An Approach to Modelling User Interests Using TF-IDF and Fuzzy Sets Qualitative Comparative Analysis

Dimitris Kardaras, Stavros Kaperonis, Stavroula Barbounaki, Ilias Petrounias, Kostas Bithas

An Approach to Modelling User Interests using TF-IDF and Fuzzy Sets Qualitative Comparative Analysis

Dimitris K. Kardaras[1] and Stavros Kaperonis[2] and Stavroula Barbounaki[3] and Ilias Petrounias[4] and Kostas Bithas[5]

[1] Athens University of Economics and Business, Patission str. 76, 10434 Athens, Greece
Kardaras@aueb.gr; dkkardaras@yahoo.co.uk
[2] Panteion University of Social and Political Sciences, Syggrou ave. 136, 17671 Athens, Greece
skap@panteion.gr
[3] Merchant Marine Academy of Aspropyrgos, 10559, Athens, Greece
sbarbounaki@yahoo.gr
[4] The University of Manchester, Oxford Rd M13 9PL, Manchester, UK
ilias.petrounias@manchester.ac.uk
[5] Panteion University of Social and Political Sciences, Syggrou ave. 136, 17671 Athens, Greece
kbithas@eesd.gr

Abstract. Modelling and understanding user interests are particularly important tasks for designing services and building systems for customized solutions in web personalization and recommender systems. User generated content (UGC) constitutes a significant source of information for capturing user interests. This paper, suggests an approach to user profiling that analyses the Term Frequency (TF) and the Inverse Document Frequency (IDF) of selected tourism services by utilising the Fuzzy set Qualitative Comparative Analysis (FsQCA). It analyses a sample of customer reviews that are collected from tourism web sites. This paper considers the amount of money that customers spent during their hotel stay, as the outcome set in the FsQCA analysis. The results produce causal combinations of services that are necessary and sufficient for building customer interests models that best lead to the outcome and argue for the applicability of the FsQCA in modelling user interests.

Keywords: User interests, Fuzzy Sets Qualitative Comparative Analysis, TF, IDF

1 Introduction

Recommender systems RC utilise techniques spreading from statistics, to AI and machine learning in order to capture user interests, build user and products/services profiles and suggest the most appropriate products or services to them. RC draw on several methods for developing user references models, with user-generated-content (UGC) to represent a source with rich customer information [1, 2]. Since social media platforms allow users to exchange experience, feedbacks, opinions, complaints, etc., they provide significant information for capturing and understanding user interests

[3]. Web personalisation is another area where user profiling is necessary for developing customised web interfaces, supporting personalised search [4] that allow users to retrieve search results according to their personal needs.

2 User Profiling in Tourism

Building user interests models has also been the focus of e-tourism research studies. Drawing on behavioural, socio-economic and demographic data analysis several researchers shed light into understanding people's travel behaviour [3]. Indeed, surveys on travellers' preferences have shown that the travel selection process is complex depending among others, on personality and mood related factors, service quality issues, the Word-Of-Mouth (WOM) and the eWOM. Customers often express their experience by publishing their reviews. Sentiment analysis of user reviews provides the means for capturing and modelling users' preferences, emotions and attitudes, thus refining market segregation by grouping customers with similar needs and incentives and predicting customers' travel behaviour more precisely [5].

Collantes and Mokhtarian [6] claim that a variety of personality factors such as: personality traits, travel-related behaviours, lifestyle characteristics, and travel trends, determine the subjective assessment of travelling and tourism services. Other researchers have noticed that travel behaviour is influenced by travel experiences and feelings [7, 8]. It is also argued that it is important to analyse human behaviour characteristics in order to understand how customers react to alternative transport policies [9]. Other travel research studies have analysed environmental factors that influence travel and tourism. Stradling and Anable [10], argue that environmental characteristics, such as workplace, shops and site topography affect travel choices.

Several approaches have been proposed for building user interests models. Kim and Chan [11], have proposed a hierarchical model for representing user interests. The user profile is constructing by analysing documents that users have visited on the web. The documents' analysis yields a list of user interests, which subsequently are grouped upon their similarity on the hierarchical interests' model. It is argued that there exist four classes of information contexts that need to be specified when attempting to understand user interests [12]. The *general information class* that refers to personal characteristics such as name, contact details, demographics of the user. The *event class* represents user's activities. The *preference class* refers to user's interests. *The social network class* explains user's connections and interactions with other users. The preference class is usually discovered by analysing various sources such as relevant documents that the user has published [12, 13].

