GeoAcademy: algorithm and platform for the automatic detection and location of geographic coordinates in scientific articles

Jesús Cascón-Katchadourian[1], Francisco Carranza-García[2], Carlos Rodríguez-Domínguez[3] and Daniel Torres-Salinas[4]

[1] *cascon@ugr.es*
University of Granada, Faculty of communication and documentation, Dept of Information and Comunication, 18071 Granada (España)

[2] *carranzafr@ugr.es*
University of Granada, ETSIIT, Dept of Software Engineering, 18014 Granada (España)

[3] *carlosrodriguez@ugr.es*
University of Granada, ETSIIT, Dept of Software Engineering, 18014 Granada (España)

[4] *torressalinas@ugr.es*
University of Granada, Faculty of communication and documentation, Dept of Information and Comunication, 18071 Granada (España)

Abstract
This document describes the GeoAcademy project, whose main objective is to automatically geolocate scientific articles, downloaded from general scientific databases such as Scopus or WoS. This geolocation is carried out on the content of the document, either through the capture of possible geographical coordinates that the document has, or toponyms that may appear in the document through an algorithm created for this purpose. In the methodology we explain the steps that have been taken in this project to create a sample database with articles that deal with Sierra Nevada (Spain) and the creation and design of the algorithm. The results show the technical data of the application of the algorithm on the database and its success rate, as well as a description of the platform created to graphically display the geolocated documents on a web map. Finally, in the discussion, we define the difficulties encountered, the possible bibliometric applications and its usefulness as a tool for viewing and retrieving information

Introduction
Geolocation is the possibility of locating information in specific geographic spaces to treat and process it (Ramos Vacca & Bucheli Guerrero, 2015). Technological advances in restitution by aerial photogrammetry and digitized cartography have contributed to this (Cortés-José, 2001), as have new procedures for cartographic writing and map printing, the appearance around 2005 of Google Maps, Bing Maps, OpenStreetMap and numerous programs to create maps such as OpenLayer, Leaflet, CartoDB or MapTiler (Cascón-Katchadourian & Ruiz Rodríguez, 2016). In the field of scientific evaluation and bibliometrics, geolocation techniques have been commonly used to locate scientific articles based on the origin of its authors (Catini et al., 2015). However, few studies have been devoted to analyzing a set of scientific publications based on their geographical coordinates and geolocating them on a map.
We would like to highlight here two current initiatives due to their importance and because they include several institutions behind their prototyping. One of them is GEOUP4, this "web portal shows on an interactive map the geolocation of the academic items of the repositories of the UPC, UPCT, UPM and UPV polytechnic universities grouped in the UP4 Association" (http://geo.up4. is/). The other project is JournalMap (https://www.journalmap.org/) a cooperative project between the USDA-ARS Jornada Experimental Range in Las Cruces, NM and the Idaho Chapter of The Nature Conservancy. This is one of the components of another tool called Landscape Toolbox. It is a project that geolocates scientific production based on the geographical location where the study is carried out to observe which areas have been studied and in which there are gaps. The present paper links with these two projects. Our study is

original and useful since it aims to geolocate the scientific articles in an automatic way through algorithms created for this purpose and not in a manual or semi-automatic way.
More specifically, there are three main objectives for this project, firstly (1) to develop an algorithm that allows the coordinates to be extracted from a collection of documents to identify exactly which places these studies deal with. The second objective (2) is that once the locations have been identified by the algorithm, they will be displayed on a map through an online platform. In the third objective (3), the platform will integrate a layer of bibliometric and/or altmetric information that will allow one to know the volume of production according to coordinates as well as different data on the scientific and social impact. It should be mentioned that in this paper will be presented only the results of objectives 1 and 2 applied to a set of scientific works that study the Sierra Nevada mountain range (Granada, Spain)

