilidades de resumen.

Ya que la mayor parte de la funcionalidad se encuentra de la parte del servidor, la parte del cliente es lo bastante ligera como para que se pueda usar en un ordenador personal sin grandes requerimientos técnicos. Además de ser ligero, el cliente está pensado para que provea un acceso intuitivo a todas las funcionalidades del servidor.

Existe una versión web del cliente, de modo que el usuario lo único que necesita es tener instalado un navegador con acceso a la red donde se encuentre el servidor. La interfaz descarga al usuario de la necesidad de conocimientos técnicos permitiendo que pueda acceder a las funcionalidades a través diversos menús y formularios.

El modelo multi-dimensional difuso sobre el que se sustenta la plataforma F-Cube Factory es un entorno natural en el que implantar nuestra propuesta de resumen lingüístico porque la incorporación de las nuevas capacidades de resumen mejora ostensiblemente su potencial para comunicarse con el usuario. Además, el modelo de datos de F-Cube Factory facilita sobremanera la configuración de un marco lingüístico adecuado para la definición de los resúmenes. En los cubos con dimensión temporal, la estructura jerárquica del tiempo que necesitamos para generar nuestros resúmenes se tiene directamente de la estructura del cubo. Además, F-Cube Factory permite la partición mediante variables lingüísticas de los dominios de la variables representadas en los hechos.

Como veremos, las operaciones OLAP de consulta sobre cubos con dimensión tiempo producen series de datos temporales que se le pueden presentar al usuario en forma de resúmenes lingüísticos. Más aún, la propia estructura de los cubos, facilita la aplicación de técnicas de comparación de series como las descritas en el Capítulo 5.

6.3. Nuestro modelo en F-Cube Factory

Una vez que hemos visto las características generales de la plataforma original, en esta sección pasaremos a presentar las nuevas capacidades que se han añadido.

F-Cube Factory implementa las operaciones OLAP usuales: *slice and dice* relacionada con la selección en los datos en un cubo, *roll-up* y *drill-down* que se encargan de cambiar la granularidad en el cubo navegando a través de los distintos niveles de abstracción de las jerarquías en las dimensiones, y la operación *pivot* para obtener representaciones alternativas de los datos. Cuando se trata con datos numéricos, se dispone de las funciones de agregación más comunes como el máximo y el mínimo, la media o la suma, entre otras. Todas estas operaciones se encuentran disponibles también para trabajar con hechos y dimensiones difusas.

En nuestro caso, hemos creado un nuevo tipo de función de agregación que se podría considerar en casos particulares durante la ya mencionada operación roll-up cuando se realiza sobre hipercubos que disponen de una dimensión tiempo. El objetivo es poder sustituir la utilización de agregadores convencionales que producen una única medida que resume la serie de datos agregada, por un nuevo operador de agregación que, basado en nuestro modelo, permita obtener un nuevo cubo de datos donde las celdas sean resúmenes lingüísticos de las series de tiempo agregadas durante el roll-up.

La Figura 6.1 muestra gráficamente el proceso haciendo uso de un ejemplo sencillo. En ella vemos la representación de un cubo de datos simple en el que se almacena información acerca de la afluencia de pacientes a diferentes centros de salud a lo largo del tiempo. Para almacenar los datos contamos con tres dimensiones básicas: *centro* para determinar el centro de salud, *género* para determinar el género de los pacientes y *tiempo* para almacenar la información temporal. Si sobre ese cubo de datos aplicamos la operación *roll-up* sobre la dimensión temporal a nivel de año, con la nueva función de agregación *resumen lingüístico* podemos obtener un nuevo cubo de datos con las mismas dimensiones pero diferente nivel de granularidad en la dimensión *tiempo*. El objetivo de la nueva capacidad es permitir al usuario generar este nuevo cubo donde las medidas numéricas que expresan la afluencia a lo largo de un año han sido sustituidas por un resumen lingüístico que describe la situación.

Como podemos observar en el cubo resultado, los hechos, en lugar de ser numéricos son textos. De esta forma se le ofrece al usuario la posibilidad de navegar en el cubo en busca de la información que le interese de una manera alternativa cercana al lenguaje natural.

De la misma manera, dentro de un cubo de datos, se puede plantear la comparación de series en relación con un determinado valor de una de las dimensiones del cubo. La Figura 6.2 muestra la forma de incorporar la capacidad de comparación descrita en el capítulo 5 aplicada dentro de un cubo OLAP. Sobre el cubo original, se selecciona

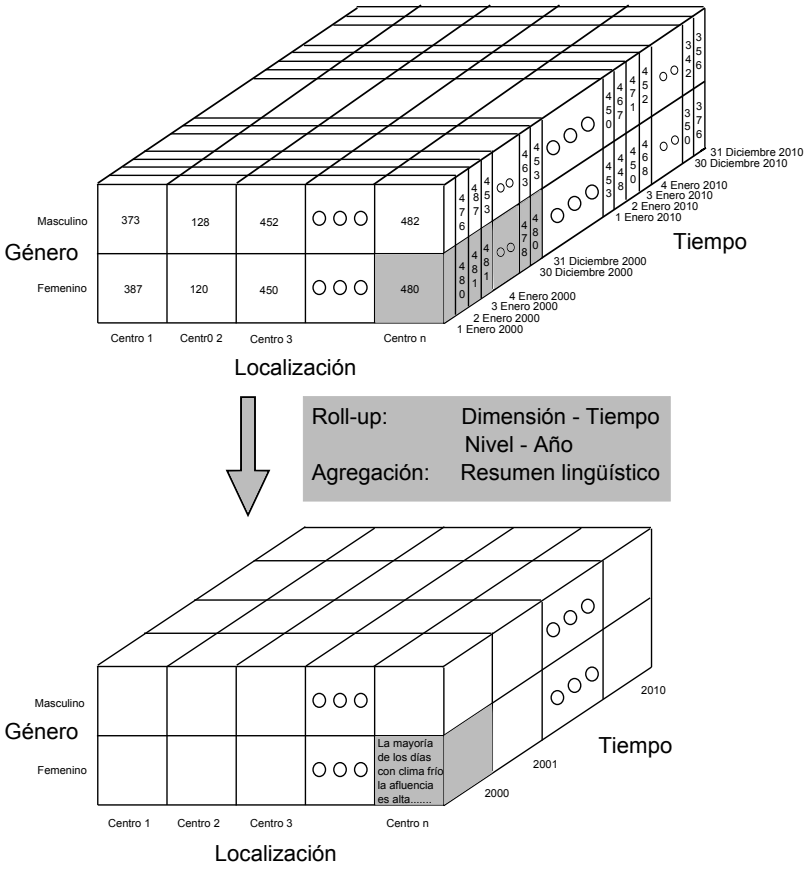


Figura 6.1: Operación *roll-up* con función de agregación *resumen lingüístico* sobre un cubo de datos con dimensiones *género*, *localización* y *tiempo*. El resultado es otro cubo de datos en el que los hechos se describen mediante resúmenes lingüísticos que sustituyen a los datos temporales agregados.

un valor (en este caso el centro n), que se usará como valor de referencia para las comparaciones. A partir de ahí, se genera un nuevo cubo de datos, ya sin este centro, en el que dentro de cada celda aparecen las distintas medidas de comparación (a saber, ΔTS_{abs}, ΔTS_{global}, ΔTS_{local}, SS, SV) en relación con el correspondiente valor de la celda equivalente del centro n seleccionado. Finalmente, se puede generar un cubo con los resúmenes lingüísticos a partir del cubo con las series de comparación. En la figura, se muestra el cubo con los resúmenes de ΔTS_{local}.

Esta nueva versión de la herramienta en la que, además de etiquetas lingüísticas en la definición de las dimensiones, se permite que aparezcan textos dentro del mismo cubo como resultado de la nueva función de agregación, aporta más flexibilidad a la plataforma. Se pone en manos del usuario decisor una herramienta interactiva con

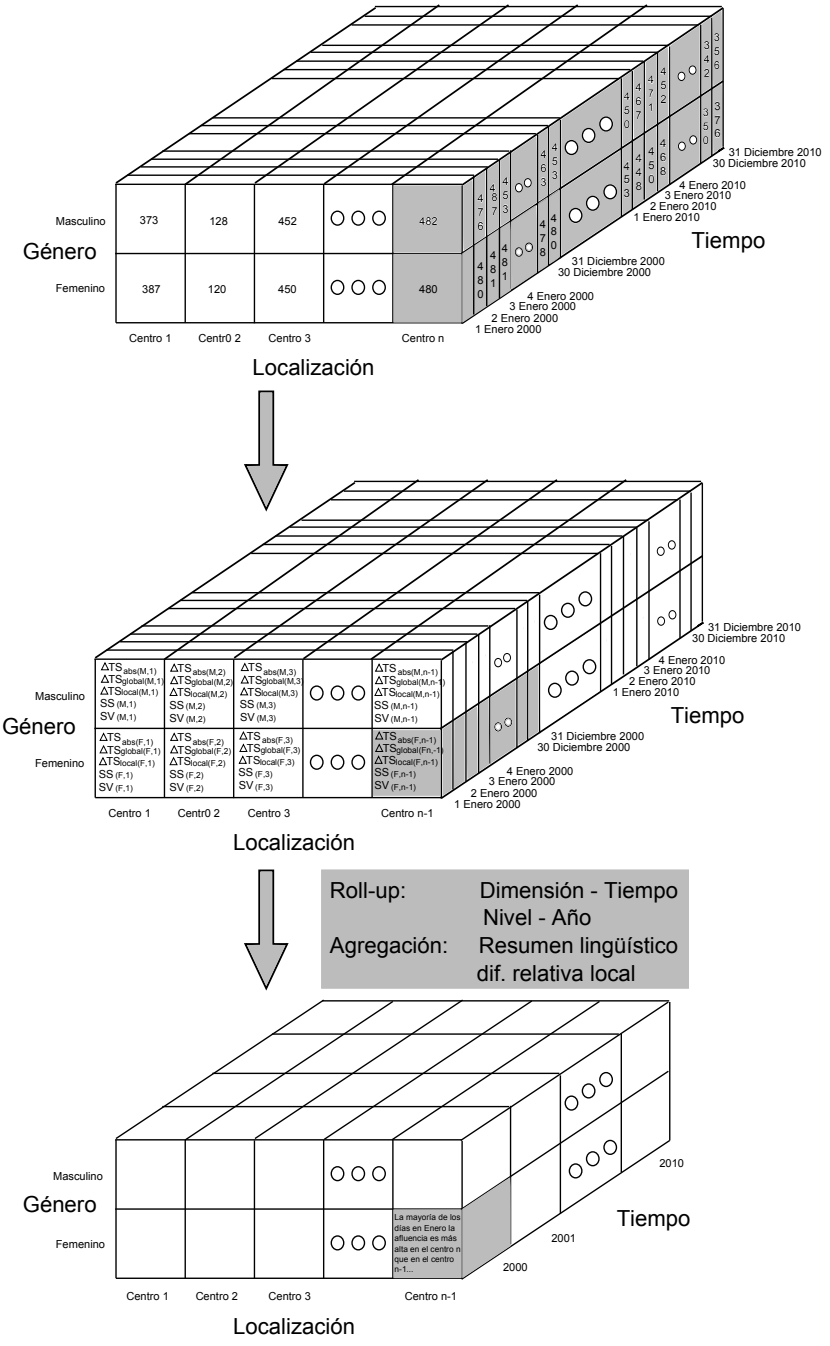


Figura 6.2: Proceso para la incorporación de la funcionalidad de comparación en Linguistic F-Cube Factory, sobre un cubo de datos con dimensiones *género*, *localización* y *tiempo*.

capacidades lingüísticas mejoradas.

6.4. Resumen lingüístico en Linguistic F-Cube Factory

En la sección actual presentaremos la herramienta Linguistic F-Cube Factory para la construcción de resúmenes lingüísticos de cubos de datos que incluyen la dimensión temporal. Gracias al uso de los distintos enfoques Greedy, la interacción con los cubos de datos al realizar resúmenes se hace de forma rápida siendo esto lo ideal para entornos interactivos.

La Figura 6.3 muestra la apariencia de la pantalla principal de la herramienta tal y como la ve el usuario final al acceder a la plataforma. Dicha pantalla se encuentra dividida en dos zonas bien diferenciadas que ofrecen información de diferente naturaleza.

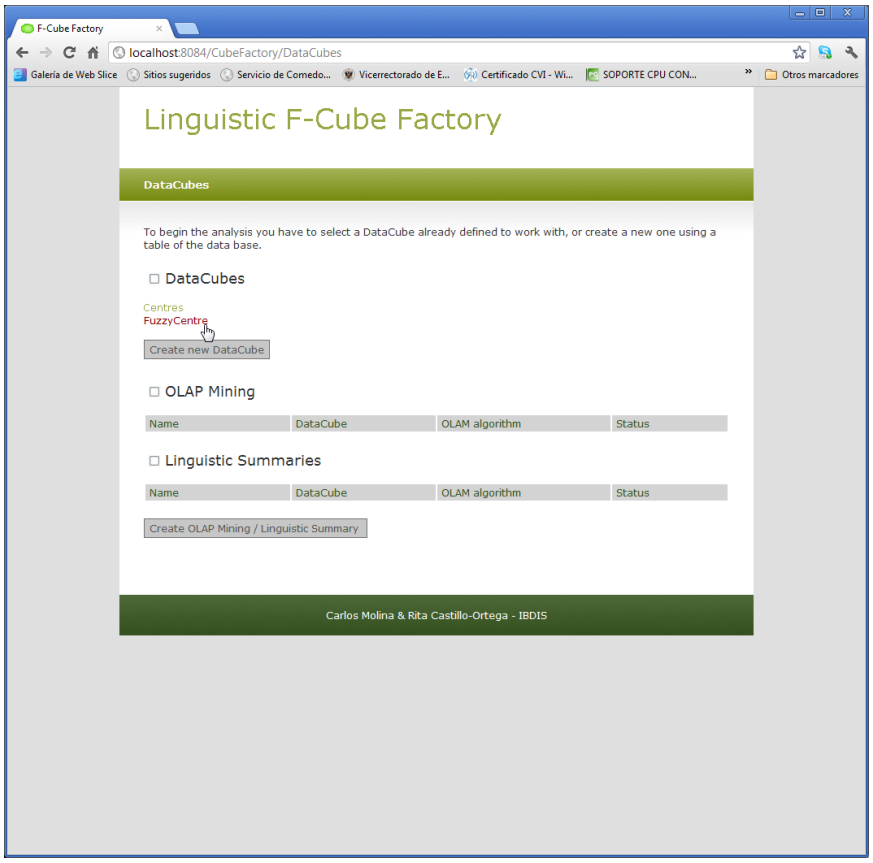


Figura 6.3: Pantalla principal de Linguistic F-Cube Factory.

La primera de las zonas se encuentra dedicada a mostrar una lista de todos los cubos de datos existentes en el sistema. En este caso concreto, el usuario cuenta con dos cubos: *Centres* y *FuzzyCentre*. Como se ha comentado con anterioridad, la herramienta, además de la gestión de cubos de datos, permite su construcción, por lo que existe un botón destinado a iniciar el proceso de creación.

La segunda zona está dedicada a mostrar los resultados fruto de los procesos OLAP de minería de datos ejecutados. Existe una diferenciación clara entre los resúmenes obtenidos como resultado de aplicar operaciones OLAP tradicionales y las basadas en el uso de resúmenes lingüísticos obtenidos al aplicar nuestro modelo de resumen. Recordemos que en el segundo tipo de resultados son cubos de datos especiales en el que se alojan resúmenes lingüísticos en los hechos en lugar de datos numéricos. También en esta zona existe un botón que permite la creación tanto de procesos OLAP clásicos como resúmenes lingüísticos.

Para continuar con la exploración de la herramienta, seleccionaremos uno de los cubos de datos, en este caso el cubo *FuzzyCentre*, para trabajar con él. La siguiente pantalla que el usuario encuentra, la cual vemos en la Figura 6.4, muestra la información relacionada con el cubo de datos seleccionado. De nuevo, esta pantalla se encuentra dividida en zonas con el objetivo de presentar la información de forma clara y ordenada.

En una primera zona se ofrece información de tipo general (*Information*) como el nombre, el tipo o el número de registros del cubo dado. En segundo lugar aparecen listadas todas las operaciones (*Operations*) que el sistema permite llevar a cabo sobre dicho cubo. El usuario puede llevar a cabo operaciones tanto de borrado, como de consulta OLAP básica o elaborada sobre los registros, así como las nuevas funcionalidades de obtención de un resumen lingüístico de series o de comparación de series. A continuación, en la zona nombrada como *Facts*, se ofrece información acerca de los hechos del cubo. Ya por último, en la zona *Dimensions* se muestran las distintas dimensiones que componen el cubo y que describen los hechos. La herramienta ofrece la posibilidad de explorar y editar tanto los hechos como las dimensiones, pudiendo añadirse por ejemplo nuevos niveles en las dimensiones.

Volvamos a la zona de operaciones, más concretamente a la que permite la obtención de resúmenes lingüísticos. Para esta tarea se ofrecen dos posibilidades: realizar la inserción de parámetros en modo experto a través de un sólo formulario, o contar con la ayuda de un asistente que, a través de una serie de pantallas, guiará al usuario no experto durante el proceso de selección de valores para los distintos parámetros que se deben suministrar para configurar adecuadamente la obtención del resumen. Si seleccionamos la opción con asistente, nos llevará a la siguiente pantalla, representada en la Figura 6.5.

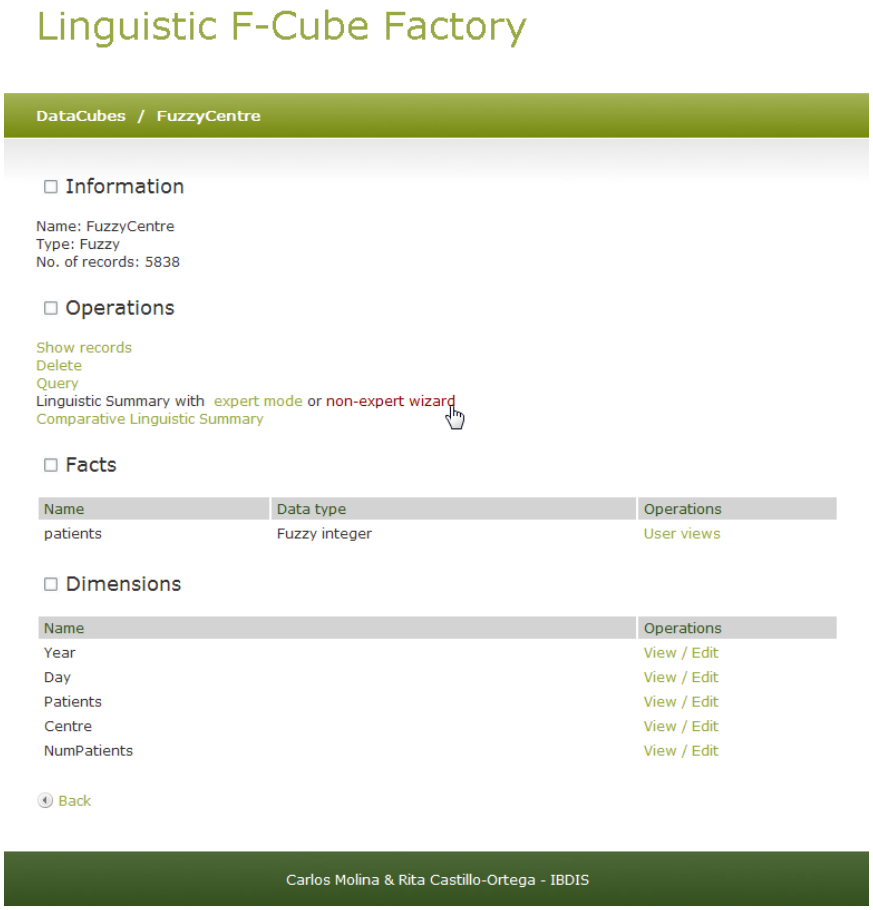


Figura 6.4: Información de un cubo de datos en Linguistic F-Cube Factory.

6.4.1. El asistente para la configuración de resúmenes

La Figura 6.5 es el primer paso de los cinco necesarios para configurar correctamente el marco lingüístico y los diversos parámetros que se usarán en el proceso de resumen. En este primer paso se ofrece una pequeña explicación acerca de la estructura tipo de la sentencia que se va a utilizar para componer el resumen y que se rellenará a lo largo del asistente. Además el usuario deberá introducir el nombre para el nuevo resumen final y elegirá el hecho sobre el que trabajar de entre un desplegable que muestra la totalidad de los hechos. En este caso concreto de ejemplo, se ha llamado al resumen *SummaryN1* y se ha seleccionado el único hecho presente en el cubo, el hecho *patients*.

En la Figura 6.6 se puede ver la pantalla correspondiente al segundo paso de la configuración. En esta ocasión se definirán los parámetros relativos al cuantificador, es decir la componente Q en la sentencia tipo.