Several representational approaches have been proposed for representing user interests. Most frequently though there are three different formats namely: keywords, semantic networks and concept-based representations [14, 15]. Keywords representing domains of interests are associated with weights indicating the strength of user interests for a particular topic. Polysemy and Synonymy are problems associated with keywords. Semantic networks, address these problems, by representing keywords with nodes that are connected with each other, including co-

occurrences. Concept-based representations resemble semantic networks in structure but they differ in having nodes to represent abstract topics rather than keywords [14, 15]. User profiles can be used in various ways such as: during personalised information retrieval, that is when a system detects relevant documents and information according to users' interests, during re-evaluating the relevance of documents taking into consideration what documents a user has retrieved and during query processing, when a user query can be modified based on user interests [16].
It is argued that filtering and clustering techniques are very useful in reducing the number of concepts that are found on the web in order to be used in formulating user profiles. However, [16], argues that these techniques lack effectiveness for they produce the same structure of interests for users with different needs. Research show that while many systems produce and use user profiles, e.g. in web personalisation, recommender systems there exists no definite procedure for deriving user interests [16–19]. This paper addresses the need for investigating alternative ways of developing user interests' models and suggests the analysis of the TF-IDF with the FsQCA.

3 Methodology

The aim of the paper is to identify the causal combinations that are necessary and sufficient to represent customer interests. This paper utilises the FsQCA in order to analyse the TF and IDF of UGC and produce causal combinations that best lead to an outcome. The FsQCA is particularly important for investigating intertwined relationships between multiple factors that affect a dependent variable or contribute to the realisation of certain outcome [20]. The FsQCA analyses the sets of relationships among causes. In FsQCA variables are modelled as sets. The FsQCA models allow a detailed analysis of how alternative conditions of causes combine and contribute to high membership scores of the outcome [21]. FsQCA may detect multiple paths, i.e. alternative causal combinations that can lead to high levels of the same outcome [20, 22]. Data in this paper is collected from customer reviews published on hotel web sites. Causal combinations may be represented by tourism services terms such as room, view, cleanliness, etc., in the set of selected documents. The outcome set in this paper, is represented by the *large amount of money spent by the customer*. Other outcome sets can also be considered. Thus, this paper aims to identify the combinations of customer hotel services interests that best reflect customer's spending. A sample of the data collected is analysed in this paper. The steps of the methodology are shown below:

1. Select documents published by user (u_i).
2. Identify the terms that will constitute the causal combinations and specify the term that will represent the outcome set.
3. Calculate the (TF) and the (IDF) for each identified term.
4. Calculate the weight of each term (t_k) using the following formula:

$$W_{tk} = TF_{tk} * \log(\frac{N_i}{d_{tk}}) \quad [23] \quad (1)$$

where, W_{tk}, represents the weight of term (t_k), TF_{tk}, is the term frequency for term (t_k), N_i, is the total number of documents published by user (u_i) and d_{tk}, represents the number of documents that contain term (t_k).

5. Apply the FsQCA and produce User Interests causal combinations.
   a. Produce the truth table of all possible permutations of the terms considered. Each permutation is a possible causal combination.
   b. Calculate membership degrees for each combination. Its calculation is performed drawing on the fuzzy sets operations theory. Assume two fuzzy sets $\tilde{A}$ and $\tilde{B}$ then:

The fuzzy union, is defined as $\mu_{(A \cup B)} = \max(\mu_A, \mu_B)$, (2)

The fuzzy intersection is defined as $\mu_{(A \cap B)} = \min(\mu_A, \mu_B)$ (3)

and the fuzzy complement is calculated as $\mu_{\neg A} 1 - \mu_A$ (4)

6. Calculate the consistency and the coverage of the solutions using formulas (2) and (3) respectively.

$$Consistency(X \ \pi \ Y) = \frac{\sum \min(X, Y)}{\sum X} \quad [24] \quad (5)$$

$$Coverage = \frac{\sum \min(X, Y)}{\sum Y} \quad [24] \quad (6)$$

where (X) is the membership degree of each causal combination and (Y) is the membership degree of the outcome set.