Material and methods

In order to create a collection of documentary information about Sierra Nevada, a search was done in the Scopus database. The Scopus database has been chosen for the facilities it offers for exporting full-text documents, thanks to Scopus Document Download Manager. With this type of search, we wanted to find scientific articles that dealt with the Sierra Nevada mountain range located in the province of Granada (Spain). As there were other mountain ranges with the same name in other parts of the world, the search for the term Spain or Granada was added. This search found 623 articles. The second step was the processing of the information for storage. After the pertinent manual verification work (the geolocation of the documents is automatic, not the selection of the sample), there are 447 documents with a complete associated pdf (not only the abstract or title) and which are openly accessible, the other 176 (623-447) records do not have an associated pdf, are incomplete or are not an open publication, of which 424 are about the Spanish Sierra Nevada and not about the Sierra Nevada of California or Peru. The third step was the identification of the coordinates. Although there are many types of coordinate systems, the most common that this project has found in the sample are the following types and subtypes of coordinates: the geographic coordinate system and the coordinate system UTM (Universal Transverse Mercator). The geographic coordinate system is subdivided into sexagesimal (degrees, minutes and seconds) and decimal (degrees and decimals). The UTM coordinate has multiple variants in that the authors express them in different ways, with or without the letters of the cardinal points, specifying the use or not (in our case it is 30S). Finally, in environmental studies, they usually use a subdivision of the use 30S, which is used in army maps and which is reflected in articles with VG, VF, WG and WF. Finally, our database contains 424 bibliographic references linked to their corresponding PDFs. This database has been used for the design and training of the algorithm.
For the geolocation of the contributions, we have developed a text mining algorithm based on the extraction of information through regular expressions and toponyms - keywords (see Figure 1). To carry out this automatic geolocation process, we design an algorithm based on 4 stages:

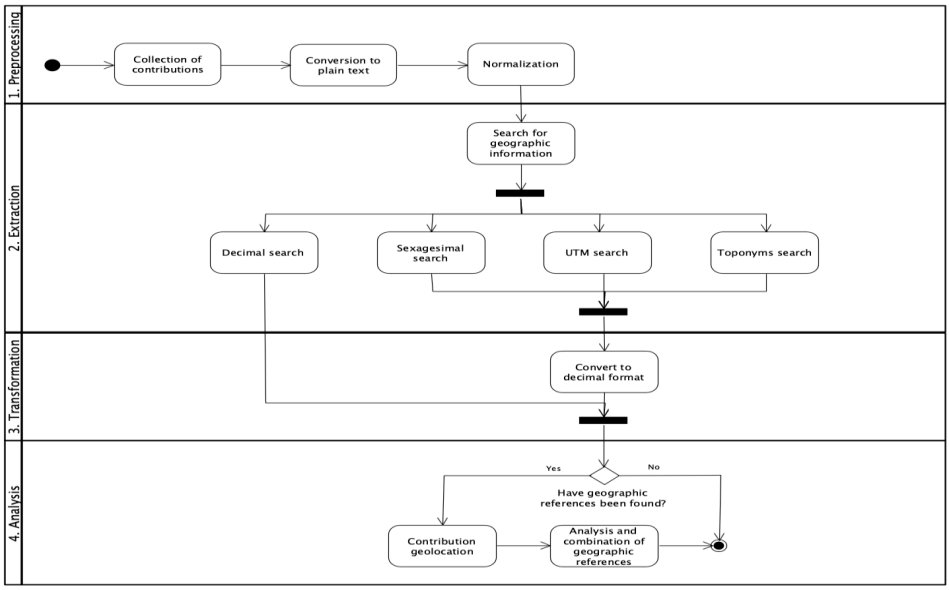
1. Preprocessing: In this stage, the contributions of the sample are processed in PDF format, converting the text of each one to a processable format (plain text), normalized and without all unnecessary information. For the conversion of PDF files to plain text, we have used pdftotext which is an open-source command-line utility.

2. Extraction: Search and extraction of geographical references in the text using two methods, (i) regular expressions for the search of geographical coordinates in their different formats (decimal, sexagesimal and UTM) and (ii) search of the frequency of appearance of place names. The database of toponyms for this study has been developed by combining all sources of public and georeferenced information of the Junta de Andalucía such as NGA and ITACA of the Andalusian Institute of Statistics and Cartography.