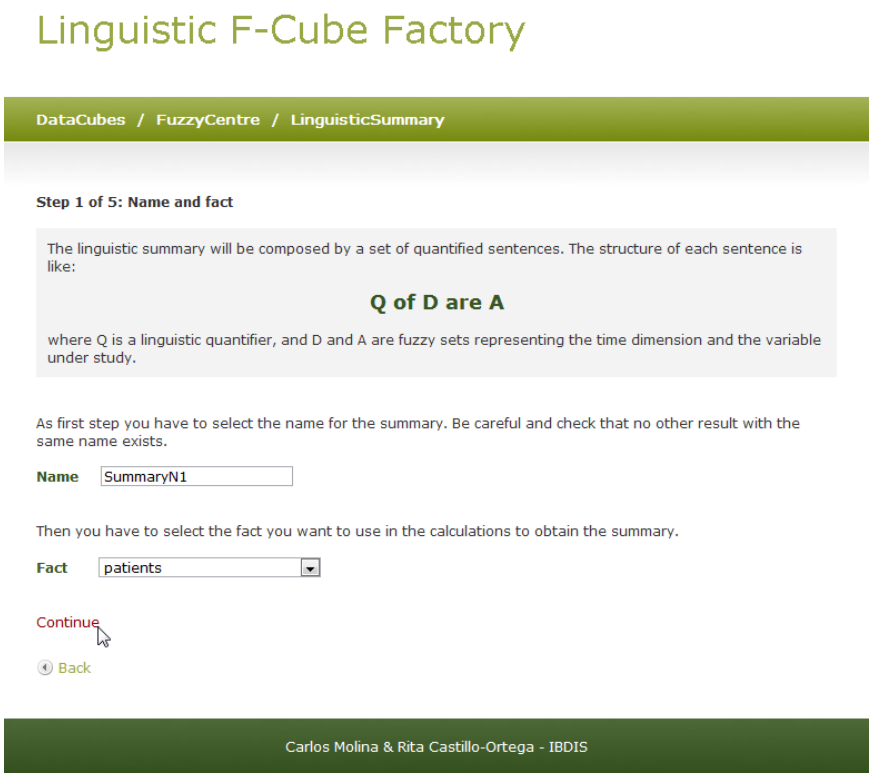


Figura 6.5: Asistente para la creación de resúmenes lingüísticos: Paso 1, información general.

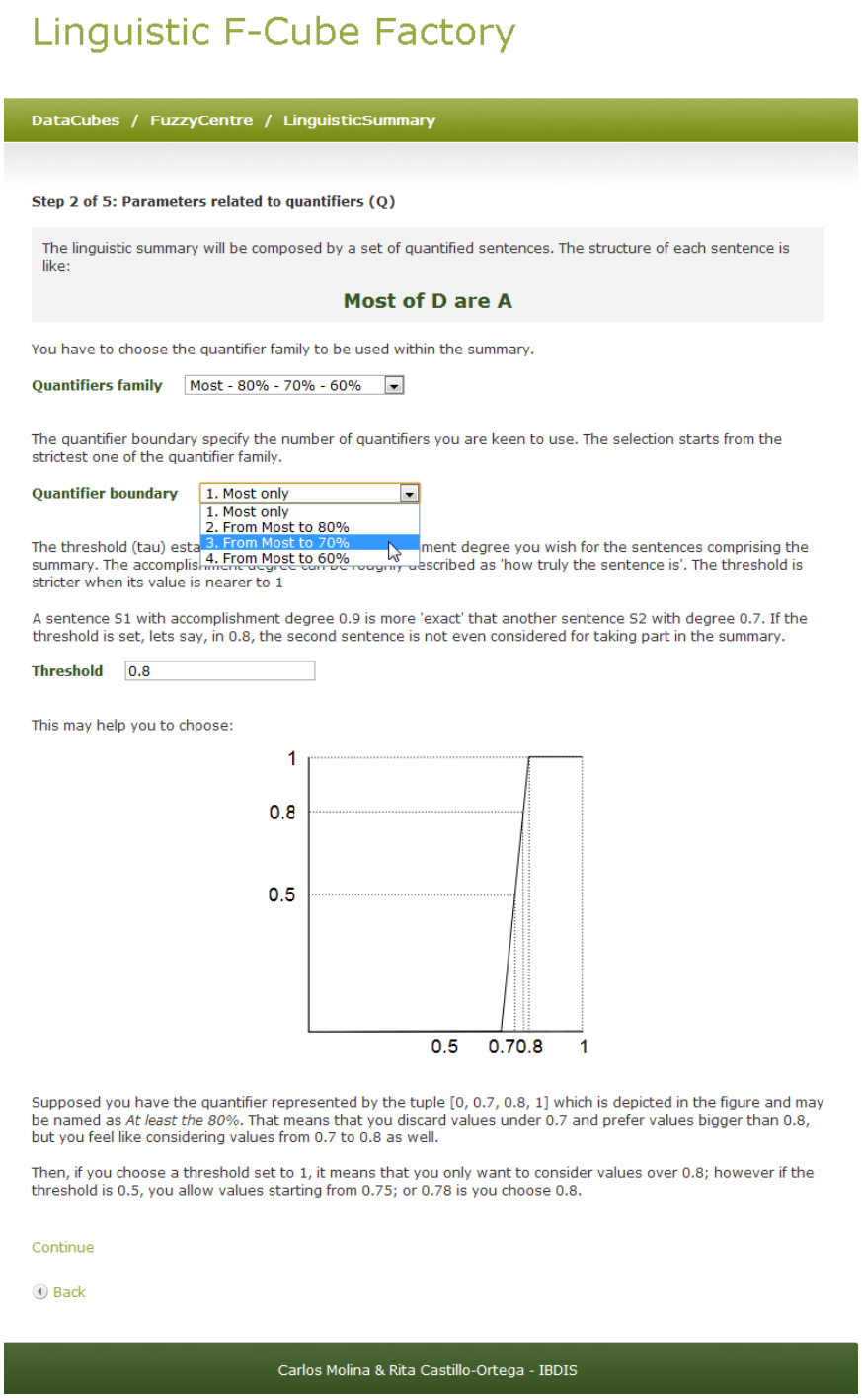


Figura 6.6: Asistente para la creación de resúmenes lingüísticos: Paso 2, parámetros relativos al cuantificador.

En primer lugar el usuario debe seleccionar la familia de cuantificadores que desea utilizar a través de una lista desplegable en la que se muestran todas las opciones disponibles. En segundo lugar, el usuario debe determinar qué cuantificadores está dispuesto a usar, es decir *Qlim* (recordemos, que se usarán desde el más estricto hasta el que menos lo es). A continuación, se debe introducir el valor umbral que determinará el mínimo cumplimiento aceptado para las sentencias que compondrán el resumen. Dicho valor aparece preseleccionado para facilitar la tarea al usuario. Además, para poder afrontar las elecciones de forma más solvente el asistente ofrece información aclaratoria acerca del significado y la repercusión de los distintos parámetros en el resumen final.

Ya en el tercer paso (ver Figura 6.7) se encuentran las decisiones relacionadas con la componente D de la sentencia tipo. En este paso el usuario debe seleccionar, de entre todas las dimensiones del cubo, la que desea considerar como dimensión temporal en el resumen. Para ello cuenta con la ayuda de un desplegable que muestra todas las dimensiones del cubo.

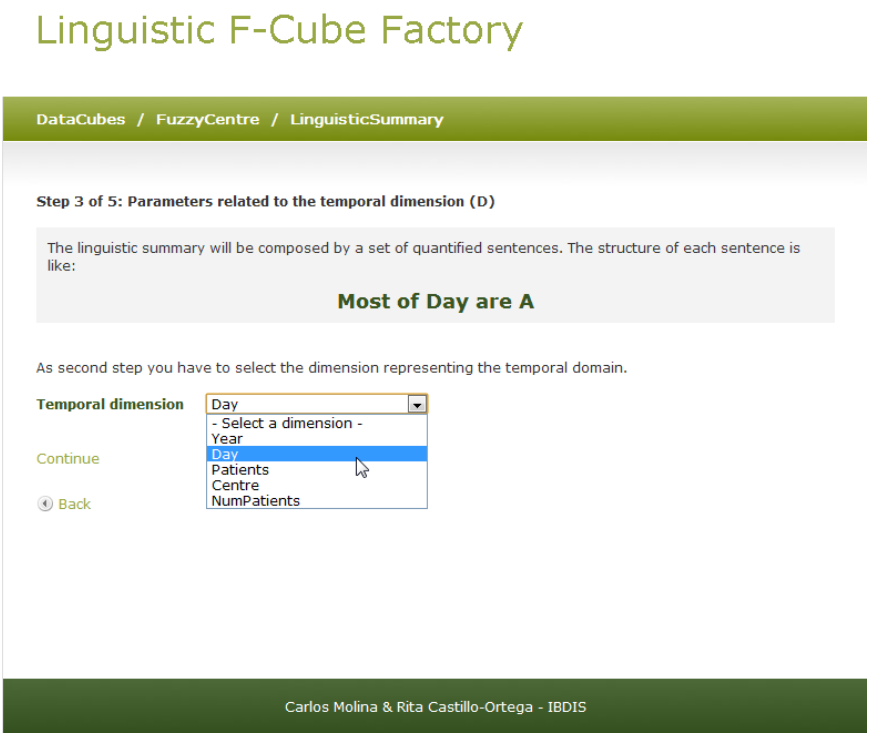


Figura 6.7: Asistente para la creación de resúmenes lingüísticos: Paso 3, parámetros relativos a la dimensión temporal.

En el cuarto paso, representado por la Figura 6.8, se da valor a los parámetros relativos a la variable que se desea describir. En primer lugar se selecciona la dimensión sobre la cual queremos obtener un resumen de forma lingüística a lo largo del tiempo. Del mismo modo que en la pantalla anterior, el usuario selecciona dicha dimensión a través de un desplegable en el que se muestran todas las dimensiones. Una vez hecho esto, el usuario debe seleccionar hasta qué punto está dispuesto a agrupar las etiquetas que se usarán en la descripción, *Glim*. Para ello existe un segundo desplegable que se inicializa con os diferentes niveles disponibles para la dimensión seleccionada en el desplegable anterior.

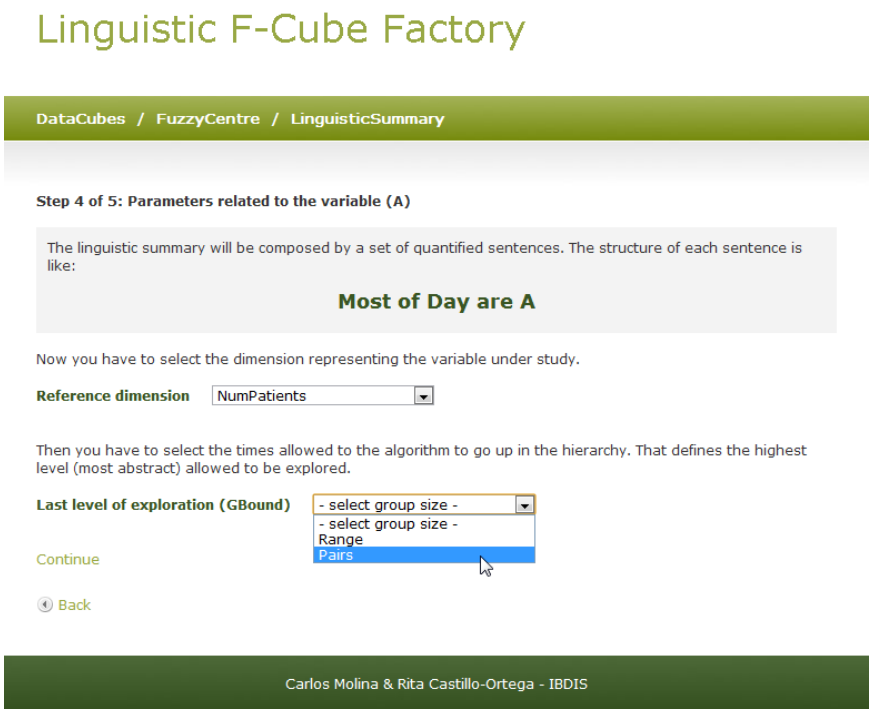


Figura 6.8: Asistente para la creación de resúmenes lingüísticos: Paso 4, parámetros relativos a la variable bajo estudio.

Ya en el quinto y último paso (Figura 6.9) el usuario debe seleccionar qué tipo de alternativa, de entre las dos Greedy que se encuentran implementadas, desea usar. Mediante la ayuda disponible en pantalla se le informa acerca de los diferentes matices semánticos que cada elección conlleva de forma que se le haga más sencillo el proceso de selección.

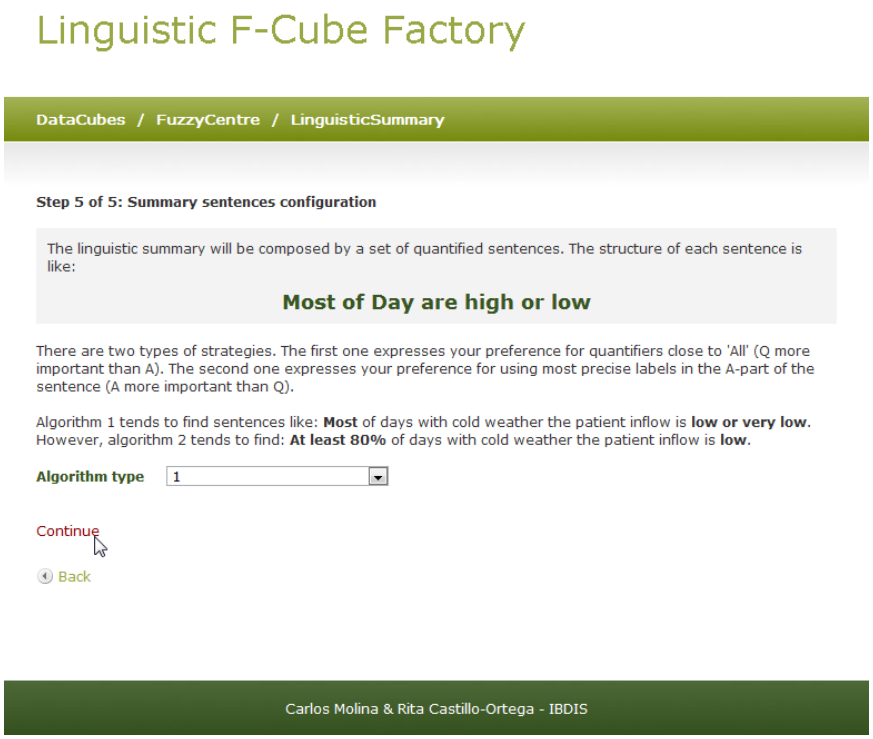


Figura 6.9: Asistente para la creación de resúmenes lingüísticos: Paso 5, preferencias semánticas en las sentencias.

Una vez que se han completado todos los pasos propuestos en el asistente, el usuario llega a una pantalla de "síntesis" en la que se le informa de los valores finales de todos los parámetros. Dicha pantalla puede verse en la Figura 6.10 y servirá tanto para labores de información como de depuración. Además, en esta misma pantalla se informa al usuario acerca de la viabilidad de la operación. Es decir, si el cubo de resúmenes se puede obtener correctamente o si, por el contrario, se observa algún error. En este segundo caso, se informa al usuario de qué tipo de error se ha producido mediante un mensaje adecuado, de modo que se pueda identificar la causa del mismo para facilitar su resolución. Para ver el resultado de la operación de resumen lingüístico ya sólo queda pulsar en *Results*.

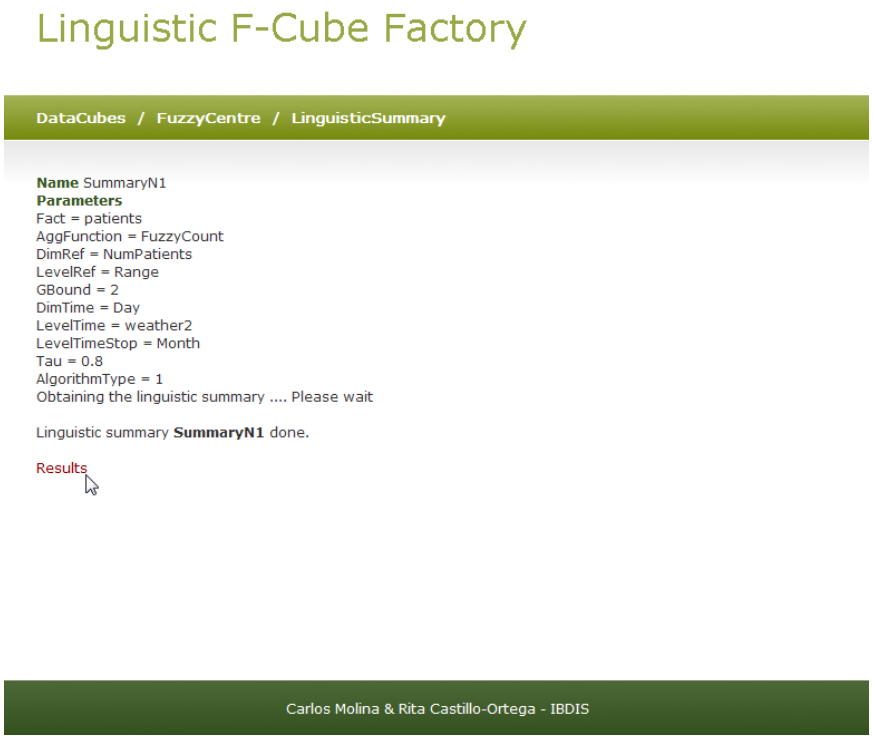


Figura 6.10: Síntesis de parámetros que se considerarán para resumir lingüísticamente.

6.4.2. Visualización de resultados

Recordemos que como resultado de la operación de resumen lingüístico sobre un cubo de datos con dimensión temporal se obtiene un nuevo cubo de datos especial en el que en lugar de tener hechos numéricos, en cada celda encontramos resúmenes lingüísticos de los mismos para el periodo de tiempo seleccionado en la dimensión tiempo. En la Figura 6.11 se muestra el contenido de las celdas de este nuevo cubo de datos.

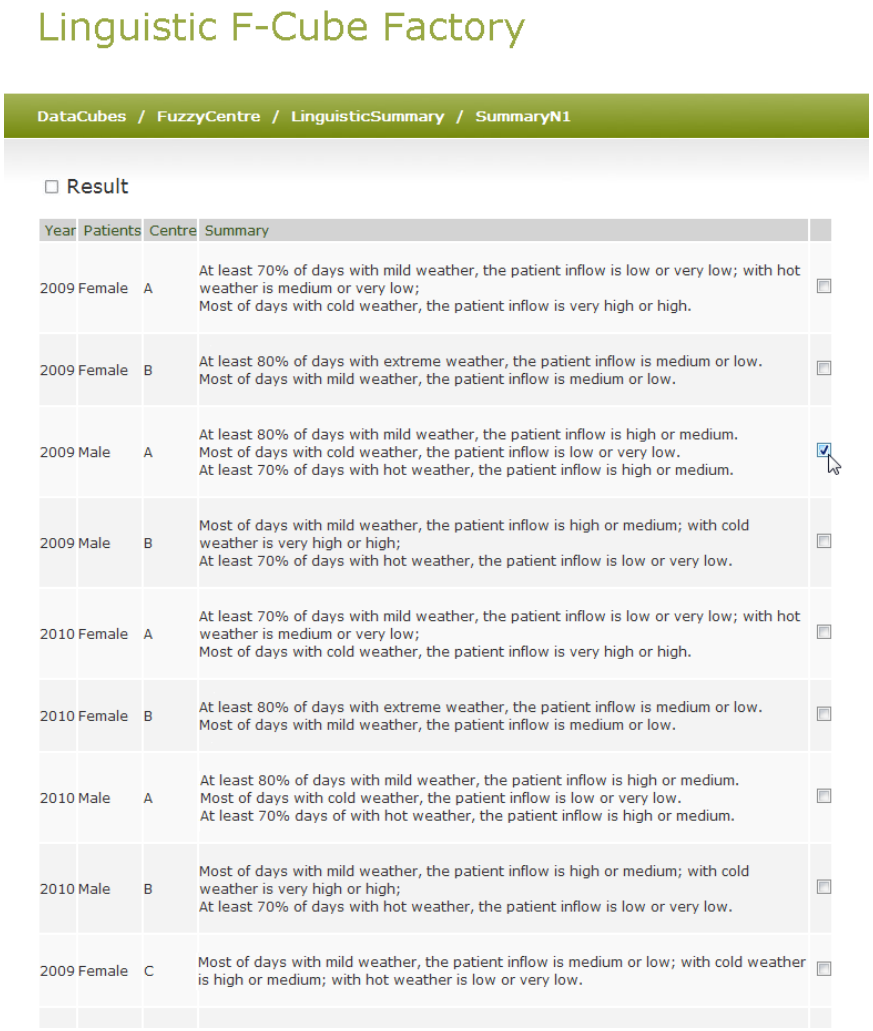


Figura 6.11: Cubo de datos con resúmenes lingüísticos en los hechos.

En este caso, el cubo de datos está compuesto por un resumen por cada año, centro y género de los pacientes. Los resultados que se muestran no son un conjunto de sentencias cuantificadas como tal, sino un párrafo fruto del post-proceso de las mismas que pretende acercar los resultados obtenidos a aquellos construidos por los seres humanos. Si el usuario se encuentra interesado en conocer más detalles de alguno de los resúmenes listados se deberá marcar con un tick en la parte derecha del resumen en cuestión y pulsar el botón *Show details*.

La Figura 6.12 muestra la pantalla resultado del click anterior y en ella se muestra más información acerca del resumen seleccionado. De nuevo en esta ocasión, la pantalla se encuentra dividida en tres grandes zonas de presentación de resultados.

En primer lugar, en la parte superior, se muestra al usuario una representación gráfica de la serie de datos temporal. El eje de las X, o de abcisas, representa la dimensión temporal mientras que el eje de las Y, o eje de ordenadas, muestra la dimensión que se desea describir a lo largo del tiempo.

A continuación, en la zona central, se presenta el resumen de texto fruto del post-proceso de las sentencias cuantificadas que componen el resumen original. Se ofrece al usuario la posibilidad de usar un pequeño reproductor de audio integrado en la página y que le permitirá escuchar el resultado.