7. Identify best combinations, by selecting the combinations that exhibit a consistently rate above a threshold (in this paper is set at 0.8) and the highest possible coverage. Simplify solutions into the final set of causal combinations.

The final causal combinations indicate the hotel services that customers who spend large amount of money consider as the most important.

4 Data Analysis: Illustrative Example

This paper analyses reviews collected from five (5) hotel customers. Then, for simplicity reasons, five (5) terms representing hotel services are selected from the total set of terms identified in the reviews. The outcome set *large amount of money spent* (LMSp) by each user during his/her hotel stay is represented as triangular fuzzy

numbers (TFN). The membership function $f_A(x)$ of TFN $\tilde{A}(a,m,b)$ can be calculated according to the following equation [25]:

$$f_A(x) = \begin{cases} \dfrac{x-a}{m-a} & , \quad a \le x \le m, \quad m \ne a \\ \dfrac{x-b}{m-b} & , \quad m \le x \le b, \quad m \ne b \\ 0 & , \quad otherwise \end{cases} \tag{7}$$

where a, m, b are real numbers. The linguistic scales which are used and their corresponding TFNs adopted in this study are shown in table 1.

Table 1. Linguistic scales and corresponding TFNs for Large Amount of Money-Spent fuzzy sets

Linguistic scale	**Triangular fuzzy scale**	**Mean of fuzzy numbers**
Very High	(0.75, 1.00, 1.00)	1.00
High	(0.50, 0.75, 1.00)	0.75
Medium	(0.25, 0.50, 0.75)	0.50
Low	(0.00, 0.25, 0.50)	0.25
Very Low	(0.00, 0.00, 0.25)	0.00

The linguistic scales represent indicate to what extent a customer is included to the set of those who spend large amount of money during their hotel stay. The TF and IDF scores (step 3) are calculated by using the KNIME tool, for all documents published by each user (u_i). Then, the weights for each term result from using formula (1). The results are shown in Table 2.

Table 2. The term weights and the membership degree for money spent for each Customer

Large Amount Spent membership degree Outcome Set (Y)		**Terms' Weights W_{tk} based on TF-IDF for each Customer**				
	Customer	**Quietness**	**Sea View**	**Staff Friendliness**	**Cultural Activities**	**Restaurant**
0.50	1	0.30	0.50	0.40	0.70	0.70
0.70	2	0.30	0.70	0.60	0.70	0.90

0.1	3	0.10	0.30	0.20	0.60	0.50
0.7	4	0.50	0.70	0.40	0.50	0.70
0.9	5	0.30	0.70	0.60	0.70	0.70

Next the FsQCA is applied. The truth table is developed. Since there are 5 terms to consider the number of permutations is $2^5 = 32$ .Table 3 shows part of the truth table.

Table 3. The truth table (part of) show all possible permutations of the terms

Causal Combination	Quietness	Sea View	Staff Friendliness	Cultural Activities	Restaurant
1	0	0	0	0	0
2	0	0	0	0	1
3	0	0	0	1	0
4	0	0	0	1	1
5	0	0	1	0	0
6	0	0	1	0	1
7	0	0	1	1	0
8	0	0	1	1	1
9	0	1	0	0	0
10	0	1	0	0	1
11	0	1	0	1	0
12	0	1	0	1	1
13	0	1	1	0	0
14	0	1	1	0	1
15	0	1	1	1	0
16	0	1	1	1	1
17	1	0	0	0	0

The cells in the truth table take the value (1) or (0) representing true or false. Thus, permutation number 3 is read (*Quietness=false, Sea View=false, Staff Friendliness=false, Cultural Activities=true, Restaurant=false*). Next the membership degrees for all combination for each user are calculated drawing on the fuzzy sets operations theory. Table 4 shows the membership degrees for the first 17 combinations.

