Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 199 (2022) 718–724

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th Information Technology and Quantitative Management

(ITQM 2020 & 2021)

Profiling clients in the tourism sector using fuzzy linguistic models based on 2-tuples

Itzcóatl Bueno[a], Ramón A. Carrasco[b], Carlos Porcel[c], Enrique Herrera-Viedma[c,1,]

[a]*Statistics Faculty, Complutense University of Madrid. Avenida Puerta de Hierro, s/n, 28040 Madrid.*
[b]*Department of Management and Marketing, Complutense University of Madrid, 28223, Madrid, Spain*
[c]*Department of Computer Science & Artificial Intelligence, University of Granada. Calle Periodista Daniel Saucedo Aranda s/n, 18071, Granada.*

Abstract

Customer segmentation is a key piece of a company's business strategy. This paper presents a segmentation of the online users of tourism platforms through the recency, frequency and helpfulness of the users. 2-tuples model is applied to these variables to be more precise without loss of information. In addition, the functionality of the proposal made by the authors is verified through a use case in which TripAdvisor opinioners are segmented in reference to the experience lived in hotels and tourist accommodation.

Keywords: fuzzy linguistic modeling, customer opinion value, Multi-Criteria Decision-Making, online customers segmentation, 2-tuples model
2010 MSC: 00-01, 99-00

1. Introduction

User publications on online platforms provide information about their consumption habits [1, 2, 3]. In the case of tourism, it can be extracted how trustworthy users are perceived by the rest of the community [4] and how often and with what periodicity they comment on their experiences in tourist accommodation and hotels. This information is of great value for hotel chains, especially in times of COVID-19 that has affected both the tourism sector, to segment [5, 6] these users and to be able to develop commercial strategies based on the type of users. In this article we intend to do this segmentation of online opinion makers in the TripAdvisor online portal through a hierarchical clustering based on 2-tuples measures showing a use case with real data and supported by the methodology presented in [7]. Although the RFM model has been widely used in the literature [8, 9, 10] for customer segmentation, the substitution of the monetary dimension for the utility dimension has not been much explored until now. In [11], clustering is applied to users with the RFH model using the k-means methodology.

*Corresponding author.
E-mail address: viedma@decsai.ugr.es.

10.1016/j.procs.2022.01.089

However, our proposal has the novelties of using the 2-tuples model on the RFH and also of using hierarchical clustering techniques that allow weighting the different dimensions of the user. The rest of the article is structured as follows: in section 2 the key concepts in the development of the model are presented, in section 3 a use case on real TripAdvisor data is presented, and finally some conclusions are presented in section 4.

2. Preliminaries

In this section we present the main components in which is based our proposal: Fuzzy Linguistic Modeling, RFM model and its 2-tuple extension, and AHP.

2.1. Fuzzy linguistic modeling

Fuzzy logic is presented as an alternative to traditional logic, with the aim to introduce grades of uncertainty to the statements that it interprets [12]. Then, Fuzzy Linguistic Modeling is a tool that allows representing qualitative aspects. It is based on the concept of linguistic variables, that is, variables whose values are not numbers, but words or statements expressed in natural or artificial language [12]. Each linguistic value its characterized by a syntactic value or label, and a semantic value or meaning. The label is a word or statement that belongs to a set of linguistic terms and the meaning is a fuzzy subset in a discourse universe.

2.1.1. 2-tuple fuzzy linguistic modeling

Herrera and Martínez developed in [13] a fuzzy linguistic representation model in which a pair of values (s,α) defined as 2-tuple represent the linguistic information. This pair is formed by s that is a linguistic label, and by α that represents the value of the symbolic translation.

In their paper, they consider that let a linguistic term set $S=\{s_0,...,s_\kappa\}$ and a value that represent the result of a symbolic aggregation operation $(\beta\in[0,\kappa])$, then the 2-tuple expressing the equivalent information to β is obtained by:

$$\begin{aligned}&\Delta[0,\kappa]\longrightarrow S\times[-0.5,0.5)\\&\Delta(\beta)=(s_i,\alpha),\ \text{with}\begin{cases}s_i & i=round(\beta)\\ \alpha=\beta-i & \alpha\in[-0.5,0.5)\end{cases}\end{aligned}\tag{1}$$

where $round(\cdot)$ is the round operation, s_i has the closest index label to β and α is the value of the symbolic translation. Moreover, from a 2-tuple is always possible to return its equivalent numerical value $\beta\in[0,\kappa]$ by a Δ^{-1} function. Other useful operator is the known as negation operator, which is described as $neg((s_i,\alpha))=\Delta(\kappa-\Delta^{-1}(s_i,\alpha))$. Finally, in our model we just use one aggregation operator which is the weighted average and it is defined as:

