Fuzzy equivalence relation based clustering and its use to restructuring websites' hyperlinks and web pages

Dimitris K. Kardaras[1,*], Xenia J. Mamakou[1], Bill Karakostas[2]

[1]Business Informatics Laboratory, Dept. of Business Administration, Athens University of Economics and Business, 76 Patission Street, Athens 10434, Greece
{kardaras,xenia}@aueb.gr
[2] Centre for HCI Design, School of Informatics, City University, Northampton Sq., London EC1V 0HB, UK
billk@soi.city.ac.uk

Abstract. Quality design of websites implies that among other factors, hypelinks' structure should allow the users to reach the information they seek with the minimum number of clicks. This paper utilises the fuzzy equivalence relation based clustering in adapting website hyperlinks' structure so that the redesigned website allows users to meet as effectively as possible their informational and navigational requirements. The fuzzy tolerance relation is calculated based on the usage rate of hyperlinks in a website. The equivalence relation identifies clusters of hyperlinks. The clusters are then used to realocate hyperlinks in webpages and to rearrange webpages into the website structure hierarchy.

Keywords: fuzzy equivalence relation, web adaptation, hyperlinks' clustering

1 Introduction

When designing a website, the way that its content is organised and how efficiently users get access to it, influence the user perceived design quality. The designers' goal is an effective and plain communication of content [2]. Website structure has been identified by many reaserch studies as an important factor that affects web design quality. The users' perception of how different parts of a web site are linked together is a strong indicator of effective design [4], [13]. Thus, the websites' hyperlinks structure should adapt to meet users' changing requirements and priorities depending e.g. on the expertise of users in navigating a website, their familiarity with its content structure, their information needs, etc. Furthermore, hyperlinks structure has been extensivley used in web search engines and web mining [12]. The number of links pointing to a web page is considered as a quality indicator that reflects the authority of the web page the links point at [1]. Many algorithms have been developed to tackle one of the greatest challenges for web design and search engines, i.e. how to specify an appropriate website structure, or how to evaluate webpages quality [8]. This paper suggests the use of fuzzy equivalence relation clustering to restucturing webpages

throught hypelinks popularity. Until recently, the web users' browsing behaviour was often overlooked in approches that attempted to manage websites' structures or determining the quality of webpages [9]. This paper considers the popularity of hypelinks as an indicator of users' browsing behaviour and classifies links into pages and subsequently pages are allocated to different website levels. An illustrative example is provide to expemlify the proposed approach.

2 Hyperlink Analysis

Although hyperlink analysis algorithms that produced significant results have been developed, they still need to tackle challenges such as how to incorporate web user behaviour in hyperlink analysis [8]. A website is considered as a graph of nodes and edges representing webpages and hyperlinks respectively. Based on the hyperlinks analysis, the importance of each node can be estimated thus leading to a website structure that reflects the relative importance of each hyperlink and each web page [7]. Two of the most representative links analysis algorithms are the HITS [5] and the Google's PageRank [1], which assume that a user randomly selects a hyperlink and then they calculate the probabilities of selecting other hyperlinks and webpages. Many other link analysis algorithms stem from these two algorithms. The basic idea behind link analysis algorithms is that if a link points to page (i) from page (j), then it is assumed that there is a semantic relation between the two pages. However, the links may not represent users' browsing behaviour, which is driven by their interests and information needs. Hyperlinks are not clicked by users with equal probabilities and they should not be treated as equally important [9]. Thus, webpages that are visited and hyperlinks that are clicked by users should be regarded as more important than those that are not, even if they belong to the same web page. It is therefore reasonable to use users' preferences to redesign the hyperlink graph of a website [8]. Google Toolbar and Live Toolbar collect user browsing information. User browsing preferences is an important source of feedback on page relevance and importance and is widely adopted in website usability [3], [20], user intent understanding [22] and Web search [9] research studies. By utilising the user browsing behaviour, website structures can be revised by deleting unvisited web pages and hyperlinks and relocating web pages and hyperlinks according to their importance as perceived by the users, i.e. as the number of clicks show. Liu et al. (2008) developed a "user browsing graph" with Web log data [10]. A website representing graph as derived from user browsing information can lead to a website structure closer to users' needs, because links in the website graph are actually chosen and clicked by users. Liu et al. (2008) also proposed an algorithm to estimate page quality, BrowseRank, which is based on continuous-time Markov process model, which according to their study performs better than PageRank and TrustRank.