Table 4. Membership degrees for combinations for each customer

Causal Combination	Customer 1	Customer 2	Customer 3	Customer 4	Customer 5
1	0.3	0.1	0.4	0.3	0.3
2	0.3	0.3	0.4	0.3	0.3
3	0.3	0.1	0.5	0.3	0.3
4	0.5	0.3	0.5	0.3	0.3
5	0.3	0.1	0.2	0.3	0.3
6	0.3	0.3	0.2	0.3	0.3
7	0.3	0.1	0.2	0.3	0.3
8	0.4	0.3	0.2	0.3	0.3
9	0.3	0.1	0.3	0.3	0.3
10	0.3	0.3	0.3	0.5	0.3
11	0.3	0.1	0.3	0.3	0.3
12	0.5	0.4	0.3	0.5	0.4
13	0.3	0.1	0.2	0.3	0.3
14	0.3	0.3	0.2	0.4	0.3
15	0.3	0.1	0.2	0.3	0.3
16	0.4	0.6	0.2	0.4	0.6
17	0.3	0.1	0.1	0.3	0.3

The membership degree of combination number 3 μ_{C3}, for customer-1, see framed cell in table 4, is calculated as follows by using formulas (3) and (4):
Consider combination number 3 membership degree μ_{C3} = μ (*Quietness=false* I *Sea View=false* I *Staff Friendliness=false* I *Cultural Activities=true* I *Restaurant=false)* = μ (not (*Quietness), not (Sea View), not (Staff Friendliness), Cultural Activities, not (Restaurant)).*

The μ (Quietness=false) = μ ((1- μ (Quietness)) = (1-0.3)=0.7. Similar calculations are performed for all terms thus, μ_{C3} =min(0.7; 0.5; 0.6; 0.3)=0.3. After all membership degrees are calculated the consistency and coverage degrees are determined. Table 5 shows the results for the first 17 combinations.

Table 5. Causal combinations' Consistency and Coverage

Causal Combination	Consistency	Coverage
1	0.785714286	0.379310345
2	0.8125	0.448275862
3	0.733333333	0.379310345
4	0.789473684	0.517241379
5	0.916666667	0.379310345
6	0.928571429	0.448275862
7	0.916666667	0.379310345
8	0.933333333	0.482758621
9	0.846153846	0.379310345
10	0.882352941	0.517241379
11	0.846153846	0.379310345
12	0.904761905	0.655172414
13	0.916666667	0.379310345
14	0.933333333	0.482758621
15	0.916666667	0.379310345
16	0.954545455	0.724137931
17	1	0.379310345

The consistency for combination number 3 is calculated, by applying formula (5) as follows: Consider the outcome column (Y) shown in Table 2 and the membership degrees (X) of combination number 3, for all users as shown in Table 4. Then,

$\sum \min(X, Y) =$

min{min(0.3;0.5)+min(0.1;0.7)+min(0.5;0.1)+min(0.3;0.7)+min(0.3;0.9)=

min(0.3+0.1+0.1+0.3+0.3)=1.1. $\sum X$ =(0.3+0.1+0.5+0.3+0.3)=1.5.

Therefore the consistency for combination number 3=0.733.

Regarding the coverage, by applying formula (6), $\sum \min(X, Y) = 1.5$ and

$\sum Y = 2.9$ thus coverage=0.37.

According to FsQCA the best causal combinations should exhibit as high as possible consistency and coverage. However, the higher the consistency is the lower the coverage. Assuming a threshold value of 0.8 for the consistency firstly and then the

higher possible coverage, the analysis results into two causal combinations; the combinations number 12 and 16 extracted from Table 3, are shown in Table 6.

Table 6. The two necessary and sufficient causal combinations

Causal Combination	Quietness	Sea View	Staff Friendliness	Cultural Activities	Restaurant
16	0	1	1	1	1
12	0	1	0	1	1

A closer look at the combinations reveals that "quietness" is not within the customers interests at all. It is not a necessary service. Thus, restructuring the causal combination the analysis results that customers who spend a large amount of money, show interest in

- (Sea View) AND (Staff friendliness) AND (Cultural activities) AND (Restaurant) OR
- (Sea View) AND (Cultural activities) AND (Restaurant).

In order to simplify the causal combinations, the "staff friendliness" could be omitted for it does not appear on both combinations.

5 Conclusions-Future Research

This study suggests that the FsQCA can be used for modelling users' interests. Data selected from customer reviews is analysed by utilising the TF and the IDF. The application of the FsQCA results into useful insights that can be used to understand customer priorities and build customer profiles. Future research can focus on examining the applicability of the FsQCA to handle multiple outcome sets and to specify terms' priorities. When applying the FsQCA method in large data sets with a long list of factors, the truth table and the set of possible causal combinations can become cumbersome to analyse. Thus, future research can focus on combining the FsQCA analysis with other techniques that will be used in pruning the size of the truth table and reduce the causal combinations to manageable size.