En tercer lugar, en la parte inferior de la pantalla, se muestran todas y cada una de las sentencias cuantificadas que componen el resumen. Junto a cada una de ellas aparecerá el grado de cumplimiento de la misma y un botón de radio que nos permitirá seleccionarla.

Al seleccionar una sentencia determinada la gráfica experimentará un cambio. Dicho cambio consistirá en el sombreado de las zonas descritas por la sentencia. Una zona vertical que describe el periodo temporal y una horizontal que representará las etiquetas utilizadas para su descripción. El resultado puede verse en la Figura 6.13. Además, como se puede observar, aparece un segundo reproductor de audio que nos permitirá escuchar la sentencia cuantificada seleccionada de forma individual y no como parte del resumen post-procesado.

La posibilidad de asociar cada una de las sentencias a los datos que soportan la afirmación en el gráfico, ofrece al usuario una herramienta para ponerla en el contexto de la gráfica, si así lo desea. Esta funcionalidad puede también usarse para contrastar la información obtenida con los datos representados en el gráfico.

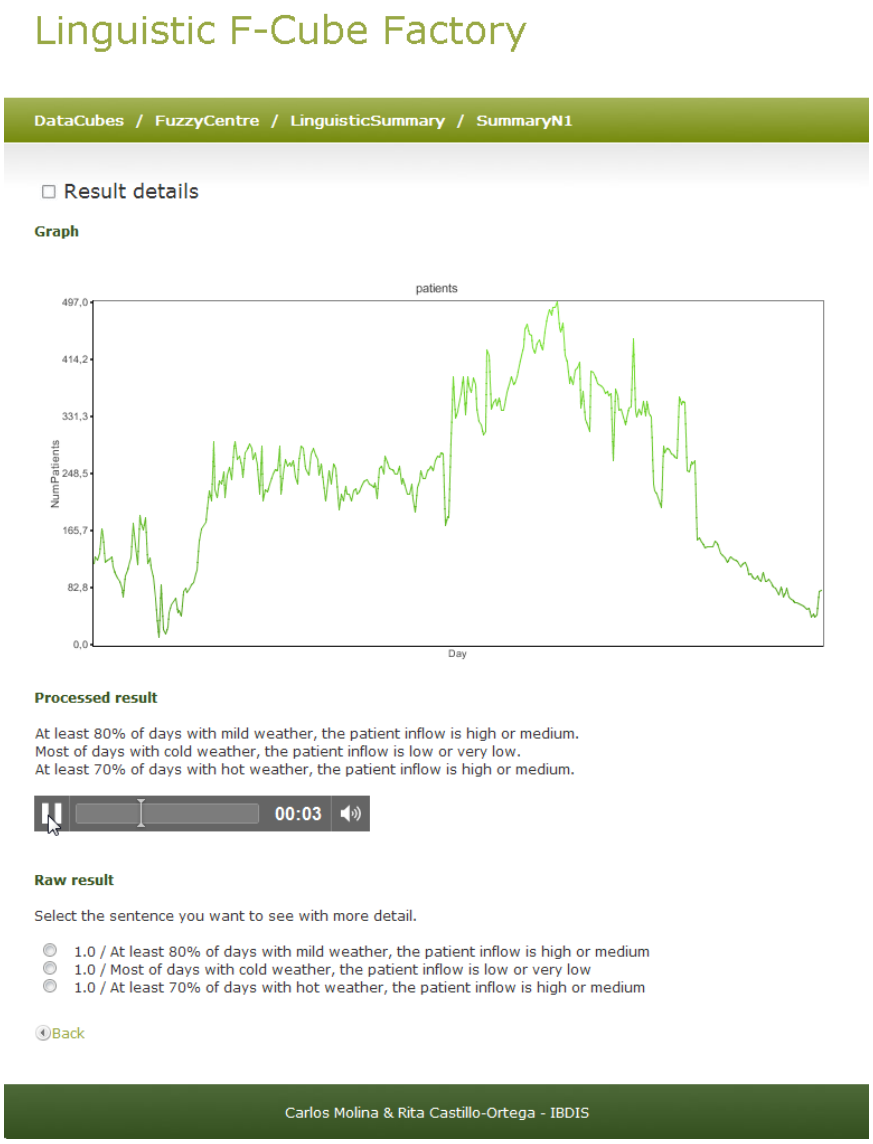


Figura 6.12: Detalles del resumen lingüístico seleccionado (1).

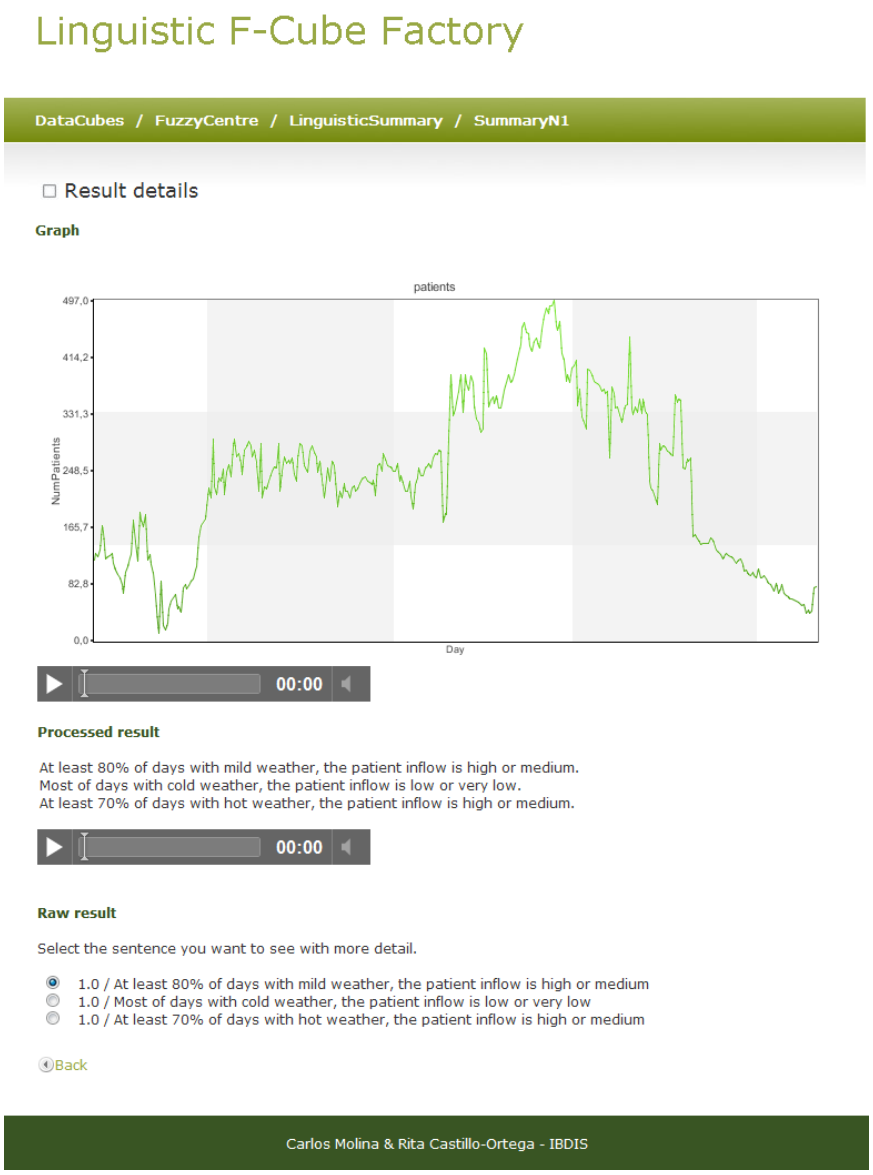


Figura 6.13: Detalles del resumen lingüístico seleccionado (2).

Al volver a la pantalla principal (Figura 6.14), comprobamos que aparece un nuevo resultado de tipo *resumen lingüístico* que efectivamente recibe el nombre de *SummaryN1*. Además del nombre, se ofrece información como el cubo de datos a partir del que se ha construido y al que, por consiguiente, describe, y el estado. En caso de que el proceso no se hubiera llevado a cabo correctamente se notificaría la existencia de un error en el proceso. En caso de que el proceso continúe su ejecución en segundo plano el estado aparecerá será *running*. Esta pantalla permite acceder directamente al cubo para su consulta sin necesidad de volverlo a generar.

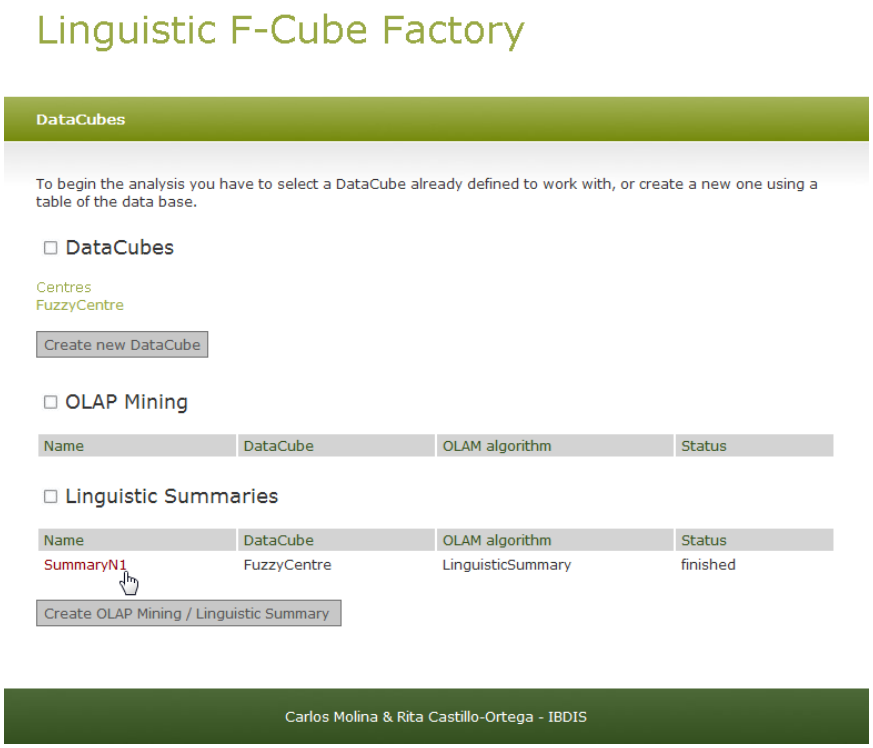


Figura 6.14: Pantalla principal de Linguistic F-Cube Factory.

6.5. Comparación en Linguistic F-Cube Factory

Continuamos trabajando con el cubo de datos *FuzzyCentre* pero en esta ocasión para obtener un cubo de datos de comparación de series de datos temporales. Para ello y, al igual que antes, se deberá dar valor a una serie de parámetros. Además, sobre este nuevo cubo de datos comparación se podrán efectuar operaciones de resumen lingüístico. Veámoslo con más detalle a continuación.

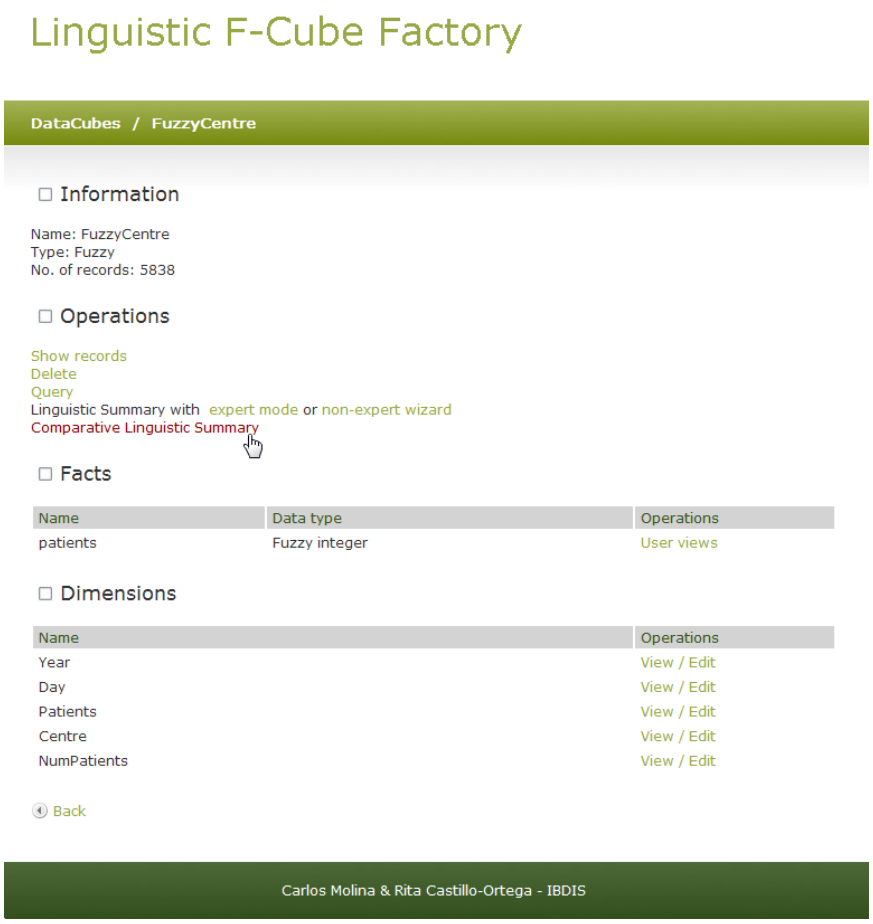


Figura 6.15: Información de un cubo de datos en Linguistic F-Cube Factory.

Como veíamos en la pantalla representada en la Figura 6.4 entre las operaciones disponibles para realizar sobre un cubo de datos dado aparece la posibilidad de construir un resumen de comparación de series. Si seleccionamos dicha operación como vemos en la Figura 6.15 aparecerá en pantalla un nuevo asistente, esta vez para la obtención del nuevo cubo de datos que obtendrá los resúmenes de comparación entre

las series (Figura 6.16).

6.5.1. El asistente de comparación

En la Figura 6.16 podemos ver la pantalla para el asistente de creación del cubo de datos con resúmenes de comparación. En primer lugar se debe introducir el nombre del nuevo cubo de datos que se creará, con cuidado de que no esté siendo usado ya.

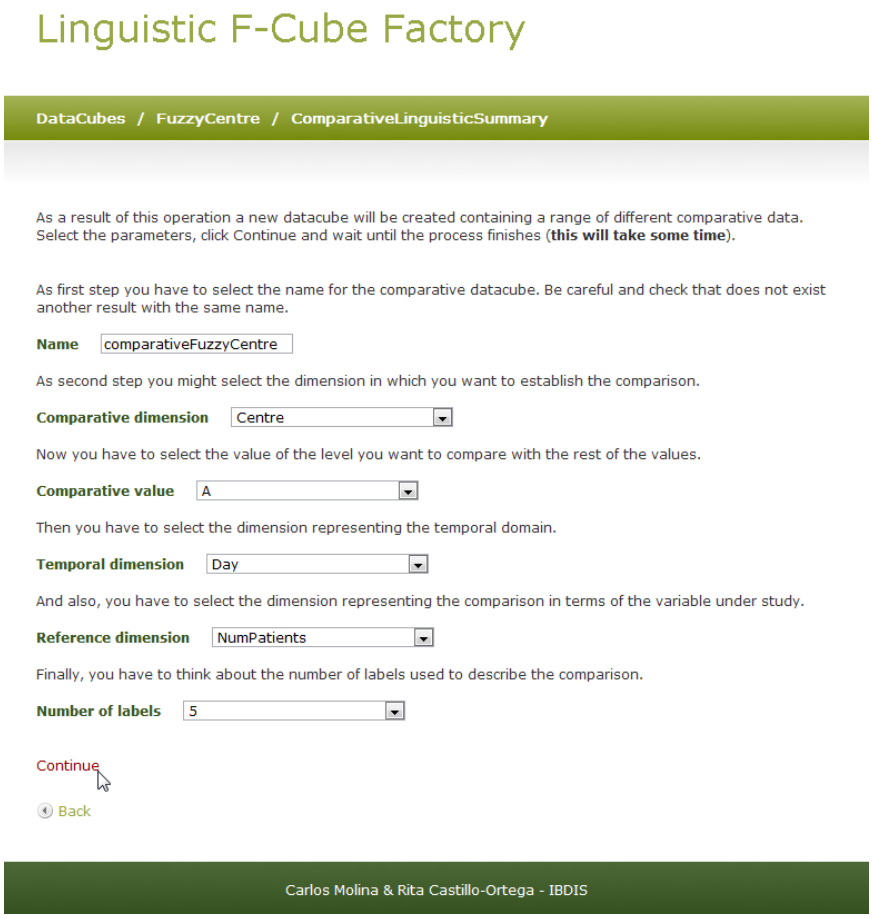


Figura 6.16: Asistente para la creación de resúmenes lingüísticos de comparación.

En segundo lugar se debe determinar cuál de las dimensiones es en la que queremos establecer la comparación, para, seguidamente, seleccionar entre los valores de dicha dimensión el que se comparará con el resto. En este caso se ha elegido la dimensión *Centre* y será el centro *A* el que se compare al resto de centros. Además también se debe elegir una dimensión temporal y una dimensión referencia entre las dimensiones

disponibles. Estas dimensiones tendrán el mismo papel que el que se explicaba para realizar resúmenes lingüísticos. En este caso, la dimensión temporal seleccionada es *Day* y la de referencia *NumPatients*, de modo que se compararán series que describan la afluencia de pacientes a los centros durante un periodo de tiempo determinado por una serie de días.

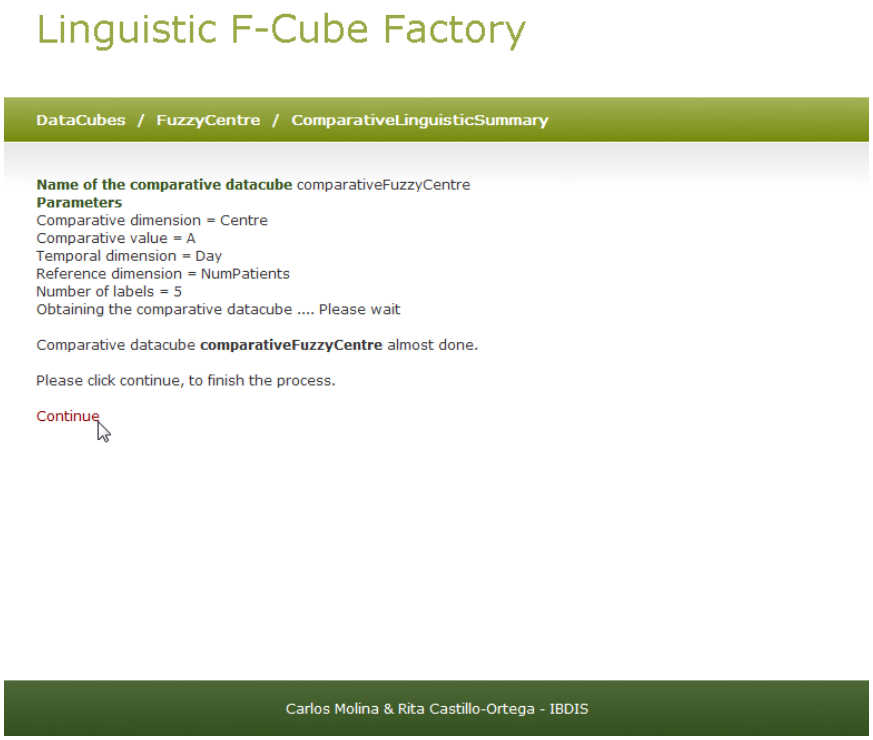


Figura 6.17: Síntesis de parámetros que se considerarán para resumir lingüísticamente la comparación de series (1).

Por último se debe seleccionar el número de etiquetas que el usuario desea utilizar para describir la comparación de series. En el desplegable habilitado para ello se ofrece la posibilidad de elegir el número de etiquetas que se quieren usar para describir el resultado de la comparación.

Al igual que en el caso de resumen lingüístico de una sola serie, la siguiente pantalla muestra una síntesis de los parámetros que se usarán para obtener el nuevo cubo de datos comparación, ver Figura 6.17. Esta misma pantalla informa al usuario sobre si se ha tenido éxito en la primera fase de la construcción del cubo de datos. Para continuar con la construcción el usuario debe pulsar *Continue*.