3 Fuzzy relations and fuzzy classification

3.1 Fuzzy relations

Fuzzy relations are important for they can describe the strength of interactions between variables [11]. Fuzzy relations are fuzzy subsets of $X \times Y$, that is mapping from $X \rightarrow Y$. Let $X, Y \subseteq R$ be universal sets. Then

$$\widetilde{R} = \{((x, y), \mu_R(x, y)) \mid (x, y) \in X \times Y\} \quad (1)$$

is called a fuzzy relation on $X \times Y$ [23].

3.2 Fuzzy Classification with equivalence relation

Numerous classification methods have been proposed so far including cluster analysis, factor analysis, discriminant analysis [16], k-means analysis [14], c-means clustering [21]. According to Ross (2010) there are two popular methods of classification namely the classification using equivalent relations and the fuzzy c-means. Cluster analysis, factor analysis and discriminant analysis are usually applied in classic statistical problem where large sample or long term data is available. When dealing with small data k-means or c-means methods are preferred [18]. In this paper, we use fuzzy equivalent relations and lambda-cuts (λ-cuts) to classify links, according to their importance in a website. Classification based on λ-cuts of equivalent relations is used in many recent studies [6], [17] and [19]. An important feature of this approach is that for its application it is not required to assume that the number of clusters is known as it is required by other methods such as in the case of k-means and c-means clustering [18].

A fuzzy relation on a single universe X is also a relation from X to X. It is a fuzzy tolerance relation if the two following properties define it:

Reflexivity: $\mu_R(x_i, x_i) = 1$

Symmetry: $\mu_R(x_i, x_j) = \mu_R(x_j, x_i)$

Moreover, it is a fuzzy equivalence relation if it is a tolerance relation and has the following property as well:

Transitivity: $\mu_R(x_i, x_j) = \lambda_1$ and $\mu_R(x_i, x_k) = \lambda_2 \rightarrow \mu_R(x_i, x_k) = \lambda$, where $\lambda \geq \min[\lambda_1, \lambda_2]$.

Any fuzzy tolerance relation can be reformed into a fuzzy equivalence relation by at most (n – 1) compositions with itself. That is:

$$\widetilde{R}_1^{n-1} = \widetilde{R}_1 \circ \widetilde{R}_1 \circ \ldots \circ \widetilde{R}_1 = \widetilde{R} \quad (2)$$

In fuzzy equivalent relations, their λ-cuts are equivalent ordinary relations.

The numerical values that characterize a fuzzy relation can be developed by a number of ways, one of which is similarity[15]. Min-max method is one of the simi-

larity methods that can be used when attempting to determine some sort of similar pattern or structure in data through various metrics. The equation of this method is given below:

$$r_{i,j} = \frac{\sum_{k=1}^{m} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{m} \max(x_{ik}, x_{jk})} \quad , \tag{3}$$

where i, j = 1,2,…,n

In this paper, we use min-max similarity method due to its simplicity and common use [15].

4 Proposed methodology

Assume a website (WS) consists of a number (p) of web pages (wp) such as WS={wp}. A website is regarded as a set of partially ordered web pages, according to their popularity in terms of users' preferences, i.e. visits. Thus, assuming a website has three levels of web pages, the web pages are allocated to a website level according to the demand users show for the web page.

The proposed methodology consists of the following steps:

1. Identify the links of each web page in a website.
2. In order to capture the users' browsing behaviour, first measure the clicks made for each link (l_i), then compute and normalise their demand using the following type:

$$Dl_i = \frac{Cl_i}{\sum_{i=1}^{n} Cl_i}, \tag{4}$$

where Dl_i is the demand for link i, i=1,2,…,n and Cl_i are the number of clicks for the link i. $Dl_i \in [0,1]$.