Definition 1. *Let $S=\{(s_1,\alpha_1),...,(s_n,\alpha_n)\}$ be a set of linguistic 2-tuple and $\Omega=\{\omega_1,...,\omega_n\}$ be their associated weights. The 2-tuple weighted average $\bar{S}^{\omega}$ is:*

$$\bar{S}^{\omega}[(s_1,\alpha_1),...,(s_n,\alpha_n)]=\Delta\left(\frac{\sum_{i=1}^{n}\beta_i\cdot\omega_i}{\sum_{i=1}^{n}\omega_i}\right)\tag{2}$$

For further information on the 2-tuple linguistic representation model, see [13] and [14].

2.2. RFM 2-tuples

This segmentation technique [15] is formed of the following three dimensions: Recency (R) represents the length of a time period since the most recent acquisition, visit to the establishment, or post in a web platform. It is measure in days, months, years, or any other time unit. Frequency (F) represents the total number of acquisitions, establishment visits or opinions posted during the studied period. Finally, Monetary (M) represents the total economic value of the acquisitions made during the studied period. Once the analysis period is chosen, the

aforementioned dimensions are gathered at a user level. In addition, the users are arranged according to each RFM measure and are grouped in equal size classes, usually in quintiles. Thus, the measures (recency, frequency and monetary) are mutated into ordinal scores. Finally, a weighted average of the R, F and M scores by the user-defined weights as the defined in Equation 3 is calculated with the aim to determine a unique judge that describes jointly each user according to their scores in a unique RFM Score.

$$RFM_i = \omega_R \times R_i + \omega_F \times F_i + \omega_M \times M_i \tag{3}$$

Incorporating the 2-tuple model it solves the main limitation of RFM model, its lack of accuracy in the scores calculation. First, the symmetric and uniformly distributed domain S is defined using linguistic labels. The labels have a semantic meaning, depending on the use case, for recency, frequency and monetary variables. So, let $S = \{s_0, ..., s_\kappa\}$ with $\kappa = 4$ its definition is showed in Figure 1.

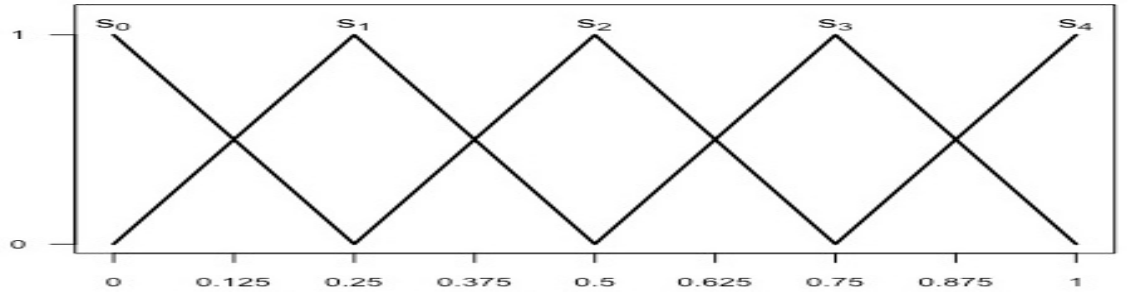

Fig. 1: Definition of the set S

Consequently, for each user we have $U_i = (U_{1i}, U_{2i}, U_{3i})$ $i = 1, ..., n$ where each one represent the score in recency, frequency and monetary for user i. Firstly, users are arranged in ascending order according to each RFM measure. Secondly, the ranking of each user regard to each of the three measures is defined as:

$$PercRank_{ij} = \frac{rank_{ij} - 1}{n - 1}$$

with $rank_{ij} \in, \{1, ..., n\}$, $n > 1$, $PercRank_{ij} \in [0, 1]$, $i = 1, ..., n$ and $j = \{1, 2, 3\}$. Therefore, the 2-tuple score U_{ij} is determined by applying Equation 1 to the the ranking of each user regard to each of the three measures. Moreover, in the case of Recency it is necessary to apply the negation operator ($neg(\cdot)$) because the highest scores mean the most recent users. Finally, the 2-tuple RFM Score, which describes jointly the RFM scores, is obtained using the Equation 2 for each user user i as $RFM_i^{Score} = \bar{S}^{\omega}[U_{ij}]$ with the weights $\Omega = \{\omega_R, \omega_F, \omega_M\}$ defined by the user.