Una vez creado el cubo de datos en sí, se construyen las etiquetas necesarias para la

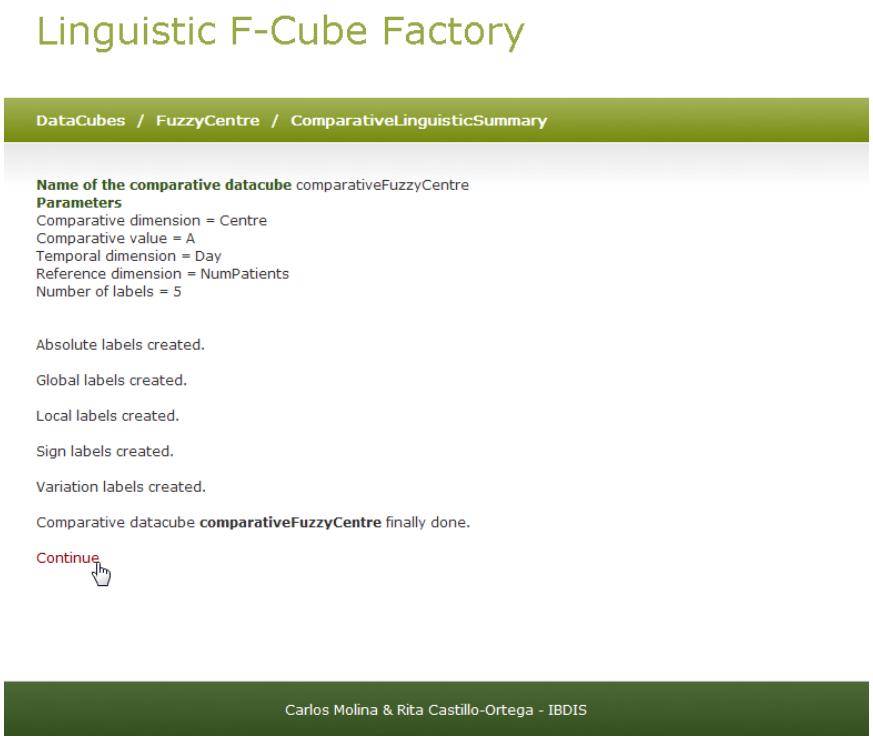


Figura 6.18: Síntesis de parámetros que se considerarán para resumir lingüísticamente la comparación de series (2).

posterior descripción de la comparación, ver Figura 6.18. En esta pantalla se informa al usuario acerca del proceso de creación de las mismas. Si existiera algún error a crear las distintas particiones se informaría al usuario en esta misma pantalla con el fin de que pudiera subsanar los errores.

6.5.2. Interacción con el cubo de resumen

La Figura 6.19 muestra la pantalla que ofrece la información del nuevo cubo de datos. Vemos que para este cubo aparecen los mismos hechos y dimensiones que ya aparecían para el cubo origen (Figura 6.4) pero que además de éstas se han creado nuevas dimensiones. Existe una nueva dimensión por cada tipo de enfoque presentado para hallar la diferencia entre series de datos. Las dimensiones han sido nombradas automáticamente a partir del nombre base de la dimensión de referencia y una partícula que identificará que tipo de datos contiene. *NumPatients_Abs*, *NumPatients_Global* y *NumPatients_Local* representan los enfoques basados en valor tanto absoluto como relativos, y *NumPatients_Sign* y *NumPatients_Magnitude* para los enfoques basados

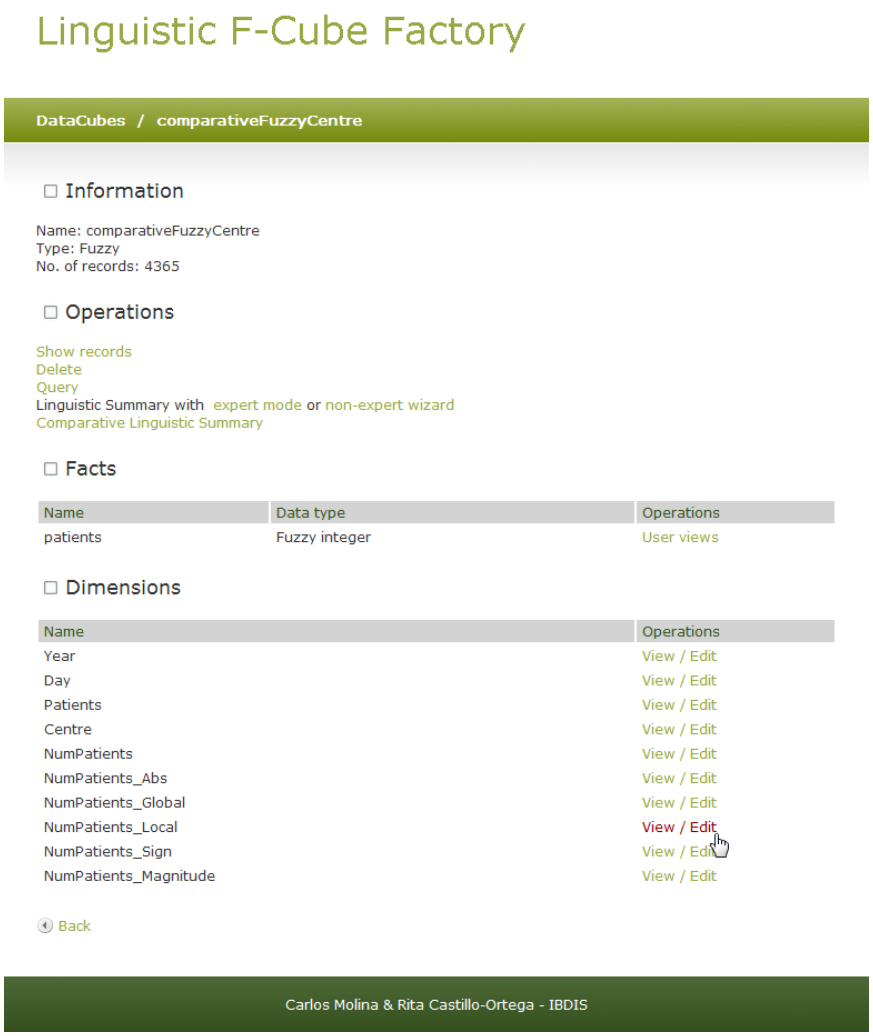


Figura 6.19: Información de un cubo de datos en Linguistic F-Cube Factory.

en cambios locales.

Si exploramos una de las dimensiones, por ejemplo *NumPatients_Local*, llegamos a la pantalla que muestra la Figura 6.20. En ella se muestran los distintos niveles definidos para la dimensión y los métodos de agregación que se utilizarán para trabajar con ella. Dichos niveles (excepto el nivel base), y las etiquetas que contienen, son los que se crearon automáticamente durante la segunda etapa del proceso de construcción.

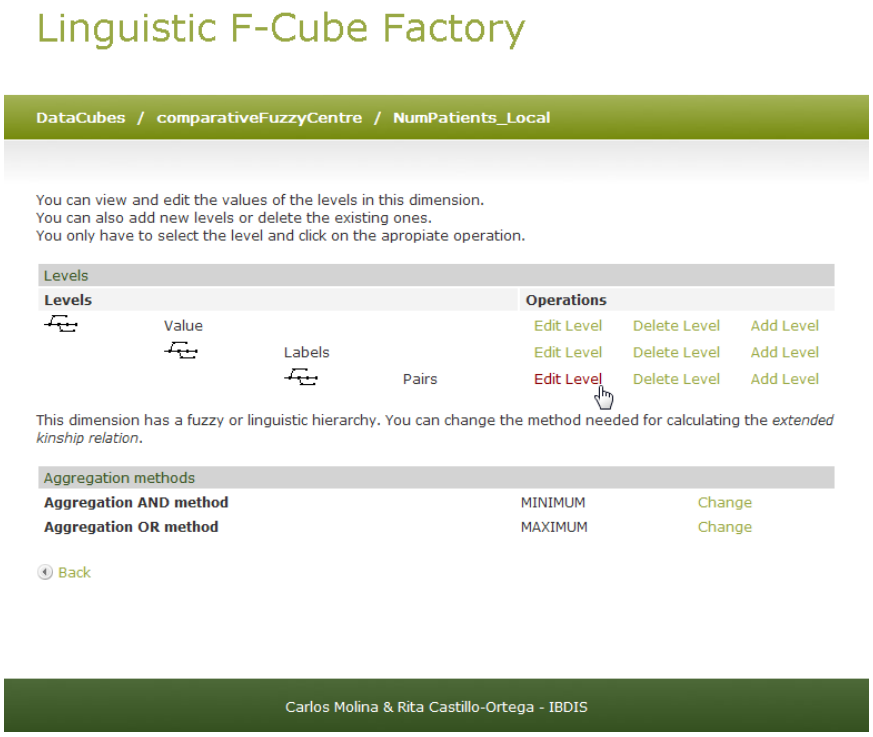


Figura 6.20: Detalle de información de una dimensión determinada en Linguistic F-Cube Factory.

En la Figura 6.21 se muestra un detalle de las etiquetas definidas para el nivel *Pairs* de la dimensión *NumPatients_ Local*. En la pantalla representada vemos una lista con las etiquetas definidas en el nivel y una serie de operaciones que se pueden realizar sobre cada una, e, incluso, la posibilidad de añadir nuevas etiquetas a la partición del nivel.

Como ya hemos comentado en varias ocasiones, el cubo de datos con las series diferencia puede ser a su vez sometido a la operación de resumen lingüístico de datos para obtener resúmenes en lenguaje natural de los mismos.

Desde la pantalla anterior, pulsando *Back*, volvemos a la pantalla de información del nuevo cubo de datos (ver Figura 6.22) y seleccionamos la opción de realizar resumen lingüístico sobre el mismo.

De nuevo decidimos usar el asistente para que nos ayude en la elección de los parámetros más relevantes. Seleccionamos un nombre para el resultado que no esté siendo utilizado para otro resultado y de nuevo seleccionamos el único hecho disponible en el cubo de datos (Figura 6.23).

El siguiente paso es el de la selección de los parámetros relacionados con el cuantificador y la cuantificación de la sentencia (ver Figura 6.24).

Volvemos a escoger entre una serie de diferentes subconjuntos de familias coherentes de cuantificadores y a continuación seleccionamos hasta qué punto estamos dispuestos a usar cuantificadores menos estrictos. Por último seleccionamos el umbral que nos marcará el mínimo valor de grado de cumplimiento que tendrán las sentencias que aparezcan en el resumen final.

El tercer paso de los cinco necesarios consiste en la selección de los parámetros relativos a la dimensión temporal (pantalla representada en la Figura 6.25). Es decir, seleccionar de entre todas las dimensiones del cubo de datos, aquella que deseamos utilizar como la dimensión que describa el tiempo en nuestro resumen.

Como penúltimo paso nos encontramos con la pantalla representada en la Figura 6.26 y que será la que nos pida los datos relativos a la dimensión de referencia o de descripción de la variable bajo estudio. En este paso se debe seleccionar la dimensión deseada de entre todas las que componen el cubo de datos y seleccionar el nivel de granularidad hasta el que estamos dispuestos a llegar. En este caso y como deseamos realizar un resumen de la diferencia de series, deberemos seleccionar aquellas que se crearon automáticamente durante el proceso de creación del cubo de datos comparación. En este ejemplo, se ha optado por la obtención de un resumen lingüístico que describa la diferencia entre series en términos relativos locales. Además se permite el uso de parejas de etiquetas en la descripción final. Las parejas de etiquetas, como ya vimos anteriormente, se encuentran definidas en el nivel *Pairs*.

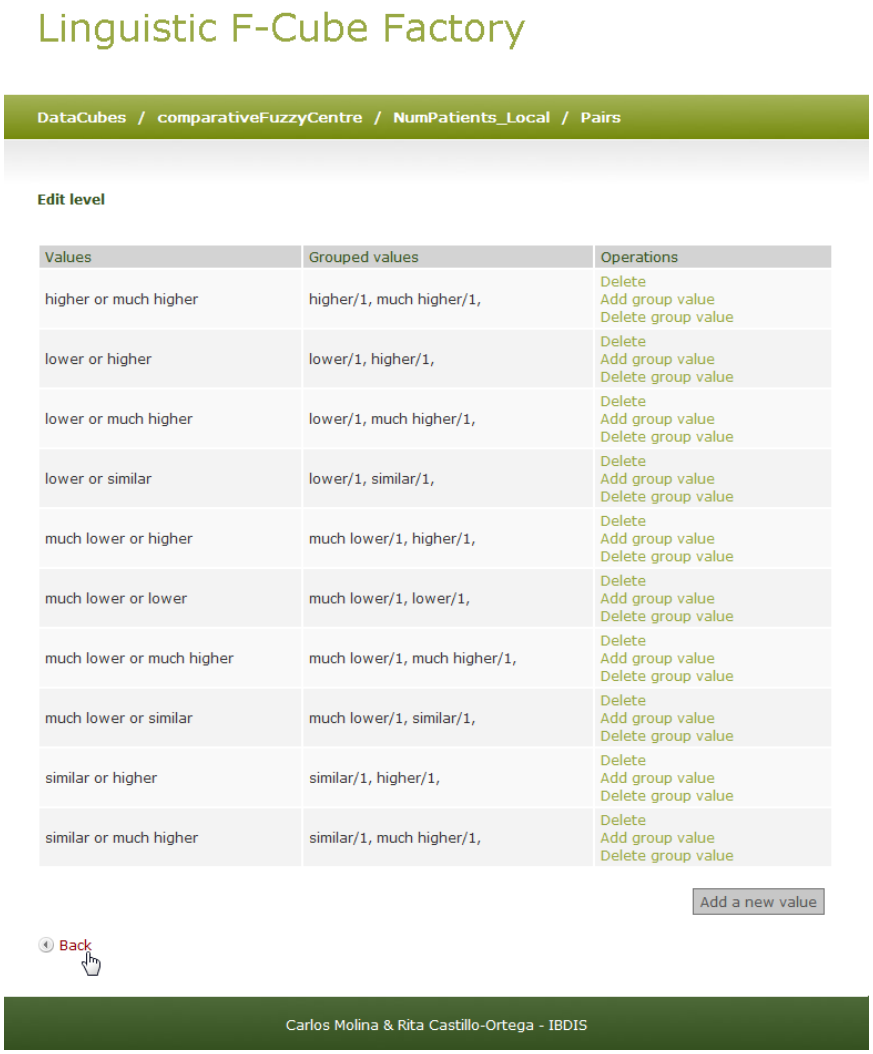


Figura 6.21: Detalle de información de un nivel determinado en Linguistic F-Cube Factory.

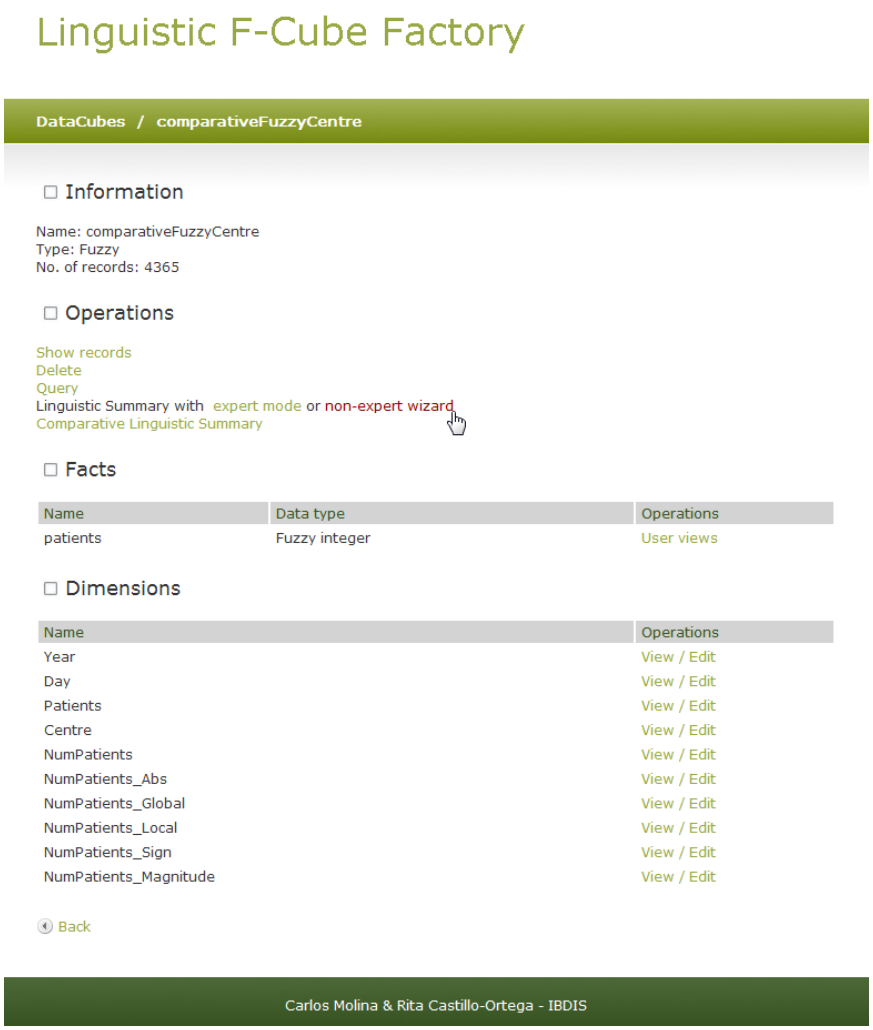


Figura 6.22: Información de un cubo de datos en Linguistic F-Cube Factory.

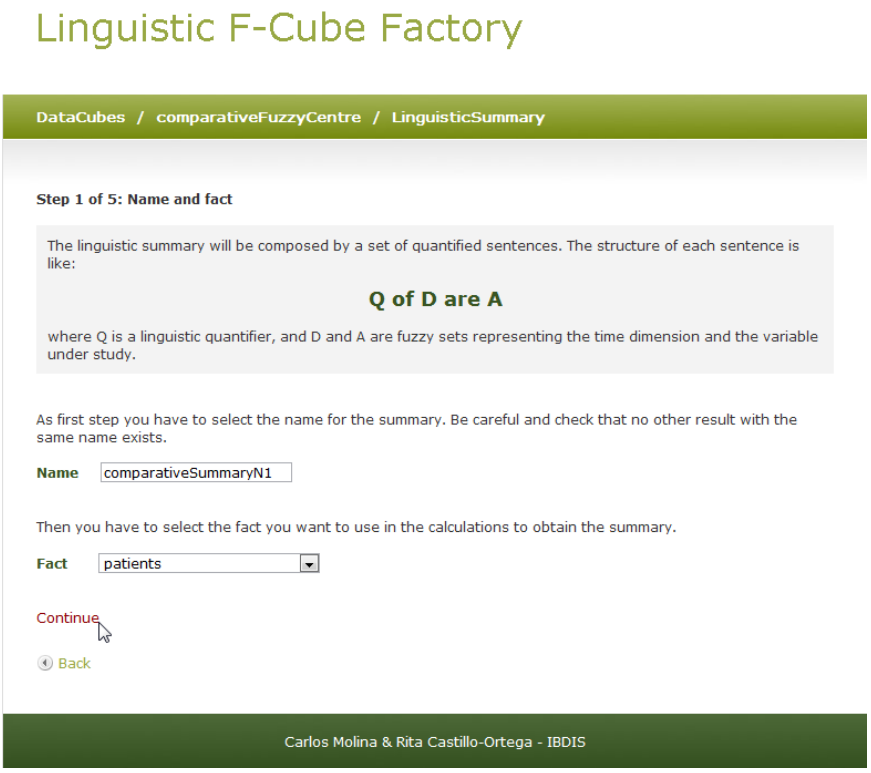


Figura 6.23: Asistente para la creación de resúmenes lingüísticos: Paso 1, información general.

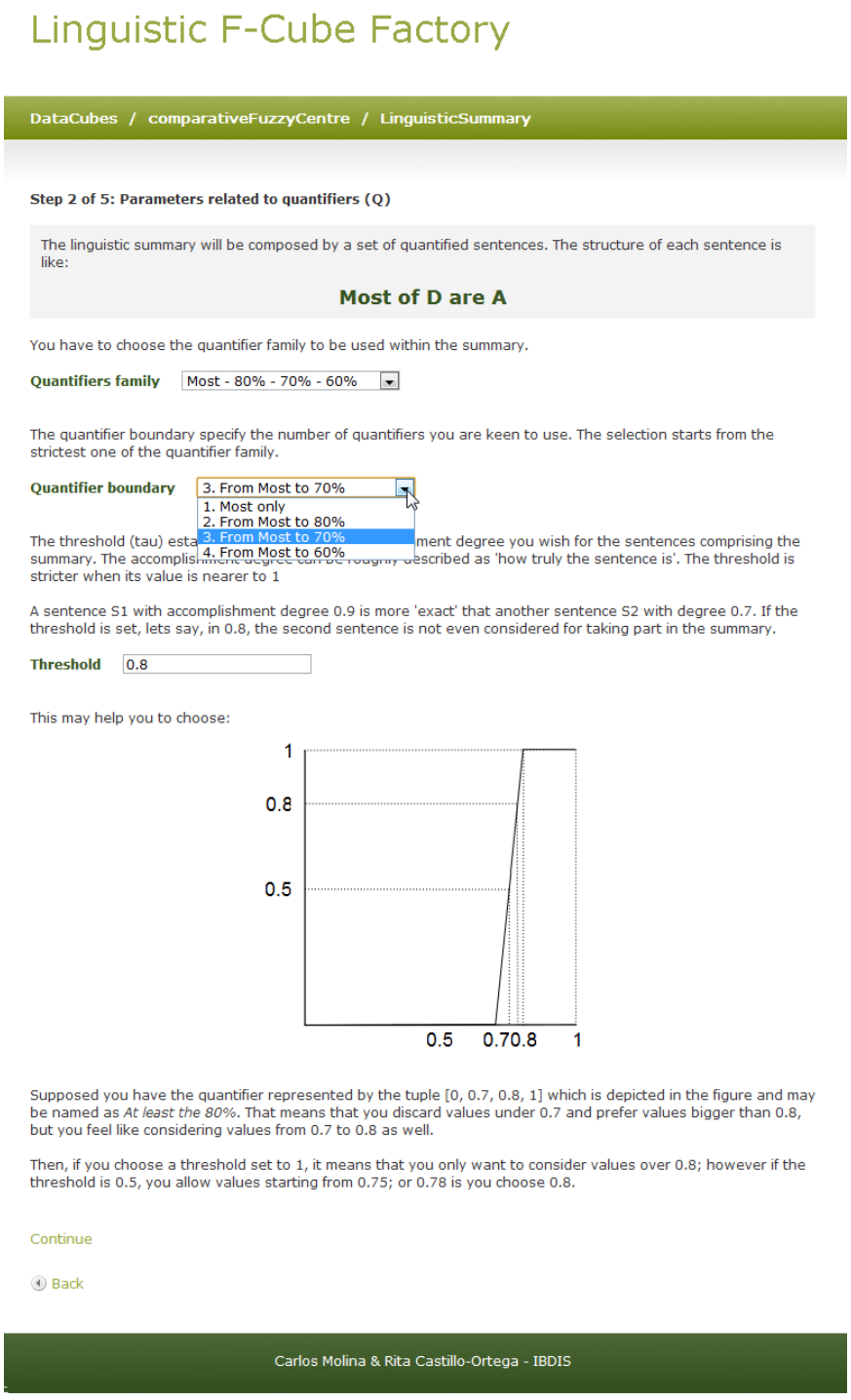


Figura 6.24: Asistente para la creación de resúmenes lingüísticos: Paso 2, parámetros relativos al cuantificador.