3. Fuzzify the Dl_i using triangular fuzzy numbers (TFNs) and specify their corresponding linguistic variables.
4. Calculate the membership degree to the corresponding linguistic variables for each link, using membership functions. The membership function of a TFN is given by the following formula:

$$\mu_A(x) = \begin{cases} 0, & x \le a \\ (x-a)/(b-a), & a \le x \le b \, and \, a < b \\ (c-x)/(c-b), & b \le x \le c \, and \, b < c \\ 0, & x \ge c \end{cases} \quad (5)$$

5. Calculate the fuzzy tolerance relation ($\tilde{R}_t$), using min-max similarity method, as in Eq. (3).
6. Calculate the fuzzy equivalent relation using Eq. (2).
7. Decide on the λ-cuts to be used.
8. Classify web page links according to λ-cuts. The derived clusters constitute the redesigned web pages.
9. Calculate the new demand for each of the newly formed web pages, according to the type:

$$Dwp_j = \sum_{i=1}^{n} Cl_i \, , \quad (6)$$

where Dwp_j is the demand for web page j, with (i) and (n) indicating the hyperlinks and the number of hyperlinks in the webpage respectively. The demand for each web page is calculated according to the total number of the clicks made to all links in this specific page.

10. Normalize the demand for the new web pages and then fuzzify the Dwp_j. Since a website is a partially ordered set of web pages, the newly formed we pages are allocated to a level a website hierarchy level according to their demand. The higher the demand the higher the level they are assigned to.
11. Calculate the validation index as shown in Eq. (7) for different λ-cuts. The λ-cut value that maximises the validation index indicated the optimum number of clusters.

$$\lambda - C(\lambda)/m \quad (7)$$

The C(λ) indicates the number of clusters and the (m) shows the number of data sequence that are subject to clustering.

5 Illustrative example

The following example illustrates the proposed methodology. Let us consider of a website that originally has a total of 10 links (i=10) in all of its 5 web pages as shown in Fig 1.

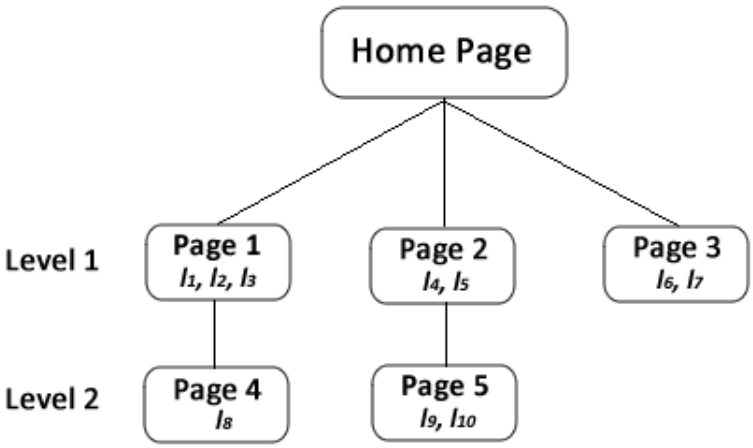

Fig. 1. Original Website with five web pages having links

By the use of cookies we can identify the number of clicks each of these links has had in a specific time period. Then according to Eq. (4) the demand of each link can be calculated. In our example, let's say that the Dl_is are: {0.31, 0.68, 0.45, 0.25, 0.76, 0.59, 0.88, 0.39, 0.77, 0.25}.

To fuzzify the Dl_i we will use three TFNs with their corresponding linguistic variables, which are Low = (0, 0.3, 0.5), Medium=(0.3, 0.5, 0.7) and High=(0.5, 0.7, 1). Then, we calculate the membership degree of each link to the corresponding linguistic variable using Eq. (5), as seen in Table 1.