2.3. Analytical Hierarchy Process (AHP)

AHP [16] was presented by Saaty [17, 18, 19]. Through a pairwise comparison matrix, this technique tries to provide a priority scale to a set of alternatives based on expert judgment on different criteria. A scale of absolute judgements presented in [19] is used to make the comparison of criteria, it is interpreted as how much more one criterion dominates another. However, these judgments could be inconsistent and therefore it is necessary to check their logic through the Consistency Ratio that is determined by $CR = \frac{CI}{RI}$ where CI is the Consistency Index obtained as $\frac{\lambda_{max} - n}{n-1}$ and RI is the Random Index, which represents consistency of a randomly generated pairwise comparison matrix. Being λ_{max} the highest eigenvalue of pairwise comparison matrix A. Particularly, if $CR \leq 0.1$ the inconsistency is tolerable, and so a trustworthy result will be awaited from AHP model. Once the consistency of each pairwise comparison matrix is checked, the priority scale in each level of the hierarchy structure is obtained. Finally, these priority scales are combined by multiplying them by the priority of their parent nodes and adding for all such nodes.

3. Proposed model and application to the segmentation of TripAdvisor users

In this section we apply the methodology presented in [7] over the dataset used in [20, 21]. The authors crawled hotel reviews from TripAdvisor for their paper, but they provide data for a longer period[1]. Although the dataset is formed by information collected from TripAdvisor[2] of hotels, we selected for the purpose of this work the following variables: user's nickname, publication date and number of helpfulness votes given by other users to an opinion. We filter the period of study to users who posted reviews between the years 2007 and 2008.

3.1. Obtaining quality customers

In this first step, to study how useful the opinions published by users are for the rest of online readers, it is necessary to filter the data set through quality criteria to maintain significant users for the model. Therefore, for this use case the same quality criteria have been established as those used in [7]: all anonymous users are discarded; a minimum of 2 reviews must be posted during the analyze period in order to take a user into account; and users with an unusual number of opinions are discarded.

3.2. Obtaining the opinion value of each customer with 2-tuple RFH model

The goal of this stage is to determine the opinioner value using the 2-tuples RFH methodology presented in [7]. Firstly, we need to obtain the three dimensions necessary for each user (recency, frequency and helpfulness) from our original dataset. In this sense, the variables will be the number of days since the last time the user posted a review on TripAdvisor as the Recency (R); the number of opinions posted on TripAdvisor by an user during the period analyzed will refer to the Frequency (F); and the number of helpfulness votes a user receives on his posts from the rest of users during the period analyzed is the Helpfulness (H). To these three measures, percentage of the ranking equation is applied and the corresponding value is obtained in the ranking percentage, so Table 1 would be the structure of the dataset used in this use case.

							2-tuples		
Author	**Recency**	**Frequency**	**Helpfulness**	**R Score**	**F Score**	**H Score**	**R Score**	**F Score**	**H Score**
OffTheBeatenPath	626	3	6	0.7705	0.8257	0.4819	$(S_3, 0.08)$	$(S_3, 0.3)$	$(S_2, -0.07)$
PieroG	268	2	3	0.1371	0	0.0981	$(S_1, -0.45)$	$(S_0, 0)$	$(S_0, 0.39)$
hdblue	642	4	8	0.8029	0.96	0.6524	$(S_3, 0.21)$	$(S_4, -0.16)$	$(S_3, -0.39)$
Calartist	447	2	4	0.3714	0	0.2409	$(S_1, 0.49)$	$(S_0, 0)$	$(S_1, -0.04)$
grcas	715	2	2	0.96	0	0	$(S_4, -0.16)$	$(S_0, 0)$	$(S_0, 0)$
btowndude	477	3	7	0.4238	0.8257	0.5771	$(S_2, -0.3)$	$(S_3, 0.3)$	$(S_2, 0.31)$