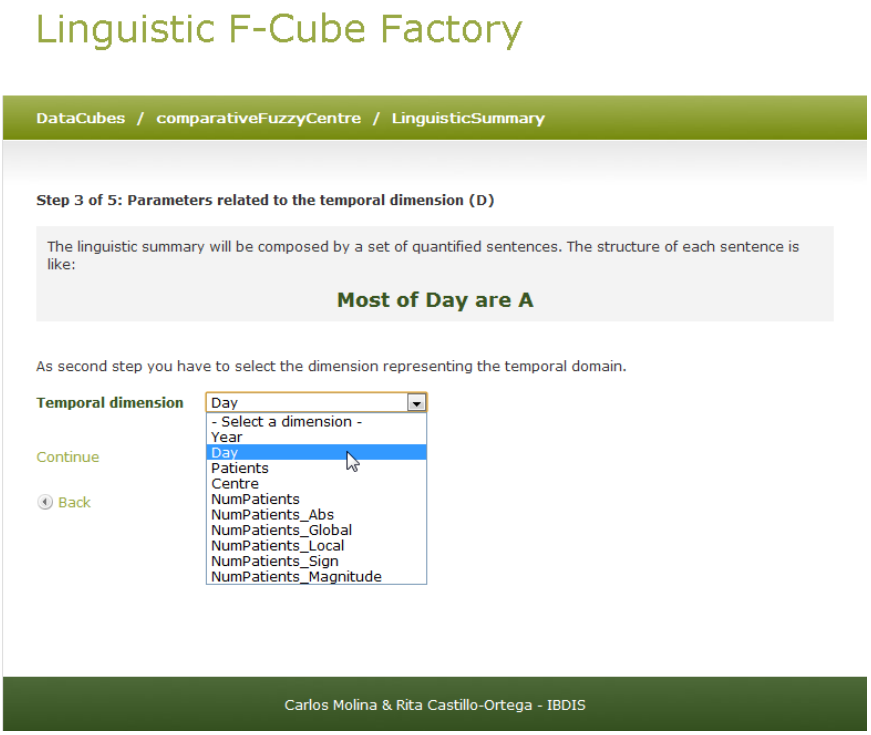


Figura 6.25: Asistente para la creación de resúmenes lingüísticos: Paso 3, parámetros relativos a la dimensión temporal.

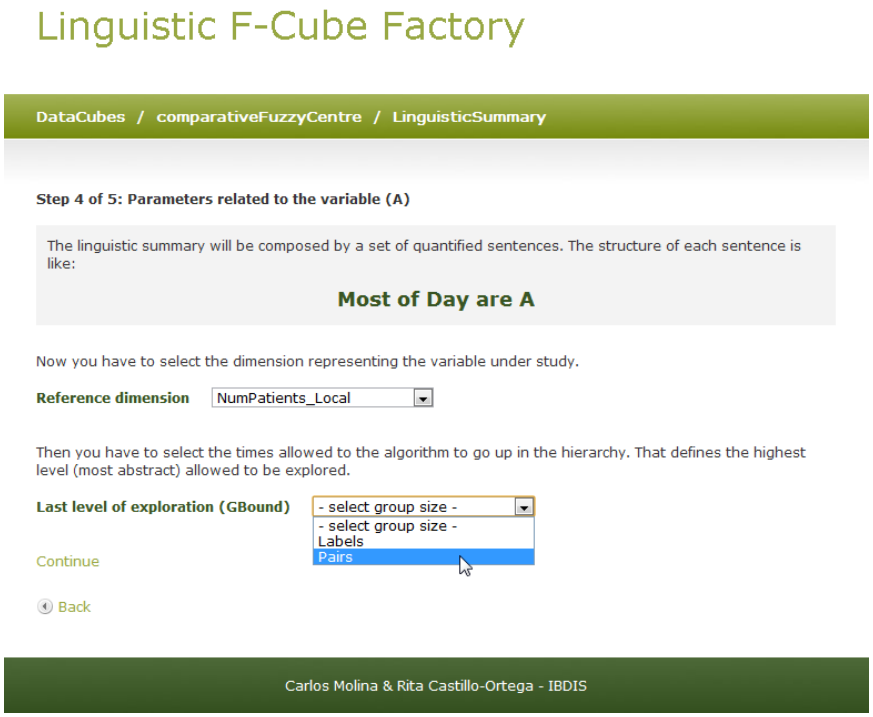


Figura 6.26: Asistente para la creación de resúmenes lingüísticos: Paso 4, parámetros relativos a la variable bajo estudio.

Como quinto y último paso, el usuario debe establecer su preferencia por uno de los dos enfoques Greedy implementados en el modelo (Figura 6.27).

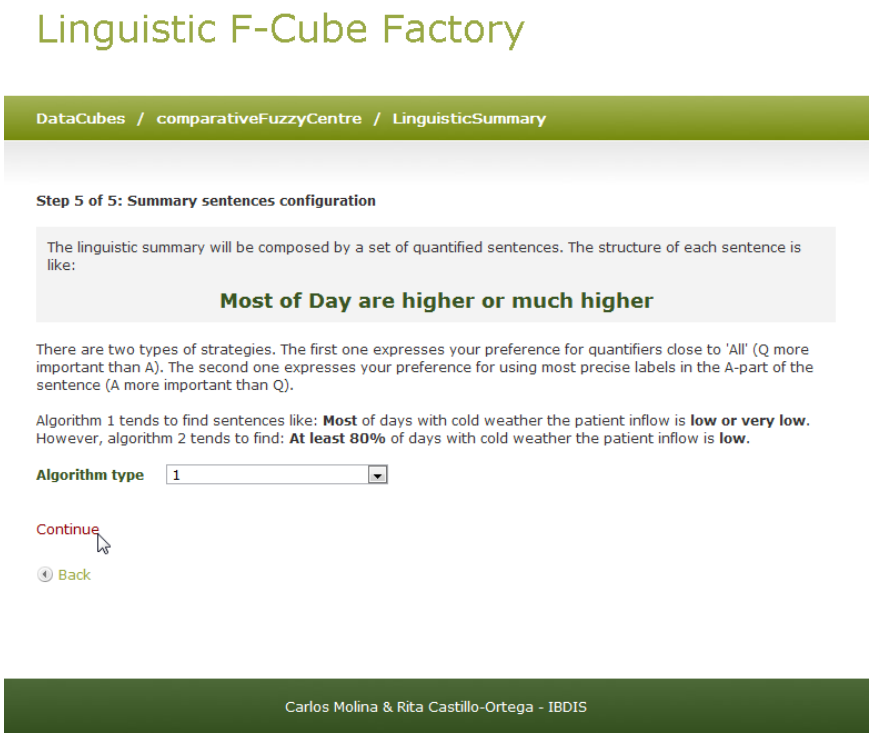


Figura 6.27: Asistente para la creación de resúmenes lingüísticos: Paso 5, preferencias semánticas en las sentencias.

Una vez terminado el proceso de ajuste de parámetros llegamos a la pantalla de síntesis, presentada en la Figura 6.28 y que nos informa de los valores que se le han asignado a los diferentes parámetros que se han tenido en cuenta para la elaboración del resumen. En este caso no se nos muestra ningún mensaje de error, por el contrario se nos informa de que el proceso se ha llevado a cabo con éxito, de modo que ya tenemos disponible para su consulta el cubo de datos resultado.

La Figura 6.29 muestra la pantalla que lista los resúmenes de comparación obtenidos y que se encuentran en el cubo de datos resultado. Los resultados se ordenan en forma de tabla en la aparecen las dimensiones en distintas columnas. Podemos ver las dimensiones *Year*, *Patients* y *Centre*, es decir, aquellas que no son la temporal (que es la que se ha usado para realizar el resumen, *Day*). Tampoco podemos ver las dimensiones de comparación ya que sólo se usa la seleccionada para hacer el resumen (que desaparece) y las demás se desechan en esta operación ya que no nos interesan. A continuación, se aprecia la columna en la que se muestra el hecho de las celdas,

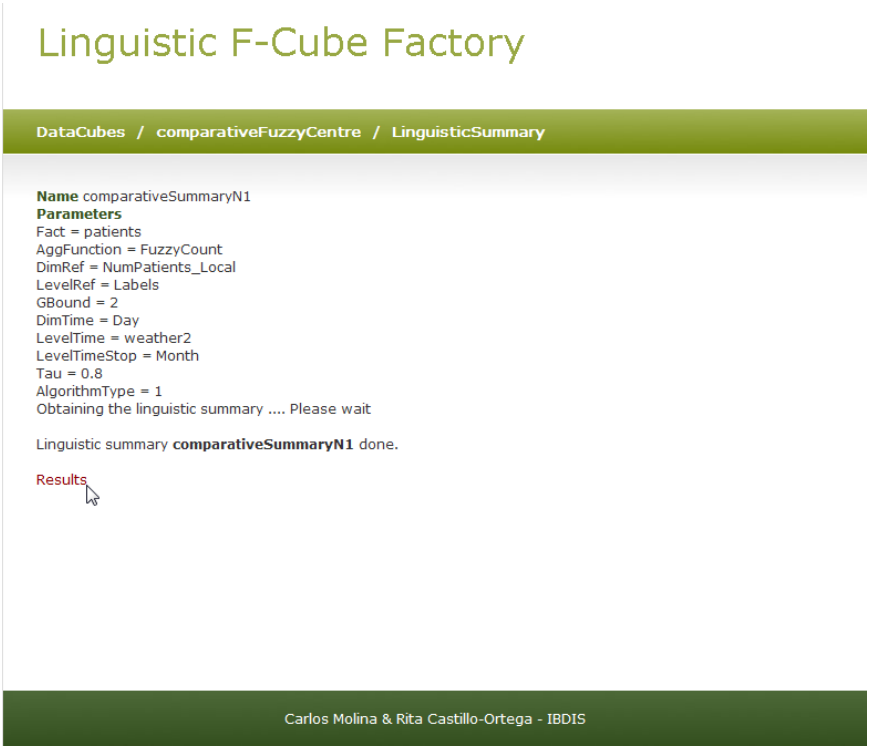


Figura 6.28: Síntesis de parámetros que se considerarán para resumir lingüísticamente.

esto es, los resúmenes lingüísticos que describen la diferencia de series. Por último, encontramos una columna con una serie de casillas en las que el usuario puede clickar para obtener más detalles acerca del resumen seleccionado.

Las Figuras 6.30 y 6.31 muestran información extra relacionada con el resumen seleccionado. En la primera de las pantallas, como ya vimos en la sección anterior (Sección 6.4) aparecen tres zonas bien diferenciadas en las que se ofrece información de distinto tipo acerca del resumen. En primer lugar se ofrece la representación gráfica de la serie que se ha resumido, es decir, la serie diferencia. A continuación se vuelve a mostrar el resumen en su versión post-procesada, pero esta vez con la utilidad adicional de contar con un reproductor audio que es capaz de reproducir dicho resumen. Por último, aparece el resumen tal y como es obtenido por el modelo, donde cada sentencia cuantificada presenta además el grado de cumplimiento de la misma. Existe la posibilidad de obtener todavía más información acerca de estas sentencias individuales. Al seleccionar una sentencia, los puntos de la gráfica que apoyan dicha información aparecen sombreados, y se ofrece la posibilidad de reproducir la sentencia.

Year	Patients	Centre	Summary	
2009	Female	B	At least 70% of days with mild weather, the patient inflow is higher or much higher in centre B than in centre A. At least 80% of days with cold weather, the patient inflow is lower or higher in centre B than in centre A; in September is higher or much higher. Most of days in June, the patient inflow is higher or much higher in centre B than in centre A; in May is higher. In August, and in July the comparison results are highly variable.	☐
2009	Female	C	At least 70% of days with mild weather, the patient inflow is higher or much higher in centre C than in centre A; with cold weather is lower or higher; with hot weather is much lower or higher.	☐
2009	Male	B	Most of days with cold weather, and in November, the patient inflow is higher or much higher in centre B than in centre A. At least 70% of days with hot weather, and in September, the patient inflow is much lower or lower in centre B than in centre A. In April, in March, in May, and in October the comparison results are highly variable.	☑
2009	Male	C	Most of days with cold weather, the patient inflow is higher or much higher in centre C than in centre A. At least 80% of days with hot weather, the patient inflow is much lower or lower in centre C than in centre A. At least 70% of days with hot to cold weather, the patient inflow is lower or higher in centre C than in centre A. In April, in March, and in May the comparison results are highly variable.	☐
2010	Female	B	At least 70% of days with mild weather, the patient inflow is higher or much higher in centre B than in centre A. At least 80% of days with cold weather, the patient inflow is lower or higher in centre B than in centre A; in September are higher or much higher. Most of days in June, the patient inflow is higher or much higher in centre B than in centre A; in May are higher. In August, and in July the comparison results are highly variable.	☐
2010	Female	C	At least 70% of days with mild weather, the patient inflow is higher or much higher in centre C than in centre A; with cold weather is lower or higher; with hot weather is much lower or higher.	☐
2010	Male	B	Most of days with cold weather, and in November, the patient inflow is higher or much higher in centre B than in centre A. At least 80% of days with hot weather, the patient inflow is much lower or lower in centre B than in centre A. At least 70% of days in September, the patient inflow is much lower or lower in centre B than in centre A. In April, in March, in May, and in October the comparison results are highly variable.	☐
2010	Male	C	Most of days with cold weather, the patient inflow is higher or much higher in centre C than in centre A. At least 80% of days with hot weather, the patient inflow is much lower or lower in centre C than in centre A. At least 70% of days with hot to cold weather, the patient inflow is lower or higher in centre C than in centre A. In April, in March, and in May the comparison results are highly variable.	☐

Figura 6.29: Cubo de datos con resúmenes lingüísticos de comparación en los hechos.

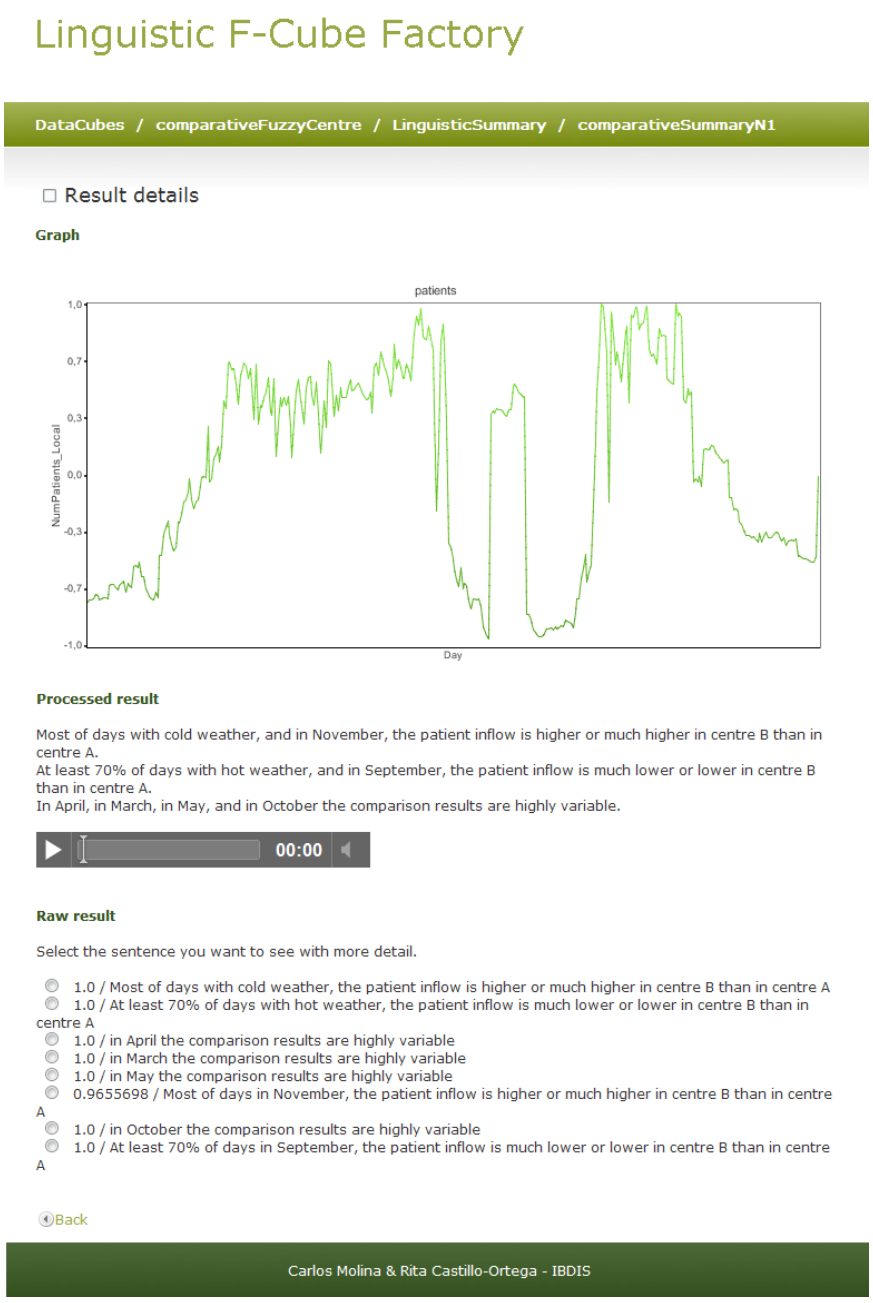


Figura 6.30: Detalles del resumen lingüístico de comparación seleccionado (1).

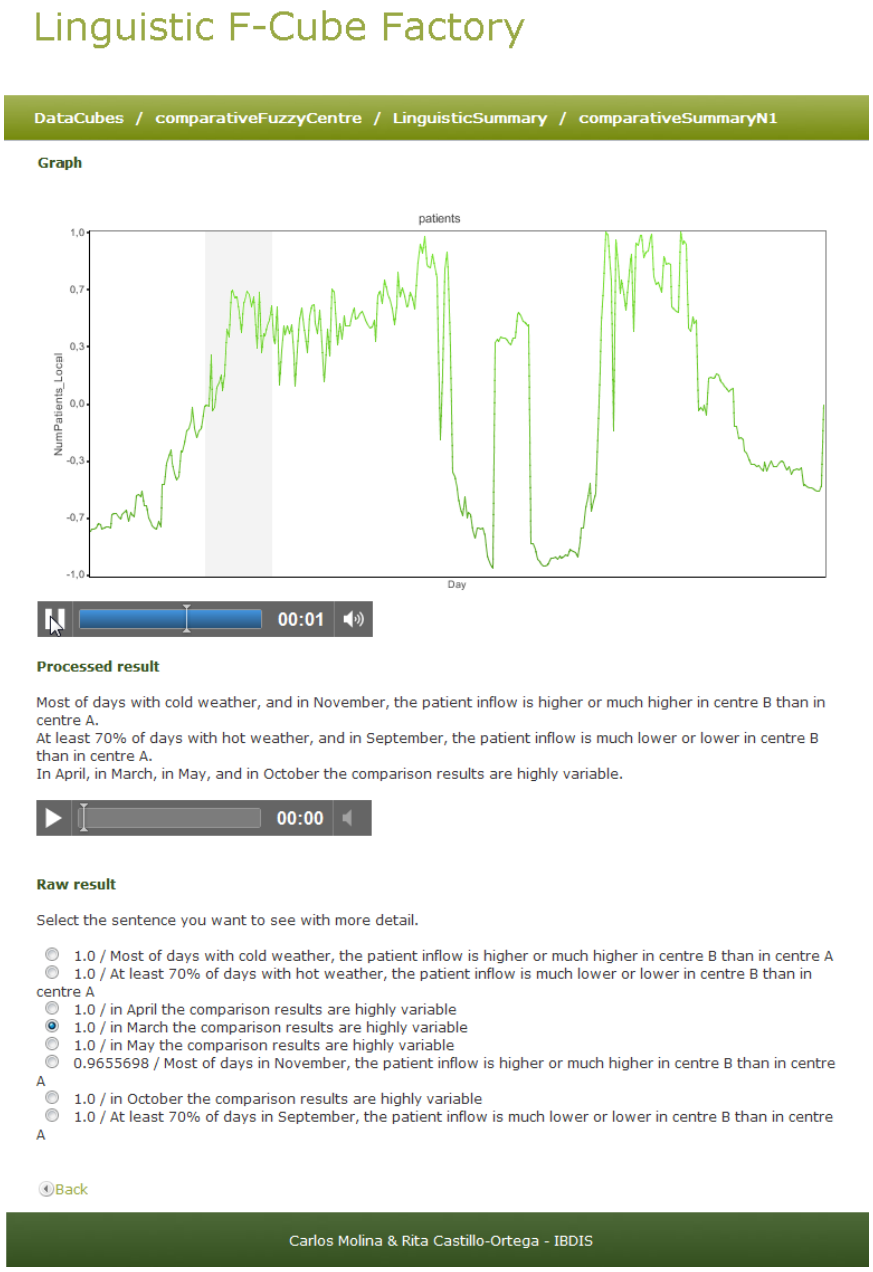


Figura 6.31: Detalles del resumen lingüístico de comparación seleccionado (2).