Table 1. Hyperlinks membership degrees to each of the corresponding linguistic variables

	l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	l_9	l_{10}
Low	0.95	0.00	0.25	0.83	0.00	0.00	0.00	0.55	0.00	0.83
Medium	0.05	0.10	0.75	0.00	0.00	0.55	0.00	0.45	0.00	0.00
High	0.00	0.90	0.00	0.00	0.80	0.45	0.40	0.00	0.77	0.00

Using Eq. (3) we form the fuzzy tolerance relation:

$$\tilde{R}_t = \begin{array}{c} \\ l_1 \\ l_2 \\ l_3 \\ l_4 \\ l_5 \\ l_6 \\ l_7 \\ l_8 \\ l_9 \\ l_{10} \end{array} \begin{array}{c} \begin{array}{cccccccccc} l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & l_{10} \end{array} \\ \begin{bmatrix} 1.00 & 0.03 & 0.18 & 0.83 & 0.00 & 0.03 & 0.00 & 0.43 & 0.00 & 0.83 \\ 0.03 & 1.00 & 0.05 & 0.00 & 0.80 & 0.38 & 0.40 & 0.05 & 0.77 & 0.00 \\ 0.18 & 0.05 & 1.00 & 0.16 & 0.00 & 0.38 & 0.00 & 0.54 & 0.00 & 0.16 \\ 0.83 & 0.00 & 0.16 & 1.00 & 0.00 & 0.00 & 0.00 & 0.43 & 0.00 & 1.00 \\ 0.00 & 0.80 & 0.00 & 0.00 & 1.00 & 0.33 & 0.50 & 0.00 & 0.96 & 0.00 \\ 0.03 & 0.38 & 0.38 & 0.00 & 0.33 & 1.00 & 0.40 & 0.29 & 0.34 & 0.00 \\ 0.00 & 0.40 & 0.00 & 0.00 & 0.50 & 0.40 & 1.00 & 0.00 & 0.52 & 0.00 \\ 0.43 & 0.05 & 0.54 & 0.43 & 0.00 & 0.29 & 0.00 & 1.00 & 0.00 & 0.43 \\ 0.00 & 0.77 & 0.00 & 0.00 & 0.96 & 0.34 & 0.52 & 0.00 & 1.00 & 0.00 \\ 0.83 & 0.00 & 0.16 & 1.00 & 0.00 & 0.00 & 0.00 & 0.43 & 0.00 & 1.00 \end{bmatrix} \end{array} \quad (8)$$

To produce the fuzzy equivalent relation ($\tilde{R}_e$) that will be used to classify the links of the website, we use Eq. (2):

$$\tilde{R}_e = \tilde{R}_t^9 = \begin{bmatrix} 1.00 & 0.38 & 0.43 & 0.83 & 0.38 & 0.38 & 0.38 & 0.43 & 0.38 & 0.83 \\ 0.38 & 1.00 & 0.38 & 0.38 & 0.80 & 0.40 & 0.52 & 0.38 & 0.80 & 0.38 \\ 0.43 & 0.38 & 1.00 & 0.43 & 0.38 & 0.38 & 0.38 & 0.54 & 0.38 & 0.43 \\ 0.83 & 0.38 & 0.43 & 1.00 & 0.38 & 0.38 & 0.38 & 0.43 & 0.38 & 1.00 \\ 0.38 & 0.80 & 0.38 & 0.38 & 1.00 & 0.40 & 0.52 & 0.38 & 0.96 & 0.38 \\ 0.38 & 0.40 & 0.38 & 0.38 & 0.40 & 1.00 & 0.40 & 0.38 & 0.40 & 0.38 \\ 0.38 & 0.52 & 0.38 & 0.38 & 0.52 & 0.40 & 1.00 & 0.38 & 0.52 & 0.38 \\ 0.43 & 0.38 & 0.54 & 0.43 & 0.38 & 0.38 & 0.38 & 1.00 & 0.38 & 0.43 \\ 0.38 & 0.80 & 0.38 & 0.38 & 0.96 & 0.40 & 0.52 & 0.38 & 1.00 & 0.38 \\ 0.83 & 0.38 & 0.43 & 1.00 & 0.38 & 0.38 & 0.38 & 0.43 & 0.38 & 1.00 \end{bmatrix} \quad (9)$$