References

1. Martínez-Garcia, E., Ferrer-Rosell, B., Coenders, G.: Profile of business and leisure travelers on low cost carriers in Europe. J. Air Transp. Manag. 20, 12–14 (2012)
2. Baka, V.: The becoming of user-generated reviews: Looking at the past to understand the future of managing reputation in the travel sector. Tour. Manag. 53, 148–162 (2016)
3. Hunecke, M., Haustein, S., Böhler, S., Grischkat, S.: Attitude-based target groups to reduce the ecological impact of daily mobility behavior. Environ. Behav. 42, 3–43 (2010)
4. Zhang, Z., Lin, H., Liu, K., Wu, D., Zhang, G., Lu, J.: A hybrid fuzzy-based personalized recommender system for telecom products/services. Inf. Sci. (Ny). 235, 117–129 (2013).

doi:10.1016/j.ins.2013.01.025
5. Wedel, M., Kamakura, W.A.: Market segmentation: Conceptual and methodological foundations. Springer Science & Business Media (2012)
6. Collantes, G.O., Mokhtarian, P.L.: Subjective assessments of personal mobility: What makes the difference between a little and a lot? Transp. Policy. 14, 181–192 (2007)
7. Handy, S., Weston, L., Mokhtarian, P.L.: Driving by choice or necessity? Transp. Res. Part A Policy Pract. 39, 183–203 (2005)
8. Sheller, M., Urry, J.: The new mobilities paradigm. Environ. Plan. A. 38, 207–226 (2006)
9. Schade, J., Schlag, B.: Acceptability of urban transport pricing strategies. Transp. Res. Part F Traffic Psychol. Behav. 6, 45–61 (2003)
10. Stradling, S.G., Anable, J.: Individual transport patterns. (2008)
11. Kim, H.R., Chan, P.K.: Learning implicit user interest hierarchy for context in personalization. In: Proceedings of the 8th international conference on Intelligent user interfaces. pp. 101–108. ACM (2003)
12. Joung, Y., El Zarki, M., Jain, R.: A user model for personalization services. In: Digital Information Management, 2009. ICDIM 2009. Fourth International Conference on. pp. 1–6. IEEE (2009)
13. Bakalov, F., König-Ries, B., Nauerz, A., Welsch, M.: A Hybrid Approach to Identifying User Interests in Web Portals. In: IICS. pp. 123–134 (2009)
14. Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A.: User profiles for personalized information access. In: The adaptive web. pp. 54–89. Springer (2007)
15. Michlmayr, E., Cayzer, S.: Learning user profiles from tagging data and leveraging them for personal (ized) information access. (2007)
16. Saleheen, S., Lai, W.: UIWGViz: An architecture of user interest-based web graph vizualization. J. Vis. Lang. Comput. 44, 39–57 (2018)
17. Magnini, B., Strapparava, C.: Improving user modelling with content-based techniques. In: International Conference on User Modeling. pp. 74–83. Springer (2001)
18. Lehmann, S., Schwanecke, U., Dörner, R.: Interactive visualization for opportunistic exploration of large document collections. Inf. Syst. 35, 260–269 (2010)
19. Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. Icwsm. 8, 361–362 (2009)
20. Chari, S., Tarkiainen, A., Salojärvi, H.: Alternative pathways to utilizing customer knowledge: A fuzzy-set qualitative comparative analysis. J. Bus. Res. 69, 5494–5499 (2016)
21. Rihoux, B., Ragin, C.C.: Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques. Sage Publications (2008)
22. Skarmeas, D., Leonidou, C.N., Saridakis, C.: Examining the role of CSR skepticism using fuzzy-set qualitative comparative analysis. J. Bus. Res. 67, 1796–1805 (2014)
23. Chen, K., Zhang, Z., Long, J., Zhang, H.: Turning from TF-IDF to TF-IGM for term weighting in text classification. Expert Syst. Appl. 66, 245–260 (2016)
24. Korjani, M.M., Mendel, J.M.: Fuzzy set qualitative comparative analysis (fsQCA): Challenges and applications. In: Fuzzy Information Processing Society (NAFIPS), 2012 Annual Meeting of the North American. pp. 1–6. IEEE (2012)
25. Lin, H.-Y., Hsu, P.-Y., Sheen, G.-J.: A fuzzy-based decision-making procedure for data warehouse system selection. Expert Syst. Appl. 32, 939–953 (2007)