6.6. Conclusiones

En este capítulo se ha presentado la herramienta de gestión de cubos de datos multi-dimensionales F-Cube Factory y se ha mostrado cómo se han incorporado nuevas funcionalidades integrando en él nuestro modelo de resumen de series de datos. Fruto de ello se ha obtenido el software *Linguistic F-Cube Factory*.

Linguistic F-Cube Factory es una plataforma web basada en un diseño cliente-servidor que facilita al usuario las tareas de creación y gestión de cubos de datos de forma transparente. La arquitectura usada favorece que el usuario no necesite instalar software en su equipo, que como consecuencia no debe presentar unos requerimientos especiales aparte de contar con un navegador web y conexión a la red de trabajo donde se encuentre el servidor.

Además de las operaciones OLAP clásicas que se encontraban ya implementadas, el usuario puede obtener ahora resúmenes lingüísticos tanto de series de datos temporales como de la comparación entre las mismas.

La herramienta web presenta al usuario la opción de obtener información más detallada sobre los resúmenes del cubo de datos obtenido. En concreto puede acceder a una representación gráfica de la serie, las sentencias cuantificadas que componen el resumen y un texto procesado. Además, se ofrece la posibilidad de reproducir los diferentes textos y contrastarlos de manera interactiva con una representación gráfica de los datos originales.

La posibilidad de contar con una herramienta visual y auditiva, sencilla de manejar que nos presente los resultados deseados siguiendo patrones cercanos al lenguaje natural mejora las capacidades de procesamiento analítico de datos de las herramientas convencionales de propósito general que se pueden encontrar para el soporte a la toma de decisiones.

7 Conclusiones y trabajo futuro

7.1. Español

7.1.1. Conclusiones

A lo largo de esta memoria se ha presentado un modelo general y configurable que permite la obtención automática de resúmenes altamente intuitivos, personalizables y de calidad, de series de datos temporales.

Con la intención de cumplir el primero de los objetivos que se comentaron en la introducción del documento, y en el cual se proponía un estudio completo del concepto de resumen y su posible aplicación al ámbito de las series de datos y el proceso de toma de decisiones, se ha llevado a cabo un estudio preliminar presentado en el Capítulo 2.

Durante el estudio se han tratado disciplinas como el resumen o el análisis de series de datos en sentido amplio y datos temporales de forma concreta. De este primer estudio general se pasó a un estudio más profundo sobre las diferentes técnicas existentes para llevarlas a cabo.

Fruto de la idea inicial y después del estudio preliminar, hemos constatado que la construcción de resúmenes lingüísticos de series de datos mediante el uso de computadores se ha consolidado como un campo de investigación interesante y prometedor. El hecho de que a partir de un gran conjunto de datos numéricos podamos obtener un resumen lingüístico que los describa es muy útil. Y lo es, en especial, en el ámbito empresarial.

Para cubrir con éxito el objetivo 2, en la memoria se han presentado los modelos de resumen propuestos en la literatura, poniendo especial atención en aquellos que usan técnicas de Soft Computing, y se ha realizado un estudio de las necesidades que se observan en este campo y que justifican el esfuerzo de investigación realizado en este trabajo (Capítulo 2).

Con el propósito de cubrir las *lagunas* identificadas, ya dentro del tercer objetivo que se fijaba en la introducción, hemos desarrollado un modelo general para el resumen lingüístico de datos que, debido a la importancia de la dimensión tiempo en el análisis de datos para el apoyo a la toma de decisiones, se ha aplicado a la descripción de series temporales aunque, como se ha visto en la memoria, también puede ser aplicado en otros campos.

Nuestro modelo de resumen hace uso de etiquetas lingüísticas durante la definición del contexto para conseguir la transformación de los datos en texto. Mientras que en la dimensión que representa la variable bajo estudio se utiliza una partición del dominio, en la dimensión temporal se introduce el uso de una jerarquía de particiones. El modelo

saca provecho de dicha jerarquía de conceptos para conseguir articular resúmenes con distintos niveles de abstracción.

La salida que ofrece nuestro modelo es un resumen lingüístico compuesto por una serie de sentencias cuantificadas lingüísticas de la forma *"Q de los D son A"* donde Q es un cuantificador de entre un subconjunto de una familia coherente, D es un periodo que describe el tiempo en la jerarquía y A es la descripción de la variable en ese periodo. Para dar cabida a los intereses del usuario tanto la definición de la familia de cuantificadores como las diferentes particiones de etiquetas se hacen de acuerdo a las necesidades concretas tanto del mismo usuario como del problema. Del mismo modo también se permite la inicialización de una serie de parámetros que el usuario puede cambiar en función de sus necesidades.

Dado un conjunto de datos existe una amplia variedad de resúmenes diferentes que lo describen. Del mismo modo, por efecto de los distintos aspectos de la calidad de un resumen, no existe el mejor resumen de todos sino un resumen que es adecuado para un usuario y una situación determinadas.

En este sentido, al igual que ocurría con respecto a la tarea de resumen de datos automatizada, también existen muchos modelos para la evaluación de la calidad del resumen. En la memoria se han presentado algunos de los más representativos.

Para avanzar en esta dirección, dentro del cuarto objetivo que nos planteamos en la introducción, esta tesis incluye como aportación un modelo de calidad que permite cuantificar la calidad de un resumen y, por tanto, diseñar algoritmos que automaticen su construcción.

Concretamente, el modelo general de calidad que nosotros proponemos aquí, está compuesto por cuatro dimensiones que se desean optimizar: *cobertura, brevedad, especificidad y exactitud*. Como decimos, los diferentes aspectos se definen de forma general para que luego se puedan adaptar a cada usuario. El modelo de calidad propuesto nos ofrece la oportunidad, no sólo de evaluar la calidad de una solución individual, sino de establecer comparaciones entre un conjunto de soluciones dadas.

Basándonos en estos modelos, hemos establecido un símil entre la búsqueda del mejor resumen de un conjunto de datos y un problema de optimización, y más concretamente de optimización multi-objetivo, lo que nos ha permitido afrontar el desarrollo del objetivo 5 de esta tesis. En estos casos no se dispone de una sola medida que se quiera optimizar sino de un conjunto de medidas de que son todas ellas igual de importantes.

Para superar el citado objetivo 5, hemos presentado diferentes implementaciones de nuestra propuesta. En primer lugar encontramos el más sencillo de todos que consiste en una búsqueda exhaustiva en el espacio de soluciones. Mediante un estudio

del tamaño del espacio de búsqueda llegamos a la conclusión de que el uso de dicho enfoque sólo es posible en un número muy reducido de problemas sencillos.

La solución en este de tipo de problemas pasa por seleccionar algunas técnicas heurísticas que sean adecuadas en relación con el modelo de calidad planteado.

Siguiendo esta idea se han implementado dos estrategias Greedy que, aunque parecidas, cuentan con diferentes matices semánticos. Los algoritmos Greedy, como es conocido, si bien es posible que no encuentren la mejor solución posible, sí que nos ofrecen una solución suficientemente buena en un periodo de tiempo bastante corto.

Las estrategias Greedy, por naturaleza, incorporan durante el diseño el "carácter" del diseñador. En nuestro caso, los algoritmos Greedy llevan en su código de manera inherente una particularización del modelo de calidad que se ha presentado en la memoria y que minimiza la brevedad sacando el máximo partido de la jerarquía de conceptos definida en el marco lingüístico.

Nuestra experimentación muestra que la estrategia Greedy, en sus dos variantes, ofrece buenas soluciones en un periodo razonable de tiempo. Con el fin de comprobar la bondad de dichas soluciones y en busca de técnicas algorítmicas más configurables y capaces de construir soluciones más diversas, hemos seleccionado una segunda técnica de búsqueda, los algoritmos genéticos multi-objetivo, para implementar nuestro modelo.

En este algoritmo se conjuga de una forma equilibrada la explotación de espacios con buenas soluciones y la exploración del espacio de búsqueda. En nuestro caso hemos desarrollado un algoritmo evolutivo multi-objetivo tipo NSGA-II tomando como referencia nuestro modelo de calidad al definir los objetivos de calidad y diseñar las primitivas evolutivas.

El enfoque evolutivo propuesto cumple con nuestro objetivo de desarrollar un algoritmo más versátil tanto desde el punto de vista de su configuración y adaptación, como desde el punto de vista de la variedad de las soluciones encontradas. Esto lo convierte en una herramienta de interés tanto para el análisis del problema de elaboración de resúmenes en un dominio concreto, como para la generación de resúmenes en aquellas situaciones donde las necesidades de tiempo de respuesta y consumo de recursos lo permiten.

En cualquier caso, en nuestros experimentos, los algoritmos Greedy dan soluciones comparables en calidad a las que se encuentran en el primer frente de Pareto de las poblaciones obtenidas mediante el NSGA-II modificado. Y además, lo hacen en un tiempo que permite su uso en herramientas interactivas de consulta como la plataforma de análisis OLAP lingüístico desarrollada en esta tesis.

Al ser el nuestro un modelo general tanto en la tarea de resumen como en el

modelo de calidad, es posible hacer las adaptaciones pertinentes para permitir su uso en diferentes campos de aplicación. Hemos demostrado este potencial con diversas particularizaciones en el Capítulo 5.

En primer lugar, hemos visto que, con una adaptación adecuada del modelo propuesto, es posible realizar resúmenes de series de datos temporales pero a través del uso de otra característica como es la tendencia o variación del valor en intervalos de tiempo. Nuestro enfoque se basa en la obtención de una nueva serie de datos y el uso de una nueva partición de etiquetas lingüísticas que nos permitan describir las tendencias de la serie.

En esta línea, también hemos instanciado nuestra propuesta con la idea de poder realizar resúmenes que describan la comparación entre dos series. Dicha comparación se calcula a través de la diferencia entre los valores de las series y se puede realizar con distintas adaptaciones del modelo utilizando marcos lingüísticos tanto de carácter absoluto como relativo en relación con los datos que se desean resumir. De igual forma, se permite también la comparación en términos de diferencia en signo y variación de las series consideradas.

Asimismo, y con motivo de mostrar la portabilidad de nuestro modelo para su aplicación en la descripción de otros conjuntos de datos, se ha presentado una propuesta que nos permite la aplicación de nuestro modelo para la descripción de imágenes almacenadas de forma digital. En el capítulo 5 mostramos como la utilización de técnicas jerárquicas de segmentación de la imagen junto con metodologías para su descripción lingüística difusa, permiten aplicar nuestro modelo para resumir lingüísticamente datos relativos a determinadas características de la imagen, como es el caso del color.

Finalmente, con la idea de llevar nuestro modelo a un dominio de aplicación concreto, y que además marcábamos como de interés estratégico en la motivación de nuestro trabajo de investigación, nos hemos trasladado al dominio de los sistemas de información en el ámbito del soporte a la toma de decisiones. De esta forma además hemos cumplido el objetivo 6 propuesto.

De este modo, nuestro modelo se ha incorporado a la plataforma F-Cube Factory para la creación y gestión de datos siguiendo un modelo multi-dimensional, dando lugar a una nueva versión mejorada con capacidades de resumen lingüístico: *Linguistic F-Cube Factory*.

Gracias a su arquitectura cliente-servidor y la sencillez de su interfaz, Linguistic F-Cube Factory permite a usuarios no expertos la gestión de cubos de datos con diversas dimensiones y capacidades lingüísticas. Una vez implementado nuestro modelo en la herramienta existente, además de una modernización de la interfaz que la hace más atractiva, se ha conseguido incorporar una nueva funcionalidad que permite la

obtención de nuevos cubos de datos cuyos hechos son resúmenes lingüísticos de series temporales sencillas o resúmenes lingüísticos que describen la comparación de series temporales.

7.1.2. Trabajo futuro

El proceso de investigación llevado a cabo en esta tesis doctoral nos ha permitido no sólo generar resultados originales para cumplir con los objetivos planteados, sino asimismo identificar diversas líneas de trabajo futuro relacionado con el trabajo desarrollado. A continuación presentaremos las líneas que en nuestra opinión se presentan como más prometedoras.

En primer lugar es interesante y necesario plantearse la extensión del marco lingüístico que hemos considerado en este trabajo, en diversos aspectos:

- Estudiar el uso de cuantificadores generalizados difusos, para cuya valoración se han propuesto diversos modelos en la última década (véase por ejemplo [39, 40, 52]).

- La descripción de un número mayor de características de la serie más allá del valor y la tendencia. En la literatura pueden encontrarse otras características que han sido empleadas para describir series de datos temporales (consultar [75]).

- Incorporar protoformas más complejas que combinen en el mismo resumen información de varias características a la vez, bien describiendo todas ellas, o bien determinando qué característica es más interesante para describir cada uno de los periodos temporales en función del interés del usuario.

- Considerar patrones utilizados en la minería de datos, como reglas de asociación [42], dependencias aproximadas [146], dependencias graduales [107], excepciones y anomalías [102], etc.

Una segunda línea interesante es profundizar en el estudio de técnicas algorítmicas de búsqueda y optimización. En este aspecto hay una gran cantidad de trabajo a realizar en distintos aspectos:

- Estudiar técnicas distintas de los enfoques Greedy y basado en Algoritmos Genéticos, en particular para disponer de técnicas que prioricen los objetivos de calidad de forma distinta a las estrategias Greedy implementadas (cuyo principal objetivo es el de brevedad) y al mismo tiempo nos permitan obtener resúmenes con rapidez suficiente para permitir la realización de consultas interactivas.

- Estudiar técnicas para realizar resúmenes en base a marcos lingüísticos extendidos, en la línea del primero de los trabajos futuros que hemos indicado anteriormente.
- Estudiar técnicas para seleccionar el conjunto de parámetros de los algoritmos más adecuado para enfrentarnos a cada problema. Debemos recordar que la optimización de este conjunto de parámetros para los algoritmos desarrollados en esta tesis no ha sido un objetivo de la misma, por lo que se han realizado pruebas con un conjunto de diferentes combinaciones de parámetros. En esta línea de trabajo contamos con el interés y la colaboración del profesor Andrea G. B. Tettamanzi, con el que se ha iniciado una estrecha colaboración en este campo que esperamos sea muy fructífera.

La tercera línea de trabajo futuro que queremos destacar se centra en profundizar en el estudio del modelo de calidad que se ha presentado, en los siguientes puntos fundamentales:

- Incorporar al modelo otros aspectos de la calidad del resumen cuya valoración es intrínsecamente subjetiva, tales como la relevancia o interés, la utilidad, etc. Muy recientemente, diversos investigadores han propuesto modelos de valoración de la calidad del resumen que consideran una gran cantidad de aspectos de este tipo, incluyendo características definidas desde el ámbito del NLG tales como las distintas dimensiones consideradas en el paradigma de Gramática Sistémica Funcional [120]. Asimismo, en el ámbito del Soft Computing existen herramientas especialmente adecuadas para definir algunos de estos criterios, tales como las relaciones de preferencia difusa.
- Determinar criterios para definir, para un contexto/usuario concreto, medidas adecuadas de las dimensiones que consideramos objetivas dentro del modelo de calidad.
- Determinar distintos mecanismos de priorización y combinación de los distintos aspectos de la calidad para el diseño de algoritmos eficientes, según las preferencias del usuario. Aunque resulta ideal desde un punto de vista teórico, el modelo puramente multi-objetivo que considera frentes de Pareto en un espacio con tantas dimensiones como aspectos de la calidad considerados plantea unas exigencias computacionales que restringen su uso potencial a procesamiento no interactivo.

En cuarto lugar nos planteamos profundizar en el estudio de la generalización de las técnicas propuestas para el resumen lingüístico de distintos tipos de datos. Como ya se indicó en el capítulo 5, hay una gran cantidad de trabajo por realizar en este punto. Entre otros aspectos podemos destacar los siguientes:

- Ampliar la comparación de series para que nos permita describir la comparación de un conjunto de series con más de dos series. De este modo podríamos obtener sentencias como por ejemplo *"en la mayoría de los años la afluencia al centro de salud durante el mes de Enero es alta"*. Esta capacidad es especialmente interesante en el ámbito de la consulta en bases de datos multi-dimensionales.

- Considerar el resumen de una variable tomando como partición la proporcionada por los valores de otra variable, o incluso una partición jerárquica proporcionada por valores de un conjunto de variables ordenadas (particionar por la primera variable, subdividir cada grupo por el valor de la segunda variable, y así sucesivamente).

- Generar resúmenes de conjuntos de datos complejos y/o poco estructurados. En particular, seguiremos trabajando en el ámbito de la descripción lingüística de imágenes en base a segmentaciones difusas jerárquicas y conceptos semánticos relativos a color, textura, forma, localización, y relaciones espaciales.

En quinto lugar podemos destacar una serie de líneas de trabajo relacionadas con el desarrollo y la explotación de aplicaciones software basadas en las técnicas de resumen que proporciona nuestro trabajo. De manera general podemos hablar de:

- Incorporar los desarrollos futuros en las líneas anteriores en la plataforma Linguistic F-Cube Factory, proporcionando nuevas posibilidades de consulta y análisis de la información a los usuarios.

- Colaborar con otros grupos de investigación para la resolución de problemas reales haciendo uso de nuestra propuesta. En la literatura se han descrito aplicaciones de técnicas de resumen de series de datos temporales en distintos ámbitos, así como aplicaciones en otros tipos de datos donde es posible aplicar versiones generalizadas de nuestras técnicas. Cabe destacar que hemos establecido contactos con distintos grupos de trabajo interesados en la aplicación de nuestras técnicas a conjuntos de datos reales para la resolución de problemas en distintos ámbitos, como por ejemplo describir la actividad física para evaluar y mejorar la calidad de vida de las personas, tanto en aplicaciones de mejora del estado físico como para aumentar la autonomía de personas mayores que viven solas mediante lo que se conoce como "independencia controlada".

Como se puede apreciar son muchas y muy variadas las líneas futuras de trabajo que se han planteado como consecuencia de nuestra investigación. Como se ve, esta lista de trabajos futuros cumple con un objetivo deseado en toda tesis de sentar las bases para un desarrollo posterior a la misma en el seno del grupo de investigación. Es nuestro deseo contar con el tiempo y los recursos para poder llevar a cabo todas y cada una de ellas.

7.2. English

7.2.1. Conclusions

Along this document, a general and configurable model enabling the automatic creation of understandable, customizable, and high quality summaries from temporal data series, has been presented.

In order to meet the first of the objectives which were mentioned in the introduction of this document, regarding a complete study of the concept of summary and its possible application to the scope of the data series and in the decision making process, a preliminary study has been implemented.

During the study, disciplines such as the summary of the analysis of data series in a broad sense and of temporal data in a specific way have been treated. From this first general study, we passed to a deeper study on the different existing techniques to realize them.

Stemming from the initial idea and after the preliminary study, we have verified that the creation of linguistic summaries of data series by means of the use of computers has consolidated as an interesting and promising research field. The fact that from a large set of numeric data we can obtain a linguistic summary describing them is really useful, especially, in the business scope.

In order to successfully cover objective 2, the summarization models proposed in the literature have been presented in this document, focusing on those which use Soft Computing techniques. Besides, the needs which can be observed in this field and which justify the research effort carried out with this work have been studied.

So as to cover the identified gaps, and already within the third objective established in the introduction, we have developed a general model for the linguistic summarization of data. Due to the relevance of the time dimension in the data analysis for supporting the decision making process, the model has been applied to the description of temporal series, although, as has been described in the report, can also be applied in other fields.

Our summarization model uses linguistic labels during the definition of the context so as to achieve the transformation of data into text. Whereas a partition of the domain is used in the dimension representing the variable which is being studied, the use of a hierarchy of partitions is introduced in the temporal dimension. The model takes advantage from that concept hierarchy to generate summaries with different levels of abstraction.

The output offered by our model is a linguistic summary consisting of a series of linguistic quantified sentences of the type *"Q of D are A"* , where Q stands for

a quantifier from a subset of a coherent family, D is a period describing time in the hierarchy, and A is the description of the variable in that period. To meet the interests of the user, both the definition of the family of quantifiers and the different label partitions are performed according to the specific needs both of the user. In the same way, a series of parameters can be specified by the user depending on his/her specific needs.

For a dataset, there is a large variety of different summaries which describe it. In the same way, due to the different aspects of the quality of a summary, there is not a "best summary", but a summary which meets the needs of a specific user at a specific situation.

In this sense, as happened with the automated data summarization task, there are also many models for the assessment of the quality of the summary. We have presented some of the most representative ones.

To progress in this direction, within the fourth objective stated in the introduction, this thesis includes as a contribution a quality model which allows the quantification of the quality of a summary, enabling as a result the design of algorithms to automate their creation.

Specifically, the general quality model proposed here consists of four dimensions to be optimized: *coverage, brevity, specificity, and accuracy.* We have contributed a general definition of this aspect and the user can accordingly formulates a suitable measures in order to meet his or her requirements. The proposed quality model offers the possibility not only of assessing the quality of an individual solution, but also of establishing comparisons among a set of given solutions.

Based on these models, we have established a comparison between the search for the best summary of a dataset and an optimization problem, more specifically, of a multi-objective optimization, which has allowed us to face the development of the 5^{th} objective of this doctoral dissertation. In these cases, there is not only a single measure to be optimized, but a set of equally important measures.

To overcome the mentioned 5^{th} objective, several implementations of our proposal have been presented. Firstly, we consider a exhaustive search in the solution space. This way, by studying the size of the search space, we reach the conclusion that the use of that approach is only possible in a very limited number of naive problems.