Table 1: RFH users table

Applying the RFH 2-tuple methodology on the data set shown in Table 1 the percentage of the ranking that each user occupies in each dimension is converted into a tuple as the defined in subsubsection 2.1.1 by applying the delta function defined in Equation 1. Finally, the weights used in Equation 3 are obtained using the AHP methodology presented in subsection 2.3. We assign weights according to our preferences in order to value the most from the users in this business case. These values are explained in [19] and the meaning of the presented in the pairwaise matrix of this use case are: equal importance (1), weak importance of an activity over other (2), and moderate importance of an activity over other (3). So, the pairwise comparison matrix is presented below.

$$A_1 = \begin{array}{c} \\ Recency \\ Frequency \\ Helpfulness \end{array} \begin{array}{c} \begin{array}{ccc} Recency & Frequency & Helpfulness \end{array} \\ \begin{pmatrix} 1 & 3 & \frac{1}{2} \\ \frac{1}{3} & 1 & \frac{1}{3} \\ 2 & 3 & 1 \end{pmatrix} \end{array}$$

[1] http://times.cs.uiuc.edu/ wang296/Data/

[2] `www.tripadvisor.com`

The meaning of this matrix is that the frequency users post an opinion is the least valued dimension given that frequently the users post low number of opinions. Helpfulness is the most important dimension since this is how we value how trustworthy the user is for the rest of the community. Finally, the recency is more important than the frequency but less than the helpfulness. Thus, the weighting vector obtained of applying the AHP is $\Omega = (\omega_R = 0.33, \omega_F = 0.14, \omega_H = 0.53)$.

3.3. User segmentation

Finally, we want to define user groups so that hotels can undertake different commercial strategies adapted to the profile of each user. For this purpose, we have applied a hierarchical clustering on the three dimensions of the user, for which the first step has been to use Δ^{-1} to transform from the linguistic to the numeric form in order to take distances between the variables. We have used the Gower's distance during the clustering process which allows to give weight for each variable and which is defined as $d_{ij} = \frac{\sum_{k=1}^{p} \omega_{ijk} d_{ijk}}{\omega_{ijk}}$ where ω_{ijk} is the weight for variable k between observations i and j, which have been obtained previously (Ω). The result of applying an average linkage clustering using Gower's distance is the formation of 7 different groups of users as can be seen in Figure 2, which the authors has defined as: Recently relevant users, Frequent users, Not meaningful users, New users, VIP users, Obsolete users and Fake users.