3. Transformation: In order to georeference the results obtained on an interactive map, all the results are processed to unify them in decimal format (latitude, longitude). To convert between the different geographic representations, we have used Proj4js1 which is a library to transform point coordinates from one coordinate system to another, including datum transformations.

4. Analysis: Finally, once we have all the information that has been extracted in the same format, a small analysis is carried out to decide, through heuristic behavior, if it is possible to geolocate the contribution. For this, we have mainly followed three criteria: (i) if geographic coordinates and place names have been found, we combine this information to select the coordinate that is textually closest to the most frequent place name; (ii) if only geographic coordinates have been identified, it is geolocated within the most referenced area; and (iii) if only toponyms have been recognized, we geolocate it in the one with the highest frequency as long as it exceeds a certain threshold (this threshold is calculated as the simple mean of the frequencies of the toponyms in the entire sample of this study)

Figure 1. Algorithm workflow.



Results

Table 1 shows a summary of the results of the algorithm (figure 1) of automatic detection of geolocations being applied to the collection of 424 scientific articles on Sierra Nevada. In total, the algorithm has been able to geolocate 157 articles with coordinates, 37% of the total. One of the tools used to increase the number of geolocated articles and elements has been the use of place names that have allowed us to geolocate 5025 place names of 189 different articles. There are 2.6 place names on average per scientific article. Finally, the number of works that have been geolocated using the algorithm is 346, that is to say 81.6% of the original document collection has been located, either by coordinates or toponyms. It should be mentioned that one of the fundamental aspects to improve the success rate of the algorithm has been the use of place names.

Table 1. Indicators and general statistics in the algorithm training process

A) General indicators related to geolocation by coordinates	
A1. Number of articles analysed in the study:	424
A2. Number of articles containing geographical coordinates (157 + 43 (images) + 5	205
A.3 Number of articles containing geographical coordinates identified by the	157
A.4. Percentage of articles geolocated through geographical coordinates (A3/A1)	37.03%
B) Indicators related to geolocation by place names	
B.1. Number of place names identified by the algorithm	5025
B.2 Number of articles geolocated by place names	189
B3. . Percentage of articles geolocated by place names (B2/A1)	44.57%
C) Findings	
C.1 Number of geolocalisations achieved (coordenates + place names) (A.3+B.2)	346
C.2. Percentage of articles geolocated from the the total number of articles (C1/A1)	81.6%

Once objective 1 had been achieved, focus moved to the creation of an application to view scientific articles on a digital map based on the different coordinates they contain. The portal that has been developed currently contains the collection of documents on Sierra Nevada that has been analysed in this paper. The portal is operational at the following web address in beta format: https://geoacademy.everyware.es/.

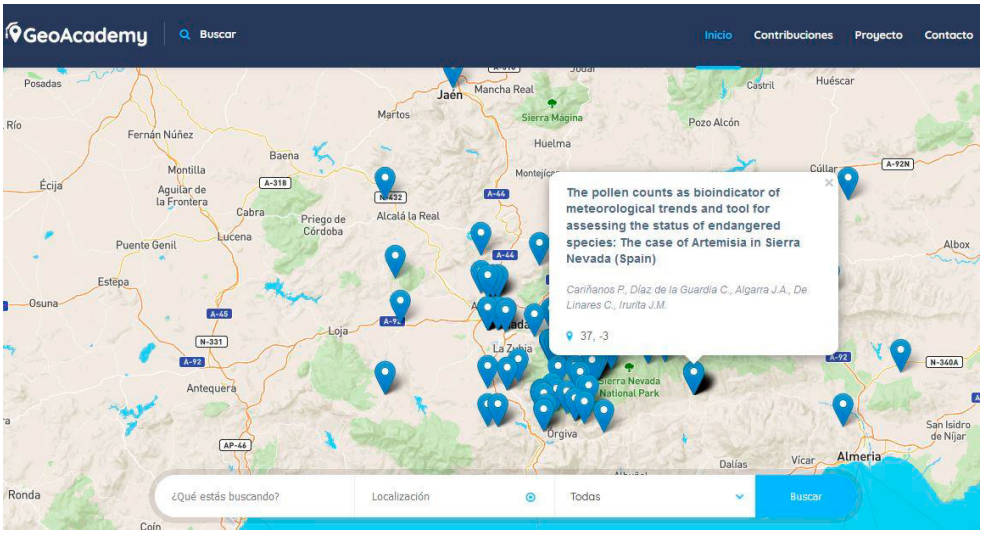


Figure 2. GeoAcademy application showing geolocated locations for the Sierra Nevada collection of articles.