The solution in this type of problems lies in selecting some heuristic search techniques which are appropriate for the used quality model.

Following this idea, two Greedy strategies, which though similar, have different semantic nuances, were implemented. Greedy algorithms, as is already known, do offer us a solution which is good enough, though they may not find the best possible one,

in a short period of time.

Greedy strategies feature, by nature, the "character" of the designer. In our case, the Greedy algorithms carry inherently a particularization of our quality model that minimizes brevity taking advantage of the concept hierarchy defined in the linguistic framework.

Our experimentation shows that the Greedy strategy, in its two modalities, offers good solutions in a reasonable period of time. In order to check it and to achieve more configurable algorithmic techniques able to create more diverse solutions, we have selected a second search technique, the multi-objective genetic algorithms, to implement our model.

In this algorithm the exploitation of spaces is combined with good solutions and with the exploration of the search space in a balanced way. In our case, we have developed a multi-objective evolutionary algorithm of the NSGA-II type, taking as reference our quality models when defining the objectives and designing our evolutionary primitives.

The proposed evolutionary approach meets our objective of developing a more versatile algorithm from the point of view of its configuration and adaptation, as well as from the point of view of the variety of found solutions. This turns it into an interesting tool both for the analysis of the summary generation problem in a specific domain as for the creation of summaries in those situations where the response time and resource consumption needs allow it.

In any case, in our experiments, Greedy algorithms render similar solutions as regards quality to the ones on the first Pareto front of the populations obtained by means of the modified NSGA-II. And besides, they do it in a time which allows its use in interactive query tools as the OLAP linguistic platform developed in this thesis.

Due to the fact that our model is a general one both as regards the summarization task and the quality model, it is possible to implement the necessary adaptations to allow its use in different application fields. We have shown this with different particularizations in Chapter 5.

First, we have seen it is possible to create summaries of time series data, with an appropriate adaptation of the proposed model, but through the use of another characteristic as the trend or variation of value between time instant pairs. Our approach is based on obtaining a new data series and on the use of a new partition of linguistic labels allowing us to describe the trends of the series.

In this line, we have also instantiated our proposal so as to be able to create summaries describing the comparison between two series. Such comparison is calculated through the difference between the values of the series and can be performed

with different adaptations of the model, using both absolute and relative linguistic frameworks, as well as difference of sign and variation of the considered series.

In the same way, and in order to show the portability of our model for its application in the description of other datasets, a proposal which allows us to apply our model for the description of digitally stored images has been presented. The use of hierarchical image segmentation techniques, together with methodologies for their fuzzy linguistic description, allow the application of our model to linguistically summarize data regarding certain characteristics of the image, as for example, colour.

Finally, with the idea of taking our model to a specific application domain and which we also marked as having a strategic interest in the motivation of our research work, we have moved to the domain of the information systems in the scope of the decision making process support. In this way, we also meet objective 6 from those proposed in the introduction.

Thus, our model has been added to the F-Cube Factory platform for the creation and management of data following a multi-dimensional model, leading to a new improved version with linguistic summarization capabilities: *Linguistic F-Cube Factory.*

Thanks to its client-server architecture and to the simplicity of its interface, Linguistic F-Cube Factory allows non expert users to manage data cubes of different dimensions and linguistic capabilities. Once our model is implemented in the existing tool, besides a modernization of the interface making it more attractive, we have managed to include a new functionality which allows us to obtain data cubes whose facts are linguistic summaries of simple temporal series or linguistic summaries describing the comparison of temporal series.

7.2.2. Future work

The research process carried out in this doctoral dissertation has allowed us not only to generate original results to meet the planned objectives, but also, to identify several future lines of research related to the developed work. Below, some of the lines which seem, in our opinion, more promising are presented.

In the first place, it is interesting and necessary to analyse the extension of the linguistic framework which has been considered in this work in certain aspects:

- Studying the use of general fuzzy quantifiers, for which assessment, different models have been proposed in the literature (see for example [39, 40, 52]).

- The description of a larger number of characteristics of the series beyond value and trend. In the literature, other characteristics which have been used to describe temporal data series can be found (see [75]).

- Adding more complex protoforms combining in the same summary information of several characteristics at the same time, either describing them all or, establishing which characteristic is the most interesting one to describe each of the temporal periods depending on the interest of the user.

- Considering patterns used in data mining, as association rules [42], approximate dependencies [146], gradual dependencies [107], exceptions and anomalies [102], etc.

A second interesting line is to study in more depth the algorithmic search and optimization techniques. There is much work to be done in several aspects:

- Studying different algorithmic techniques which prioritize the quality objectives in a different way to the implemented Greedy strategies (mainly aiming at brevity) and which allow us at the same time to obtain summaries quick enough so as to allow the implementation of interactive queries.

- Studying techniques to create summaries based on extended linguistic frameworks, following the first of the future works which has been previously mentioned.

- Studying techniques to select the most appropriate set of parameters to face each problem. We must bear in mind that the optimization of this set of parameters for the algorithms which have been developed in this thesis has not been its objective. Due to this, some tests with a set of different combinations of parameters have been carried out. In this line of work, we count on the interest and collaboration of Professor Andrea G. B. Tettamanzi, with whom a tight collaboration, which we expect to be highly productive, has been formed in this field.

The third line of future work which we would like to highlight is focused in analysing the quality model which has been presented in more depth, in particular, the following topics:

- Adding to the model other aspects of the quality of the summary which have an intrinsically subjective assessment, such as its relevance or interest, usefulness, etc. Very recently, several researchers have proposed assessment models for the quality of the summary which take into consideration a large amount of aspects of this type, including characteristics defined from the NLG scope, such as the different dimensions considered for the *Functional Systemic Grammar* paradigm [120]. In the same way, within the Soft Computing area, there are particularly

appropriate tools to define some of these criteria, such as the fuzzy preference relationships.

- Determining criteria to define, for a specific context/user, appropriate measures of the dimensions which we consider within the quality model.

- Determining different mechanisms for the prioritization and combination of the different aspects of quality for the design of efficient algorithms, depending on the preferences of the user. Although ideal from a theoretical point of view, the purely multi-objective model which considers Pareto fronts in a space with as many dimensions as considered quality aspects bears some computational requirements which restrict its potential use to non interactive processing.

In the fourth place, we are thinking about studying the generalization of the proposed techniques for the linguistic summarization of different types of data in more depth. As pointed out in Chapter 5, there is a large amount of work to be done along this line. Among other points, the following can be highlighted:

- Enhancing the comparison of series so as to be able to describe the comparison of a set of series with more than two of them. In this way, we could obtain sentences such as *"in most of the years, the number of people going to the healthcare centre during January is high"*. This capability is particularly interesting in the scope of queries in multi-dimensional databases.

- Considering the summary of a variable taking as partition the one provided by the values of another variable, or even a hierarchical partition provided by values of a set or ordered variables (partitioning by the first variable, subdividing each group by the value of the second variable, and so on).

- Generating summaries of complex and/or almost unstructured data sets. In particular, we will continue our work in the area of the linguistic description of images based on hierarchical fuzzy segmentations and semantic concepts regarding colour, texture, shape, location, and spatial relationships.

In the fifth place, we can highlight a series of lines of work related to the development and operation of software applications based on the summarization techniques provided by our work. In general, we can point out:

- Adding the future developments in the previously described lines in the Linguistic F-Cube Factory platform, providing the users with new possibilities for the query and information analysis.

- Collaborating with other research teams in the solution of real problems using our proposal. In the literature, applications of summarization techniques of temporal data series in several areas have been described, as well as applications in other types of data where it is possible to apply generalized versions of our techniques. It is worth highlighting that we have established contact with different work groups who are interested in the application of our techniques to real datasets for the resolution of problems in different areas, as for example, to describe physical activity to assess and improve the quality of life of people, both in applications for the improvement of the physical condition and for increasing the autonomy of the elderly living on their own by means of what is known as "controlled independence".

As can be seen, there are many different future lines of work which have been raised from our research. It can be appreciated that the list of future tasks meets the desired aim of every thesis. It allows us to establish the bases of a subsequent development within our investigation group. We would like to have the necessary time and resources to tackle all of them

Referencias

[1] Hernán E. Aguirre and Kiyoshi Tanaka. Adaptive -ranking on many-objective problems. *Evolutionary Intelligence*, 2(4):183–206, 2009.

[2] D. Anderson, R. H. Luke III, J. M. Keller, M. Skubic, M. Rantz, and M. Aud. Linguistic summarization of video for fall detection using voxel person and fuzzy logic. *Computer Vision and Image Understanding*, 113(1):80–89, 2009.

[3] D. T. Anderson, R. H. Luke III, and J. M. Keller. Segmentation and linguistic summarization of voxel environments using stereo vision and genetic algorithms. In *FUZZ-IEEE 2010, IEEE International Conference on Fuzzy Systems, Barcelona, Spain, 18-23 July, 2010, Proceedings*, pages 1–8, 2010.

[4] D. T. Anderson, J. M. Keller, M. Anderson, and D. J. Wescott. Linguistic description of adult skeletal age-at-death estimations from fuzzy integral acquired fuzzy sets. In *FUZZ-IEEE 2011, IEEE International Conference on Fuzzy Systems, Taipei, Taiwan, 27-30 June, 2011, Proceedings*, pages 2274–2281, 2011.

[5] J. Barwise and R. Cooper. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4:159–219, 1981.

[6] I. Z. Batyrshin. On linguistic representation of quantitative dependencies. *Expert Syst. Appl.*, 26(1):95–104, 2004.

[7] I. Z. Batyrshin and L. Sheremetov. Perception-based approach to time series data mining. *Appl. Soft Comput.*, 8(3):1211–1221, 2008.

[8] I. Z. Batyrshin and T. Sudkamp. Perception based data mining and decision support systems. *Int. J. Approx. Reasoning*, 48(1):1–3, 2008.

[9] I. Z. Batyrshin and M. Wagenknecht. Towards a linguistic description of dependencies in data. *International Journal Appl. Math. Comput. Sci.*, 12(3):391–401, 2002.

[10] B. Berlin and P. Kay. *Basic color terms: their Universality and Evolution.* Berkeley: University of California Press, 1969.

[11] P. Bosc, D. Dubois, O. Pivert, H. Prade, and M. De Calmes. Fuzzy summarization of data using fuzzy cardinalities. In *Int. Conf. Inf. Process. Manag. Uncertainty Knowl. Based Syst.*, pages 1553–1559, 2002.

[12] P. Bosc, A. HadjAli, H. Jaudoin, and O. Pivert. Flexible querying of multiple data sources through fuzzy summaries. In *DEXA Workshops*, pages 350–354, 2007.

[13] D. R. Brillinger. *Time Series. Data Analysis and Theory.* Siam, Society for Industrial and Applied Mathematics, 2001.

[14] P. J. Brockwell. *Introduction to time series and forecasting.* Springer-Verlag, 1996.

[15] P. Brophy and K. Coulling. *Quality management for information and library managers.* Gower, London, 1996.

[16] R. A. Carrasco and P. Villar. A new model for linguistic summarization of heterogeneous data: an application to tourism web data sources. *Soft Comput.*, 16(1):135–151, January 2012.

[17] R. Castillo-Ortega, N. Marín, D. Sánchez, and A.G.B. Tettamanzi. Linguistic summarization of time series data using genetic algorithms. In *EUSFLAT-LFA 2011 European Society for Fuzzy Logic and Technology*, pages 416–423, 2011.

[18] R. Castillo-Ortega, N. Marín, D. Sánchez, and A.G.B. Tettamanzi. A multi-objective memetic algorithm for the linguistic summarization of time series. In *GECCO, Genetic and Evolutionary Computation Conference 2011*, pages 171–172, 2011.

[19] J. Chamorro-Martínez, D. Sánchez, B. Prados-Suárez, E. Galán-Perales, and M.A. Vila. Segmenting colour images on the basis of a fuzzy hierarchical approach. *Mathware & Soft Computing*, 10:101–115, 2003.

[20] C. Chatfield. *The analysis of time series: an introduction.* Chapman and Hall, 2004.

[21] Chun-Hao Chen, Tzung-Pei Hong, and Vincent S. Tseng. Fuzzy data mining for time-series data. *Appl. Soft Comput.*, 12(1):536–542, January 2012.

[22] G. Chen, Q. Wei, and E. E. Kerre. *Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques, Massive Computing Series*, chapter 14 Fuzzy Logic in Discovering Association Rules: An Overview, pages 459–493. Massive Computing Series. Springer, Heidelberg, Germany, 2006.

[23] H. D. Cheng and J. Li. Fuzzy homogeneity and scale-space approach to color image segmentation. *Pattern Recognition*, 36(7):1545–1562, 2003.

[24] D. Chiang, L. R. Chow, and Y. Wang. Mining time series data by a fuzzy linguistic summary system. *Fuzzy Sets Syst.*, 112:419–432, June 2000.

[25] C. A. Coello Coello, G. B. Lamont, and D. A. Van Veldhuizen. *Evolutionary Algorithms for Solving Multi-Objective Problems (Genetic and Evolutionary Computation).* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[26] A.G. Cohn and S.M. Hazarika. Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 46(1–2):1–29, 2001.

[27] J. D. Cryer and Kung-Sik Chan. *Time Series Analysis with applications in R*. Springer, 2008.

[28] C. Darwin. *On the Origin of Species by Means of Natural Selection*. John Murray, 1859.

[29] G. Das, K.I. Lin, H. Mannila, G. Renganathan, and P.Smyth. Rule discovery from time series. In *4th Conf. on Knowledge Discovery an Data Mining*, pages 16–22, 1998.

[30] Instituto Nacional de Estadística. http://www.ine.es/.

[31] POOLRed Sistema de Información de precios en origen de mercado de contado del aceite de oliva. http://www.oliva.net/poolred/.

[32] Bolsa de Madrid. http://www.bolsademadrid.es/esp/portada.htm.

[33] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii. pages 849–858. Springer, 2000.

[34] Kees Van Deemter. Utility and language generation: the case of vagueness, 2009.

[35] Kenneth A. DeJong. *Evolutionary Computation: A unified approach*. MIT Press, Cambridge, MA, 2002.

[36] M. Delgado, C. Molina, L. Rodríguez Ariza, D. Sánchez, and M. A. Vila Miranda. F-cube factory: a fuzzy olap system for supporting imprecision. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 15(Supplement-1):59–81, 2007.

[37] M. Delgado, M.D. Ruiz, D. Sánchez, and M.A. Vila. Quantified sentences and evaluation methods: a state of the art. *Sometido a International Journal of Approximate Reasoning*.

[38] M. Delgado, D. Sánchez, and M.A. Vila. Fuzzy cardinality based evaluation of quantified sentences. *International Journal of Approximate Reasoning*, 23:23–66, 2000.

[39] F. Díaz-Hermida and A. Bugarín. Linguistic summarization of data with probabilistic fuzzy quantifiers. In *ESTYLF 2010, XV Congreso Español Sobre Tecnologías y Lógica Fuzzy, Huelva, Spain, 3-5 Frebruary, Proceedings*, pages 255–260, 2010.

[40] F. Díaz-Hermida and A. Bugarín. Semi-fuzzy quantifiers as a tool for building linguistic summaries of data patterns. In *Proceedings of the IEEE Symposium on Foundations of Computational Intelligence, FOCI 2011, part of the IEEE Symposium Series on Computational Intelligence 2011, Paris, France, 11-15 April 2011*, pages 45–52, 2011.

[41] F. Díaz-Hermida, A. Ramos-Soto, and A. Bugarín. On the role of fuzzy quantified statements in linguistic summarization of data. In *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*, pages 166 –171, nov. 2011.

[42] Wu Dongrui and J. M. Mendel. Linguistic summarization using if-then rules and interval type-2 fuzzy sets. *IEEE Transactions on Fuzzy Systems*, 19(1):136–151, 2011.

[43] Agoston E. Eiben and J. E. Smith. *Introduction to Evolutionary Computing*. Springer-Verlag, Berlin, 2003.

[44] Asociación española de normalización y certificación (AENOR). *Documentación: Preparación de resúmenes*. 1990.

[45] M. Pinto Molina et al. *Aprendiendo a resumir. Prontuario y resolución de casos*. TREA, Gijón, 1 edition, 2005.

[46] Yahoo finance. http://finance.yahoo.com/.

[47] Carlos M. Fonseca, Peter J. Fleming, Eckart Zitzler, Kalyanmoy Deb, and Lothar Thiele, editors. *Evolutionary Multi-Criterion Optimization*, volume 2632 of *LNCS*. Springer-Verlag, Berlin, 2003.

[48] Tak-chung Fu. A review on time series data mining. *Eng. Appl. Artif. Intell.*, 24(1):164–181, February 2011.

[49] J. Moreno García, J. J. Castro-Schez, and L. Jiménez. A fuzzy inductive algorithm for modeling dynamical systems in a comprehensible way. *IEEE T. Fuzzy Systems*, 15(4):652–672, 2007.

[50] S. Garcia-Talegon and J. Moreno García. A linguistic fuzzy method to study electricity market agents. In *ICEIS 2005, Proceedings of the Seventh International Conference on Enterprise Information Systems, Miami, USA, May 25-28, 2005*, pages 394–399, 2005.

[51] A. Gatt, F. Portet, E. Reiter, J. Hunter, S. Mahamood, W. Moncur, and S. Sripada. From data to text in the neonatal intensive care unit: Using nlg technology for decision support and information management. *AI Commun.*, 22(3):153–186, August 2009.

[52] I. Glöckner. Evaluation of quantified propositions in generalized models of fuzzy quantification. *Int. J. Approx. Reasoning*, 37(2):93–126, 2004.

[53] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky. *Analysis of time series structure*. Chapman and Hall, 2003.

[54] J. D. Hamilton. *Time Series Analysis*. Princeton Univesity Press, 1994.

[55] R. Harris and R. Sollis. *Applied time series modelling and forecasting*. John Wiley and Sons, 2003.

[56] A. C. Harvey. *Forecasting, structural time series models and Kalman filter*. Cambridge University Press, 1989.

[57] Yll Haxhimusa and Walter Kropatsch. Hierarchical image partitioning with dual graph contraction. In *Proc. of 25th DAGM Symposium LNCS*, pages 338–345. Springer, 2003.

[58] R. Hayek, G. Raschia, P. Valduriez, and N. Mouaddib. Peersum: a summary service for p2p applications. *Int. J. Pervasive Computing and Communications*, 4(4):390–410, 2008.

[59] R. Hayek, G. Raschia, P. Valduriez, and N. Mouaddib. Summary management in unstructured p2p systems. *Ingénierie des Systèmes d'Information*, 13(5):83–106, 2008.

[60] Reyhaneh Hesami, Alireza BabHadiashar, and Reza HosseinNezhad. Range segmentation of large building exteriors: A hierarchical robust approach. *Computer Vision and Image Understanding*, 2010. In press, doi:10.1016/j.cviu.2009.12.004.

[61] F. Höppner. Discovery of temporal patterns - learning rules about the qualitative behaviour of time series. In *PKDD'01. Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, volume 2168 of *LNAI*, pages 192–203. Springer-Verlag, 2001.

[62] F. Höppner. Handling feature ambiguity in knowledge discovery from time series. In *DS'02: Proceedings of the 5th International Conference on Discovery Science*, volume 2534 of *LNCS*, pages 398–405. Springer-Verlag, 2002.

[63] F. Höppner. Learning dependencies in multivariate time series. In *Proceedings of the ECAI'02 Workshop on Knowledge Discovery in (Spatio-) Temporal Data*, pages 25–31, 2002.

[64] F. Höppner. Time series abstraction methods – a survey. In *Informatik bewegt: Informatik 2002 - 32. Jahrestagung der Gesellschaft für Informatik e.v.*, pages 777–786. GI, 2002.

[65] F. Höppner and F. Klawonn. Finding informative rules in interval sequences. In *IDA'01: Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*, volume 2189 of *LNCS*, pages 125–134. Springer-Verlag, 2001.

[66] M. Humano, M. Okamura, and K. Seta. Improved method for linguistic expression of time series with global trend and local features. In *FUZZ-IEEE 2009*, pages 1169–1174, 2009.

[67] Hypervolume indicator. http://iridia.ulb.ac.be/ manuel/hypervolume.

[68] M. Jeon, M. Alexander, W. Pedrycz, and N. Pizzi. Unsupervised hierarchical image segmentation with level set and additive operator splitting. *Pattern Recognition Letters*, 26:1461–1469, 2005.

[69] E. Uriel Jiménez. *Análisis de series temporales. Modelos ARIMA*. Paraninfo, 1985.

[70] J. Kacprzyk. Fuzzy logic for linguistic summarization of databases. In *IEEE International Fuzzy Systems Conference*, pages 813–818, 1999.

[71] J. Kacprzyk and A. Wilbik. Linguistic summarization of time series using fuzzy logic with linguistic quantifiers: A truth and specificity based approach. In Leszek Rutkowski, Ryszard Tadeusiewicz, Lotfi A. Zadeh, and Jacek M. Zurada, editors, *Artificial Intelligence and Soft Computing - ICAISC 2008, 9th International Conference, Zakopane, Poland, June 22-26, 2008, Proceedings*, pages 241–252, 2008.

[72] J. Kacprzyk and A. Wilbik. Linguistic summaries of time series using a degree of appropriateness as a measure of interestingness. In *Ninth International Conference on Intelligent Systems Design and Applications, ISDA 2009, Pisa, Italy , November 30-December 2, 2009*, pages 385–390, 2009.

[73] J. Kacprzyk and A. Wilbik. Using fuzzy linguistic summaries for the comparison of time series: an application to the analysis of investment fund quotations. In U. Kaymak J. P. Carvalho, D. Dubois and J. M. C. Sousa, editors, *IFSA-EUSFLAT 2009*, pages 1321–1326, 2009.