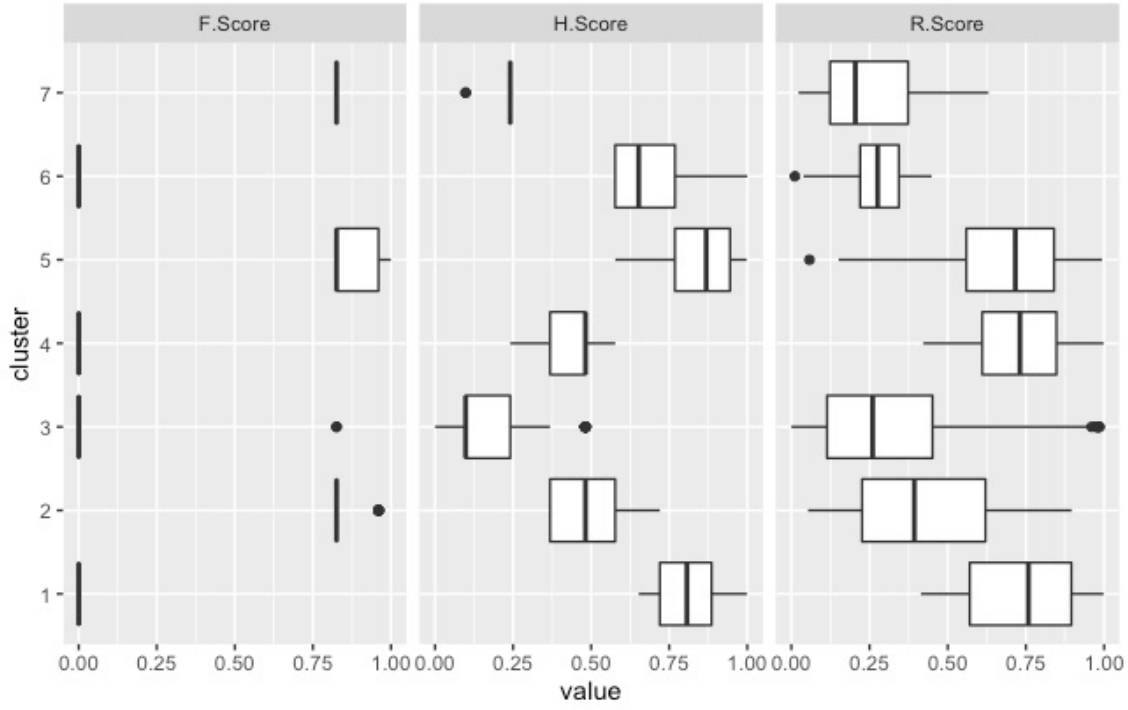

Fig. 2: Distribution of each dimension score within the clusters

In Table 2, we see how each dimension is distributed within each cluster. The x-axis of Figure 2 show the cluster id and in the y-axis it is measure the score for each of the three measures. Moreover, in Table 2 these clusters are explained in more detail.

Table 2: Segmentation description

Cluster ID	Cluster Name	Description
1	Recently relevant	It is formed by users perceived as useful by other users that published a review recently but do not post frequently. The reviews are usually exhaustive and interesting that denote that the user has recently visited the reviewed hotel.
2	Frequent	It is formed by users that post frequently but they are not perceived as useful, although they do generate trust in the rest of the users, having medium importance in the community. They are very active users but they do not finish penetrating the user community
3	Not meaningful	It is formed by users who publish infrequently, that the rest of the community does not find their opinions useful and that they have not published for a long time since their last review.
4	New	It is formed by users who publish infrequently but they have posted their last reviews recently and have been perceived as useful enough to be taken into account in the community. It includes new users in the comunnity of opinioners.
5	VIP	It is formed by users who publish very frequently, which keeps them active over time and who are perceived as very useful members in the user community.
6	Obsolete	It is formed by users who are considered useful to other users, however they have stopped publishing with very little frequency, which has made them obsolete.
7	Fake	It is formed by users who publish a lot but who do not penetrate the community as useful reviewers. This group would include fake users and trolls.

Finally, in Table 3 we can see the centroids of each cluster and the number of users in our dataset included in each one. The centroids are expressed in the 2-tuple terminology and refute the types of users and descriptions provided in Table 2.

Table 3: RFH 2-tuple cluster centroids

Cluster ID	Recency	Frequency	Helpfulness	Number of users
1	$(S_2, -0.06)$	$(S_0, 0)$	$(S_2, 0.23)$	185
2	$(S_1, -0.25)$	$(S_2, 0.39)$	$(S_1, -0.06)$	56
3	$(S_0, 0.25)$	$(S_0, 0.02)$	$(S_0, -0.34)$	423
4	$(S_2, -0.09)$	$(S_0, 0)$	$(S_1, -0.31)$	172
5	$(S_2, -0.25)$	$(S_2, 0.48)$	$(S_2, 0.38)$	115
6	$(S_0, 0.09)$	$(S_0, 0)$	$(S_2, -0.18)$	89
7	$(S_0, 0.09)$	$(S_2, 0.3)$	$(S_0, -0.14)$	11

4. Conclusions and Future Work

In this paper we present the way to obtain a segmentation of online users for the tourism sector using the RFH 2-tuples model, applying a hierarchical clustering on it. Its functionality has been tested through a use case in the TripAdvisor user community for the segmentation of hotel reviewers. The results obtained in the example presented in this work show the consistency of how online users are segmented in very well differentiated groups. These results brings the opportunity to hotels, restaurants and other services companies to take strategies in order to make changes and improve their services to attract or keep customers from the different clusters, as well as ignoring those that the model identifies as fake or trolls. The authors consider that there is a path for this model and its applications in the study of other clustering techniques as well as in the use of multigranular fuzzy linguistic modeling on the data to be segmented. Moreover, it can be applied over users who give their opinion on a set of hotels located in a geographical area, on a certain hotel group, etc.

Acknowledgments

This work has been funded by the Spanish State Research Agency through the project PID2019-103880RB-I00 / AEI / 10.13039/501100011033.