The next step is to decide on the λ-cut. Assuming that the λ-cut that maximises the validation index shown in formula (7) is 0.5 we have:

$$\tilde{R}_{0.5} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (10)$$

Eq. (10) is then used to classify each of the 10 links in our example. In this case, we find four classes, each of which includes specific links: $class_1=\{l_1,l_4,l_{10}\}$, $class_2=\{l_2,l_5,l_7,l_9\}$, $class_3=\{l_3,l_8\}$ and $class_4=\{l_6\}$. We then re-create the website, having this time only four (out of the original five) web pages. The next step is to calculate the new demand for each of the four newly created web pages, using Eq. (6). In our case we find: $Dwp_1=395$, $Dwp_2=44$, $Dwp_3=442$ and $Dwp_4=62$. Then we normalize each of the Dwp_i and we get $Dwp_1=395/943=0.42$, $Dwp_2=44/943=0.05$, $Dwp_3=442/943=0.47$ and $Dwp_4=62/943=0.07$.

To fuzzify the Dwp_i we use the same three TFNs we used when fuzzifying the Dl_i. That is Low = (0, 0.3, 0.5), Medium = (0.3, 0.5, 0.7) and High = (0.5, 0.7, 1). By calculating the membership functions for each Dwp_i to each TFN, we find that $\mu_{Low}(Dwp_1) = 0.4$, $\mu_{Medium}(Dwp_1) = 0.6$, $\mu_{High}(Dwp_1) = 0$, $\mu_{Low}(Dwp_2) = 0.17$, $\mu_{Medium}(Dwp_2) = 0$, $\mu_{High}(Dwp_2) = 0$, $\mu_{Low}(Dwp_3) = 0.15$, $\mu_{Medium}(Dwp_3) = 0.85$, $\mu_{High}(Dwp_3) = 0$, $\mu_{Low}(Dwp_4) = 0.23$, $\mu_{Medium}(Dwp_4) = 0$ and $\mu_{High}(Dwp_4) = 0$. We can then easily understand that wp_1 and wp_3 belong to the "Medium" category, whereas wp_2 and wp_4 belong to the "Low" category. To decide on the level of each web page to the website, we take all web pages found in the "High" category and put them in Level 1, then the web pages found in the "Medium" category are organized in Level 2 and web pages that belong to the "Low" category are left in Level 3. In case a category does not exist we put in the specific Level, the web pages that are found in the next category. In our example, no web page is found in the "High" category, so Level 1 will include the pages of the "Medium" category, which are wp_1 and wp_3. Wp_2 and wp_4 are then put in Level 2 as shown in Fig. 2.

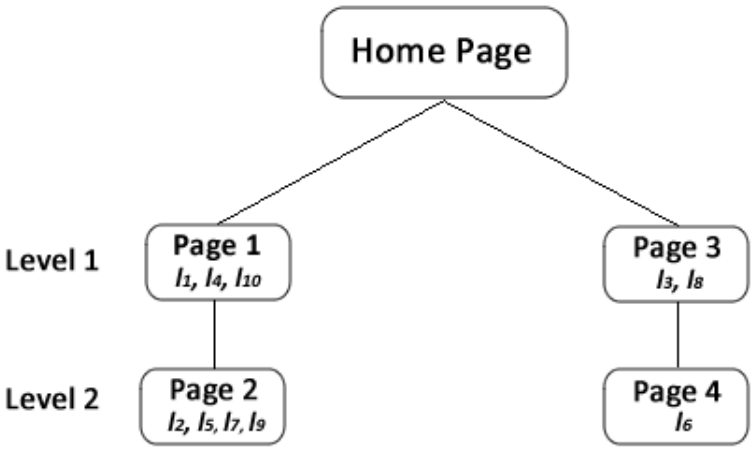

Fig. 2. Restructured website with three web pages having links in two levels

6 Conclusions

This paper suggests an approach to websites structuring. A fuzzy equivalence relation based clustering is been adopted, for it does not assumed a known number of clusters as other clustering techniques do. Further it is a clustering technique that has been recently used in other domains with satisfactory results. In order to derive an appropriate web structure, this paper considers the data that reflect the users' browsing behaviour. Research studies claim that it is important to take into account users' browsing information in determining websites structure and webpages' quality. This paper suggests that clustering of hyperlinks' usage data can be used in order to cluster links of similar popularity into the same web page and then similar demand web page to be grouped into the same web page hierarchy level. An illustrative example has shown the applicability of the proposed approach.