[74] J. Kacprzyk and A. Wilbik. A comprehensive comparison of time series described by linguistic summaries and its application to the comparison of performance of a mutual fund and its benchmark. In *FUZZ-IEEE 2010, IEEE International Conference on Fuzzy Systems, Barcelona, Spain, 18-23 July, 2010, Proceedings*, pages 1–8, 2010.

[75] J. Kacprzyk, A. Wilbik, and S. Zadrozny. Capturing the essence of dynamic behaviour of sequences of numerical data using elements of quasi-natural language. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 3365–3370, 2006.

[76] J. Kacprzyk, A. Wilbik, and S. Zadrozny. Linguistic summarization of time series using a fuzzy quantifier driven aggregation. *Fuzzy Sets and Systems*, 159(12):1485–1499, 2008.

[77] J. Kacprzyk, A. Wilbik, and S. Zadrozny. An approach to the linguistic summarization of time series using a fuzzy quantifier driven aggregation. *Int. J. Intell. Syst.*, 25(5):411–439, 2010.

[78] J. Kacprzyk and R. R. Yager. Linguistic summaries of data using fuzzy logic. In *International Journal of General Systems*, volume 30, pages 133–154, 2001.

[79] J. Kacprzyk and R. R. Yager. Linguistic summarization of data association rules sets using association rules. In *The IEEE International Conference on Fuuy Systems*, pages 702–707, 2003.

[80] J. Kacprzyk, R. R. Yager, and S. Zadrozny. A fuzzy logic based approach to linguistic summaries in databases. *International Journal of Applied Mathematical Computer Science*, 10:813–834, 2000.

[81] J. Kacprzyk and S. Zadrozny. Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools. *Inf. Sci. Inf. Comput. Sci.*, 173(4):281–304, 2005.

[82] J. Kacprzyk and S. Zadrozny. Protoforms of linguistic database summaries as a human consistent tool for using natural language in data mining. *IJSSCI*, 1(1):100–111, 2009.

[83] J. Kacprzyk and S. Zadrozny. Computing with words is an implementable paradigm: Fuzzy queries, linguistic data summaries, and natural-language generation. *IEEE T. Fuzzy Systems*, 18(3):461–472, 2010.

[84] M.W. Kadous. Learning comprehensible descriptions of multivariate time series. In *Int. Conf. on Machine Learning*, pages 454–463, 1999.

[85] K.L. Kelly and D.B. Judd. The ISCC-NBS method of designating colors and a dictionary of color names. *National Bureau of Standards (USA)*, (NBS Circular 553), 1955.

[86] K.L. Kelly and D.B. Judd. Color universal color language and dictionary of names. *National Bureau of Standards (USA)*, (440), 1976.

[87] Sir M. Kendall and J. Keith Ord. *Time Series, Third Edition.* Edward Arnold, 1990.

[88] G. Kirchgassner and J. Wolters. *Introduction to modern time series analysis.* Springer Verlag, 2008.

[89] I. Kobayashi, M.Noumi, and A. Hiyama. A study on verbalization of human behaviors in a room. In *FUZZ-IEEE*, pages 1–6, 2010.

[90] I. Kobayashi and N. Okumura. Verbal explaining of the behavior of time-series data. In *Web Intelligence/IAT Workshops*, pages 139–142, 2008.

[91] I. Kobayashi and N. Okumura. Verbalizing time-series data: With an example of stock price trends. In *IFSA/EUSFLAT Conf.*, pages 234–239, 2009.

[92] Arjan Kuijper and Luc M. J. Florack. The hierarchical structure of images. *IEEE Transactions on Image Processing*, 12:1067–1079, 2003.

[93] A. Laurent. A new approach for the generation of fuzzy summaries based on fuzzy multidimensional databases. *Intell. Data Anal.*, 7:155–177, April 2003.

[94] Doheon Lee and Myoung-Ho Kim. Database summarization using fuzzy isa hierarchies. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, pages 671–680, 1997.

[95] Stefaan Lhermitte, Jan Verbesselt, Inge Jonckheere, Kris Nackaerts, Jan A.N. van Aardt, Willem W. Verstraeten, and Pol Coppin. Hierarchical image segmentation based on similarity of NDVI time series. *Remote Sensing of Environment*, 112:506–521, 2008.

[96] Tung-Kuan Liu, Yeh-Peng Chen, and Jyh-Horng Chou. Extracting fuzzy relations in fuzzy time series model based on approximation concepts. *Expert Syst. Appl.*, 38(9):11624–11629, September 2011.

[97] Marc Liévin and Franck Luthon. Nonlinear color space and spatiotemporal mrf for hierarchical segmentation of face features in video. *IEEE Transactions on Image Processing*, 13:63–71, 2004.

[98] J. Maeda, C. Ishikawa, S. Novianto, N. Tadehara, and Y. Suzuki. Rough and accurate segmentation of natural color images using fuzzy region-growing algorithm. In *15th International Conference on Pattern Recognition*, volume 3, pages 638–641, October 2000.

[99] S. Mahamood and E. Reiter. Generating affective natural language for parents of neonatal infants. In *In proceeding of: 13th European Workshop on Natural Language Generation*, pages 12–21, 2011.

[100] Nicolas Eric Maillot and Monique Thonnat. Ontology based complex object recognition. *Image and Vision Computing*, 26:102–1131, 2008.

[101] S. Makrogiannis, G. Economou, and S. Fotopoulos. A region dissimilarity relation that combines feature-space and spatial information for color image segmentation. *IEEE Transactions on Systems, Man & Cybernetics, Part B Cybernetics*, 35 (1):44–53, 2005.

[102] Dragos Margineantu, Stephen Bay, Philip Chan, and Terran Lane. Data mining methods for anomaly detection kdd-2005 workshop report. *SIGKDD Explor. Newsl.*, 7(2):132–136, December 2005.

[103] S. Mitra, Senior Member, Fellow, S. K. Pal, and P. Mitra. Data mining in soft computing framework: A survey. *IEEE Transactions on Neural Networks*, 13:3–14, 2001.

[104] D. Mladenic, N. Lavrac, M. Bohanec, and S Moyle, editors. *Data Mining and Decision Support*, volume 745 of *The Springer International Series in Engineering and Computer Science*. 2003.

[105] A. Moghaddamzadeh and N. Bourbakis. A fuzzy region growing approach for segmentation of color images. *Pattern Recognition*, 30(6):867–881, 1997.

[106] C. Molina, L. Rodríguez Ariza, D. Sánchez, and M. A. Vila Miranda. A new fuzzy multidimensional model. *IEEE T. Fuzzy Systems*, 14(6):897–912, 2006.

[107] Carlos Molina, José-María Serrano, Daniel Sánchez, and María Amparo Vila Miranda. Measuring variation strength in gradual dependencies. In *EUSFLAT Conf. (1)'07*, pages 337–344, 2007.

[108] M. Pinto Molina. *El resumen documental:paradigmas, modelos y métodos*. Fundación Germán Sánchez Ruipérez, Salamanca, 2 edition, 2001.

[109] U. Neisser. *Psicología cognoscitiva*. Trillas, México, 1 edition, 1976.

[110] Bernd Neumann and Ralf Möller. On scene interpretation with description logics. *Image and Vision Computing*, 26:82–101, 2008.

[111] A. Niewiadomski. Six new informativeness indices of data linguistic summaries. In *AWIC*, pages 254–259, 2007.

[112] A. Niewiadomski. A type-2 fuzzy approach to linguistic summarization of data. *IEEE T. Fuzzy Systems*, 16(1):198–212, 2008.

[113] A. Niewiadomski. On finity, countability, cardinalities, and cylindric extensions of type-2 fuzzy sets in linguistic summarization of databases. *IEEE T. Fuzzy Systems*, 18(3):532–545, 2010.

[114] A. Niewiadomski and Oskar Korczak. Methods of evaluating degrees of truth for linguistic summaries of data: A comparative analysis. In *ICAISC (1)*, pages 160–167, 2010.

[115] S. Méndez Nuñez and G. Triviño. Combining semantic web technologies and computational theory of perceptions for text generation in financial analysis. In *FUZZ-IEEE 2010, IEEE International Conference on Fuzzy Systems, Barcelona, Spain, 18-23 July, 2010, Proceedings*, pages 1–8, 2010.

[116] J. A. O'Brien and G. M. Marakas. *Management information systems*. McGraw-Hill, 8 edition, 2008.

[117] M. Ortolani, H. Hofer, D. Patterson, F. Hoeppner, and M. Berthold. Fuzzy information granules in time series data. In *World Congress on Computational Intelligence*, pages 695–699, 2002.

[118] D. Peña. *Analisis de Series Temporales*. Alianza, 2005.

[119] Zheng Pei, Yang Xu, Da Ruan, and Keyun Qin. Extracting complex linguistic data summaries from personnel database via simple linguistic aggregations. *Inf. Sci.*, 179(14):2325–2332, June 2009.

[120] M. Pereira-Fariña, L. Eciolaza, and G. Triviño. Quality assessment of linguistic description of data. In *ESTYLF 2012, XVI Congreso Español Sobre Tecnologías y Lógica Fuzzy, Valladolid, Spain, 1-3 Frebruary, Proceedings*, pages 608–613, 2012.

[121] F. E. Petry and Lei Zhao. Data mining by attribute generalization with fuzzy hierarchies in fuzzy databases. *Fuzzy Sets Syst.*, 160(15):2206–2223, August 2009.

[122] S. Philipp-Foliguet, M. Bernardes Viera, and A. Albuquerque Araujo. Segmentation into fuzzy regions using topographic distance. *Proceedings of the XIV Brazilian Symposium on Computer Graphics and Image Processing*, pages 282–288, 2001.

[123] D. Pilarski. Linguistic summarization of databases with quantirius: a reduction algorithm for generated summaries. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 18(3):305–331, 2010.

[124] B. Prados-Suarez, J. Chamorro-Martínez, D. Sánchez, and J. Abad. Region-based fit of colour homogeneity measures for fuzzy image segmentation. *Fuzzy Sets and Systems*, 158(3):215–229, 2007.

[125] B. Prados-Suárez, D. Sánchez, and J. Chamorro-Martínez. A similarity measure between fuzzy regions to obtain a hierarchy of fuzzy image segmentations. In *Proceedings WCCI 2008*, pages 1647–1654, 2008.

[126] A. Ramos-Soto, F. Díaz-Hermida, and A. Bugarín. Construcción de resúmenes lingüísticos informativos sobre series de datos meteorológico: Informes climáticos de temperatura. In *ESTYLF 2012, XVI Congreso Español Sobre Tecnologías y Lógica Fuzzy, Valladolid, Spain, 1-3 Frebruary, Proceedings*, pages 642–649, 2012.

[127] G. Raschia and N. Mouaddib. Saintetiq: a fuzzy set-based approach to database summarization. *Fuzzy Sets Syst.*, 129(2):137–162, 2002.

[128] E. Reiter. Task-based evaluation of nlg systems: Control vs real-world context. In *Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop*, pages 28–32, 2011.

[129] E. Reiter and A. Belz. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558, 2009.

[130] E. Reiter and R. Dale. Building applied natural language generation systems. *Journal of Natural Language Engineering*, 12:57–87, 1997.

[131] E. Reiter and R. Dale. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press, 2000.

[132] E. Reiter, S. Sripada, J. Hunter, J. Yu, and I. Davy. Choosing words in computer-generated weather forecasts. *Artif. Intell.*, 167(1-2):137–169, September 2005.

[133] Mª A. Moreno Reques. *El Resumen documental : Normas de elaboración : Textos de Archivística, Biblioteconomía, Museología y Documentación*. Madrid: Estudio de Técnicas Documentales, 1-13 edition, 2007.

[134] S. Richard. Quality-driven service agreement as performance indicators. In Newcastle University P. Wressell, editor, *Proceedings of the 1st Northumbria International Conference on Performance Measurement in Libraries an Information Services*. Springer, Heidelberg, 1995.

[135] V. Rieser and O. Lemon. Natural language generation as planning under uncertainty for spoken dialogue systems. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 683–691, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[136] M. Ros, M. Pegalajar, M. Delgado, A. Vila, D. T. Anderson, J. M. Keller, and M. Popescu. Linguistic summarization of long-term trends for understanding change in human behavior. In *FUZZ-IEEE 2011, IEEE International Conference on Fuzzy Systems, Taipei, Taiwan, 27-30 June, 2011, Proceedings*, pages 2080–2087, 2011.

[137] R. Sambaraju, E. Reiter, R.t Logie, A. Mackinlay, C. McVittie, A. Gatt, and C. Sykes. What is in a text and what does it do: Qualitative evaluations of bt-nurse using content analysis and discourse analysis. In *In proceeding of: 13th European Workshop on Natural Language Generation*, pages 22–31, 2011.

[138] D. Sanchez-Valdes, A. Alvarez-Alvarez, and G. Triviño. Linguistic description of the traffic evolution in roads. In *ESTYLF 2012, XVI Congreso Español Sobre Tecnologías y Lógica Fuzzy, Valladolid, Spain, 1-3 Frebruary, Proceedings*, pages 614–619, 2012.

[139] D. Sankoff and J.B. Kruskal. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison.* Addison-Wesley, 1983.

[140] D. Saupe, M. Ruhl, R. Hamzaoui, L. Grandi, and D. Marini. Optimal hierarchical partitions for fractal image compression. In *IEEE Int. Conf. on Image Processing ICIP'98*, 1998.

[141] I. Savnik, G. Lausen, H.P. Kahle, H. Spieckers, and S. Hein. Algorithm for matching sets of time series. In *Int. Conf. on Principles of Data Mining and Knowledge Discovery*, pages 277–288, 2000.

[142] S. Schockaert, M. De Cock, , and E. E. Kerre. Spatial reasoning in a fuzzy region connection calculus. *Artificial Intelligence*, 173(2):258–298, 2009.

[143] P. Sebastiani, M. Ramoni, P.R. Cohen, J. Warwick, and J. Davis. Discovering dynamics using bayesian clustering. In *3rd International Sympposium on Intelligent Data Analysis*, pages 199–209. Springer, Berlin, 1999.

[144] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications.* Springer, 2000.

[145] P. Smyth. Clustering sequences with hidden markov models. In *Advances in Neural Information Processing Systems*, volume 9, pages 648–654, 1997.

[146] Daniel Sánchez, José Serrano, Ignacio Blanco, Maria Martín-Bautista, and María-Amparo Vila. Using association rules to mine for strong approximate dependencies. *Data Mining and Knowledge Discovery*, 16:313–348, 2008. 10.1007/s10618-008-0092-3.

[147] Daniel Sánchez and Andrea G. B. Tettamanzi. *Fuzzy quantification in fuzzy description logics*, pages 135 – 159. Capturing intelligence ; 1. Elsevier, Amsterdam, 2006.

[148] P. Sobrevilla and E. Montseny. Fuzzy sets in computer vision: An overview. *Mathware & Soft Computing*, 10:71–83, 2003.

[149] José M. Soto-Hidalgo, Jesús Chamorro-Martínez, and Daniel Sánchez. A new approach for defining a fuzzy color space. In *FUZZ-IEEE 2010, IEEE International Conference on Fuzzy Systems, Barcelona, Spain, 18-23 July, 2010, Proceedings*, pages 1–6, 2010.

[150] J. C. Sprott. *Chaos and time-series analysis*. Oxford University Press, 2003.

[151] S. M. Stigler. *The history of statistics : the measurement of uncertainty before 1900*. Belknap Press of Harvard University Press, 1986.

[152] Umberto Straccia. Towards spatial reasoning in fuzzy description logics. In *Proceedings Fuzz-IEEE 2009*, pages 512–517. 2009.

[153] Fardin Akhlaghian Taba, Golshah Naghdya, and Alfred Mertins. Scalable multiresolution color image segmentation. *Signal Processing*, 86:1670–1687, 2006.

[154] James C. Tilton. Method for recursive hierarchical segmentation by region growing and spectral clustering with a natural convergence criterion, 2000. Disclosure of Invention and New Technology: NASA Case No. GSC 14,328-1.

[155] James C. Tilton, Giovanni Marchisio, and Mihai Datcu. Knowledge discovery and data mining based on hierarchical segmentation of image data, 2000. a research proposal submitted October 23, 2000 in response to NRA2-37143 from NASA's Information Systems Program.

[156] G. Triviño and G. Bailador. Linguistic description of human body posture using fuzzy logic and several levels of abstraction. In *CIMSA 2007 - IEEE International Conference on Computational Intellingence for Measurement Systems and Applications, Ostuny, Italy, 27-29 June 2007*, pages 105–109, 2007.

[157] G. Triviño, A. Sanchez, A. S. Montemayor, J. J. Pantrigo, R. Cabido, and E. G. Pardo. Linguistic description of traffic in a roundabout. In *FUZZ-IEEE 2010, IEEE International Conference on Fuzzy Systems, Barcelona, Spain, 18-23 July, 2010, Proceedings*, pages 1–8, 2010.

[158] G. Triviño and A. van der Heide. Linguistic summarization of the human activity using skin conductivity and accelerometers. In *Proceedings of IPMU 2008, Torremolinos, Málaga, June 22-27, 2008*, pages 1583–1589, 2008.

[159] Z.W. Tu and S.C. Zhu. Parsing images into regions, curves and curve groups. *International Journal of Computer Vision*, 69:223–249, 2006.

[160] L. Ughetto, W. A. Voglozin, and N. Mouaddib. Database querying with personalized vocabulary using data summaries. *Fuzzy Sets Syst.*, 159(15):2030–2046, August 2008.

[161] E. Uriel and A. Peiró. *Introducción al análisis de series temporales*. Editorial AC, 2000.

[162] A. van der Heide and G. Triviño. Automatically generated linguistic summaries of energy consumption data. In *Ninth International Conference on Intelligent Systems Design and Applications, ISDA 2009, Pisa, Italy , November 30-December 2, 2009*, pages 553–559, 2009.

[163] M. A. Vila, J. C. Cubero, J. M. Medina, and O. Pons. The generalized selection: an alternative way for the quotient operations in fuzzy relational databases. In B. Bouchon-Meunier, R. Yager, and L. Zadeh, editors, *Fuzzy Logic and Soft Computing*. World Scientific Press, 1995.

[164] W. A. Voglozin, G. Raschia, L. Ughetto, and N. Mouaddib. Querying a summary of database. *J. Intell. Inf. Syst.*, 26(1):59–73, 2006.

[165] W. W. S. Wei. *Time series analysis. Univariate and Multivariate methods*. Addison Wesley, 1990.

[166] A. Wilbik, J. M. Keller, and G. L. Alexander. Linguistic summarization of sensor data for eldercare. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Anchorage, Alaska, USA, October 9-12, 2011*, pages 2595–2599, 2011.

[167] R. R. Yager. A new approach to the summarization of data. *Information Sciences*, (28):69–86, 1982.

[168] R. R. Yager. General multiple-objective decision functions and linguistically quantified statements. *International Journal of Man-Machine Studies*, 21(5):389–400, 1984.

[169] R. R. Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Trans. Syst. Man Cybern.*, 18(1):183–190, January 1988.

[170] R. R. Yager. On linguistic summaries of data. In *Knowledge Discovery in Databases*, pages 347–366. 1991.

[171] R. R. Yager. Families of owa operators. *Fuzzy Sets and Systems*, 59:125–148, 1993.

[172] R. R. Yager. Linguistic summaries as a tool for database discovery. In *FQAS*, pages 17–22, 1994.

[173] R. R. Yager. Toward a language for specifying summarizing statistics. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 33(2):177–187, 2003.

[174] R. R. Yager. A human directed approach for data summarization. In *IEEE International Conference on Fuzzy Systems*, pages 707–712, 2006.

[175] R. R. Yager, K. M. Ford, and A. J. Cañas. An approach to the linguistic summarization of data. In *Uncertainty in Knowledge Bases, 3rd International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU '90, Paris, France, July 2-6, 1990, Proceedings*, pages 456–468, 1990.

[176] R. R. Yager and F. E. Petry. A multicriteria approach to data summarization using concept ontologies. *IEEE T. Fuzzy Systems*, 14(6):767–780, 2006.

[177] J. Yu, E. Reiter, J. Hunter, and C. Mellish. Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, 13(1):25–49, 2007.

[178] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

[179] L A Zadeh. The concept of a linguistic variable and its application to approximate reasoning. *Information Sciences*, 8(3):199–249, 1975.

[180] L. A. Zadeh. A computational approach to fuzzy quantifiers in natural languages. *Computing and Mathematics with Applications*, 9(1):149–184, 1983.

[181] L. A. Zadeh. Soft computing and fuzzy logic. *IEEE Software*, 11(6):48–56, 1994.

[182] L. A. Zadeh. Generalized theory of uncertainty (GTU)–principal concepts and ideas. *Computational Statistics & Data Analysis*, In Press, Uncorrected Proof, 2006.

[183] L. A. Zadeh. Is there a need for fuzzy logic? *Inf. Sci.*, 178(13):2751–2779, 2008.

[184] L.A. Zadeh. A prototype-centered approach to adding deduction capability to search engines-the concept of protoform. In *Fuzzy Information Processing Society, 2002. Proceedings. NAFIPS. 2002 Annual Meeting of the North American*, pages 523 – 525, 2002.

[185] L. Zhang, Z. Pei, and H. Chen. Extracting fuzzy linguistic summaries based on including degree theory and fca. In *Proceedings of the 12th international Fuzzy Systems Association world congress on Foundations of Fuzzy Logic and Soft Computing*, IFSA '07, pages 273–283, Berlin, Heidelberg, 2007. Springer-Verlag.