References

[1] I. Bueno, R. A. Carrasco, R. Ureña, E. Herrera-Viedma, Application of an opinion consensus aggregation model based on owa operators to the recommendation of tourist sites, Procedia Computer Science 162 (2019) 539–546.

[2] R. A. Carrasco, I. Bueno, Applying an opinion consensus aggregation model based on 2-tuple owa operators to the recommendation of accommodation, Journal of Development Research 13 (2) (2020) 42–49.

[3] M. A. Rahim, M. Mushafiq, S. Khan, Z. A. Arain, Rfm-based repurchase behavior for customer classification and segmentation, Journal of Retailing and Consumer Services 61 (2021) 102566.

[4] T. L. Ngo-Ye, A. P. Sinha, The influence of reviewer engagement characteristics on online review helpfulness: A text regression model, Decision Support Systems 61 (2014) 47–58.

[5] B. Oztaysi, M. Kavi, Fuzzy rfm analysis: An application in e-commerce, in: International Conference on Intelligent and Fuzzy Systems, Springer, 2020, pp. 1225–1232.

[6] İ. KABASAKAL, Customer segmentation based on recency frequency monetary model: A case study in e-retailing, Bilişim Teknolojileri Dergisi 13 (1) (2020) 47–56.

[7] I. Bueno, R. A. Carrasco, C. Porcel, G. Kou, E. Herrera-Viedma, A linguistic multi-criteria decision making methodology for the evaluation of tourist services considering customer opinion value, Applied Soft Computing 101 (2021) 107045.

[8] Y. Huang, M. Zhang, Y. He, Research on improved rfm customer segmentation model based on k-means algorithm, in: 2020 5th International Conference on Computational Intelligence and Applications (ICCIA), IEEE, 2020, pp. 24–27.

[9] B. Rizki, N. G. Ginasta, M. A. Tamrin, A. Rahman, Customer loyality segmentation on point of sale system using rfm and k-means, Jurnal Online Informatika 5 (2).

[10] E. Ernawati, S. Baharin, F. Kasmin, A review of data mining methods in rfm-based customer segmentation, in: Journal of Physics: Conference Series, Vol. 1869, IOP Publishing, 2021, p. 012085.

[11] A. Aakash, A. Jaiswal, Segmentation and ranking of online reviewer community: The role of reviewers' frequency, helpfulness, and recency, International Journal of E-Adoption (IJEA) 12 (1) (2020) 63–83.

[12] L. A. Zadeh, The concept of a linguistic variable and its application to approximate reasoning, in: Learning systems and intelligent robots, Springer, 1974, pp. 1–10.

[13] F. Herrera, L. Martínez, A 2-tuple fuzzy linguistic representation model for computing with words, IEEE Transactions on fuzzy systems 8 (6) (2000) 746–752.

[14] F. Herrera, L. Martínez, A model based on linguistic 2-tuples for dealing with multigranular hierarchical linguistic contexts in multi-expert decision-making, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 31 (2) (2001) 227–234.

[15] R. G. Martínez, R. A. Carrasco, J. García-Madariaga, C. P. Gallego, E. Herrera-Viedma, A comparison between fuzzy linguistic rfm model and traditional rfm model applied to campaign management. case study of retail business., Procedia Computer Science 162 (2019) 281–289.

[16] R. A. Carrasco, L. N. Forero, S. X. López, E. Herrera-Viedma, C. Porcel, Using the ahp model to improve the measurement of satisfaction in the ict sector., in: SoMeT, 2018, pp. 299–311.

[17] T. L. Saaty, Axiomatic foundation of the analytic hierarchy process, Management Science 32 (7) (1986) 841–855. doi:10.1287/mnsc.32.7.841.
URL <GotoISI>://WOS:A1986D106300006

[18] T. L. Saaty, How to make a decision - the analytic hierarchy process, Interfaces 24 (6) (1994) 19–43. doi:10.1287/inte.24.6.19.
URL <GotoISI>://WOS:A1994PZ14800002

[19] T. L. Saaty, Decision making with the analytic hierarchy process, Int. J. Services Sciences 1 (1) (2008) 83–97.

[20] H. Wang, Y. Lu, C. Zhai, Latent aspect rating analysis on review text data: a rating regression approach, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 783–792.

[21] H. Wang, Y. Lu, C. Zhai, Latent aspect rating analysis without aspect keyword supervision, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, pp. 618–626.