As we can see in Figure 2, the geolocations detected by the algorithm are distributed, by clicking on each one we see the reference information of the paper and a link to view it. At the bottom you can filter by type of document you are looking for (article, conference, thesis, etc.) or you can search by keyword. The different tabs give us access to a traditional search engine with filters and a search radius, as well as a tab where the project and its members are described, as well as a contact form. Likewise, each scientific work that has been processed includes both the metadata of the databases (in this case Scopus) and all the metadata generated automatically by

the algorithm. In our portal, the coordinates in decimal format as well as a complete list of the identified place names are added to the bibliographic description.

Discussion

This study contains a number of limitations that we proceed to list and which have mainly to do with processing. The first problem relates to the processing of PDF documents. We found several that did not have an associated pdf, for others the pdf requires payment, or only the title or the title and abstract are available. This limits the number of results you can cover. Secondly, the coordinates have some formal standardization, whereas in practice each field of knowledge has different guidelines for writing the coordinates, with greater variability with the UTMs that the algorithm has overcome without problems. Another problem is that several studies show the coordinates in image form, that makes them more difficult to extract, although in the future through OCR technology they could be captured. At the linguistic level, words with multiple means in the toponymy can be problematic, which can give both false positives and false negatives.

The success rate of the Geoacademy algorithm on all documents is 81.6%, a high percentage for this type of study. We will continue to train the algorithm with much larger documentary collections related to other geographical features. Likewise, it will be applied in other contexts, such as archaeological where most of the articles include precise geographic coordinates. It could also be applied to digital humanities, since interesting projects are being carried out where historical works are geolocated through the place names that appear in them.

The third objective of this work is to provide the platform with a layer of information capable of representing information of a bibliometric and altmetric nature, that is, of scientific and social impact. In this sense, future development implies providing the GeoAcademy platform with the following functionalities 1) Filter locations based on indicators (Impact Factor, Number of citations, Altmetric Attention Score ...) 2) Use of geoposition markers differentiated according to values of the indicators. Likewise, the platform will present a bibliometric summary of the different coordinates and locations, offering bibliometric information at a level not currently seen. Future developments could include its inclusion with an industry tool such as bibliometrix.

It would also be very interesting if users of large scientific databases such as Scopus or Web of Science, having performed a search with a list of results, could see said list geographically on a map, applying our algorithm for a better user experience. That is to say, to integrate our project in their platforms. Finally, numerous studies do not work with geographical points, we are currently working on how to show these studies on our platform

References

Cascón-Katchadourian, J., & Ruiz Rodríguez, A. Á. (2016). Descripción y valoración del software MapTiler: Del mapa escaneado a la capa interactiva publicada en la web.

Catini, R., Karamshuk, D., Penner, O., & Riccaboni, M. (2015). Identifying geographic clusters: A network analytic approach. Research policy, 44(9), 1749-1762.

Cortés-José, J. (2001). El documento cartográfico. En J. Jiménez Pelayo, J. Monteagudo López-Menchero, & F. J. Bonachera Cano. La documentación cartográfica : Tratamiento, gestión y uso (pp. 37-113). Huelva: Universidad de Huelva.

Ramos Vacca, I. D., & Bucheli Guerrero, V. A.. (2015). Automatic geolocation of the scientific knowledge: Geolocarti. Paper presented at the - 2015 10th Computing C*olombian Conference (10CCC),* pp. 416-424. doi:10.1109/ColumbianCC.2015.7333454