TOWARDS UNIVERSAL RATING OF ONLINE MULTIMEDIA CONTENT

Lawrence Nderu[1], Nicolas Jouandeau[2], and Herman Akdag[2]

[1]Jomo Kenyatta University of Agriculture and Technology, Kenya
[1]nderu@jkuat.ac.ke
[2]University of Paris 8 –LIASD, France
{n, akdag}@ai.univ-paris8.fr

ABSTRACT

Most website classification systems have dealt with the question of classifying websites based on their content, design, usability, layout and such, few have considered website classification based on users' experience. The growth of online marketing and advertisement has lead to fierce competition that has resulted in some websites using disguise ways so as to attract users. This may result in cases where a user visits a website and does not get the promised results. The results are a waste of time, energy and sometimes even money for users. In this context, we design an experiment that uses fuzzy linguistic model and data mining techniques to capture users' experiences, we then use the k-means clustering algorithm to cluster websites based on a set of feature vectors from the users' perspective. The content unity is defined as the distance between the real content and its keywords. We demonstrate the use of bisecting k-means algorithm for this task and demonstrate that the method can incrementally learn from user's profile on their experience with these websites.

KEYWORDS

Website Classification, Fuzzy Linguistic Modeling, K-Means Clustering, Web Mining.

1. INTRODUCTION

The Internet has become a major source of information. Individuals and organizations depend on the Internet in one way or another. It can be asserted that the web is the largest available repository of data with the largest number of users [1]. Therefore, the web can be viewed as a meeting point of providers of information and services and their consumers.

Since its early days the Internet has seen remarkable growth. This growth is fueled by millions who provide high quality, trustworthy content. However, in this favorable landscape a good number of providers may seek to profit by promising users resources which eventually they cannot or do not provide. This leads to cases whereby a user might end up wasting time, energy and even sometimes money. This situation may also create a disillusioned user. Our goal is to develop a website clustering system that takes into consideration the previous users' experiences.

The paper is set out as follows: First, we introduce some literature on web mining and fuzzy linguistic modeling. Secondly, an experiment is proposed to compute the agreement between what a web site claims it provides and what in-fact it provides based on users browsing patterns. Third, a discussion is provided to demonstrate the feasibility and effectiveness of the proposed model.

David C. Wyld et al. (Eds) : CST, ITCS, JSE, SIP, ARIA, DMS - 2014
pp. 305–310, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4128

Finally, some conclusions are presented at the end of this paper.

2. WEB MINING AND FUZZY LINGUISTIC MODELING

2.1 Web Mining

Web mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web. This includes the automatic search of information resources available on-line, i.e. Web content, data mining and discovery of users access patterns from Web servers [2].

As information technology grows, users can aggregate and store mass data more easily and efficiently [3]. Data mining can help users analyze and retrieve valuable information from mass data sources available online [4]. Due to the existence of a variety of websites that claim to provide the information that the users are looking for, users find themselves dealing with the question of identification of sources that deliver what they claim. Web mining can be used to analyze user's browsing behavior and provide suitable information for future users of the same websites. Important data can be obtained from Web servers or proxy servers such as log files, user profiles, registration data and user sessions or transactions [5]. From these data we can discover valuable information about the websites visited.

2.2 Fuzzy Linguistic Modeling

Fuzzy linguistic modelling [6] is a very useful kind of fuzzy linguistic approach used for modeling the computing with words process as well as linguistic aspects [7]. When collecting details from the user's on-line activities some aspects are qualitative while others are quantitative. Fuzzy linguistic modelling has been widely used and has provided very good results, it deals with qualitative aspects that are presented in qualitative terms by means of linguistic variables [8], [9], [10], [1]. The 2-tuple Fuzzy Linguistic representation model provides the following advantages over classical models [1].

1) The linguistic domain can be treated as continuous, while in the classical models it is treated as discrete.
2) The linguistic computational model based on linguistic 2-tuples carries out processes of computing with words easily and without loss of information.
3) The results of the process of computing with words are always expressed in the initial linguistic domain.

The model used in this paper to evaluate the conformance of what the web site says and what it actually delivers uses a set of quality criteria related to the Web sites and a computation instrument of quality assessments. We assume that the quality of a web site is measured through users perceptions on the services offered through its Web site [11] and that users have an objective to achieve when they decide to visit a certain website.