7 References

1. Brin, S. and Page, L. The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems. 30(1/7), 107-17. (1998)
2. De Marsico, M. & Levialdi, S. Evaluating web sites: exploiting user's expectations. Int. J. Human Computer Interactions, 60(3), 381-416 (2004)
3. Fang, X. and Holsapple, C.W. An empirical study of web site navigation structures' impacts on web site usability. Decision Support Systems 43(2), 476–491 (2007)
4. Henzinger, M., Motwani, R. and Silverstein, C. Challenges in web search engines, paper presented at the SIGIR Forum. (2002)
5. Kleinberg, J. M. Authoritative sources in a hyperlinked environment. In Proceedings of the 9th ACM SIAM Symposium on Discrete Algorithms, Baltimore, MD, USA, 668–677 (1998)
6. Liang, G.,S., Chou, T.,Y., and Han, T.,C. Cluster analysis based on equivalence relation. European Journal of Operational Research, 166, 160-171 (2005)
7. Lin, S.H., Chu, K.P., and Chiu, C.M. Automatic sitemaps generation: Exploring website structures using block extraction and hyperlink analysis. Expert Systems with Applications, 38, 3944-3958 (2011)
8. Liu, Y., Xue, Y., Xu, D., Cen, R., Zhang, M., Ma, S., & Ru, L. Constructing a reliable Web graph with information on browsing behaviour. Decision Support Systems, 54, 390-401 (2012)
9. Liu, Y., Chen, F., Kong, W., Yu, H., Zhang, M., Ma, S., Ru, L. Identifying Web spam with the wisdom of the crowds. ACM Transaction on the Web 6 (1), 1-30 (2012)
10. Liu, Y., Gao, B., Liu, T., Zhang, Y., Ma, Z., He, S., Li, H. BrowseRank: letting web users vote for page importance. In: Proc. of 31st ACM SIGIR Conference, 451–458 (2008)
11. Luukka, P.: Similarity classifier using similarities based on modified probabilistic equivalence relations. Knowledge-Based Systems, 22(1), 57-62 (2009)
12. Mandl, T. The impact of web site structure on link analysis. Internet Research, 17(2), 196-206 (2007)
13. Muylle,S, Moenaertb, R., & Despontin, M. The conceptualization and empirical validation of web site user satisfaction. Information & Management, 41(5), 541-560 (2004)
14. Ralambondrainy, H, A conceptual version of K-means algorithm. Pattern Recognition Letters, 16, 1147-1157 (1995)

15. Ross, T. J.: Fuzzy logic with engineering applications. Wiley-Blackwell, 3rd Edition (2010)
16. Sharma, S. Applied multivariate techniques. JohnWiley (1996)
17. Sun, P.G., Gao, L., Han, S.,S. Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks. Information Sciences, 181, 1060-1071 (2011)
18. Wang, Y.L. and Lee, H.S. A clustering method to identify representative financial ratios. Information Sciences, 178, 1087-1097 (2008)
19. Wang, Y.L. A clustering method based on fuzzy equivalence relation for customer relationship management. Expert Systems with Application, 37, 6421-6429 (2010)
20. Wu, H. Gordon, M. DeMaagd, K. Fan, W. Mining web navigations for intelligence. Decision Support Systems 41(3), 574–591 (2006)
21. Wu, K.L. and Yang,M.S. Alternative c-means clustering algorithms. Pattern Recognition 35, 2267-2278 (2002)
22. Xu, D., Liu, Y., Zhang, M., Ru, L., Ma, S. Predicting Epidemic Tendency through Search Behavior Analysis. In Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI-11) (Barcelona, Spain), 2361–2366 (2011)
23. Zimmermann, H. J.: Fuzzy set theory-and its applications. Springer (2001)