Users are invited to fill in a survey built on a set of quality criteria. To measure quality, conventional measurement tools used by the customers are devised on cardinal or ordinal scales. However, the scores do not necessarily represent user preferences. This is because respondents have to internally convert preference to scores and the conversion may introduce distortion of the preference [3]. For this reason, we use fuzzy linguistic modeling to represent the user's perceptions and tools of computing with words to compute the quality assessments [7]. The subjectivity and vagueness in the assessment process is dealt with using the fuzzy logic [12]. Multiple raters are often preferred rather than a single rater to avoid the bias and to minimize the partiality in the decision process. Figure 1 shows an example of the way the Fuzzy computation with words was used in the experiment in one of the question to the users. Does the website provide accurate information? The choices were None-N, Very Little- VL, Little- L, Medium- M, High- H, Very High- VH and Perfect- P.

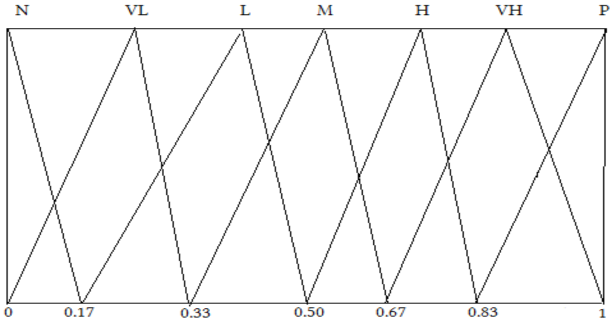

Figure 1. Linguistic variable “Accurate”.

3. Clustering Fuzzy Web Access Patterns Based on K-Means

By making use of the log files details and analysis of the results obtained from the online questioners, it is possible to adapt the paradigm of clustering. The key effectiveness of the clusters is an intuitive distance function [13]. Since the content of each single page p ε W can be represented by a feature vector of term frequencies, the whole details measurable properties as seen from the users is represented by set of feature vectors. One assumption made by this approach is that sites that fulfills a user needs as seen from the users pattern mining of related content but varying size will become very similar with respect to the Sum of Minimum Distances (SMD) [14]. Let U_1, U_2 be two users and let f: P → N^d be a feature transformation that returns the feature vector of a users' profile p ε P where P is the set of all profiles. The concept of SMD is discussed by Hans et al in [15]. Using these details we create S_1, S_2…S_3 as centroids that represent the various log features. Figure 2 shows the modeling of the clusters based on the fact that websites are different, based on the fact that users have different expectations when visiting the websites.

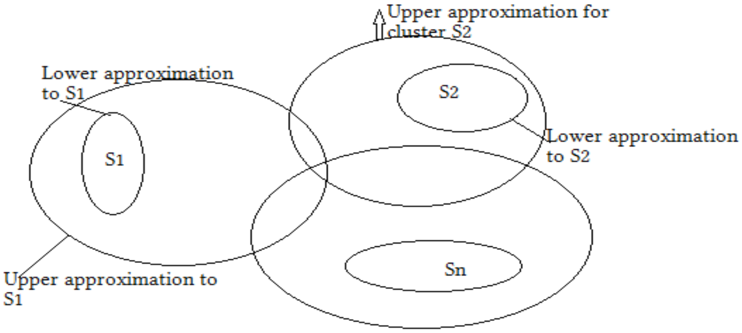
Upper approximation for cluster S2
S2
Lower approximation to S1
S1
Lower approximation to S2
Sn
Upper approximation to S1

Figure 2. The Clustering Model.

4. EXPERIMENTAL SETUP

An experiment was performed with the aim of collecting data from two sources, log analysis and online questionnaires. Students were used as experimental subjects. It is a known fact that this could lead to results that are artificial, particularly when students are asked to perform tasks which they have little or no experience. After identifying this as a potential problem, a number of websites were selected. The selection criterion was that a specific objective needed to be achieved by the website users. The time users took on this websites was then read from the freely provided Google API. A total of 199 randomly selected samples were collected, this included the time taken in a website and the time it would take to carry out the specific objective on the website.

Several methodologies have been suggested for rating websites [17], [2], since our objective was to cluster websites based on what it promises, a new criteria needed to be developed. The criteria developed identified the following as the important features with respect to our objective: accuracy, believable, relevant, details, value, revisit, and deliverability. These factors were obtained from the literature. Table 1, shows the online questionnaire with results from one rater of the websites. In this example on Table 1, the user is rating a website that he has already visited. The questionnaires were administered for ten visited URLs for which already the time taken by users while visiting this websites had already been noted. The fact that we had 199 log details for users and only selected 10 URLs to evaluate was dictated by a number of factors, the major one being the evaluation that we had carried out about the nature of the websites that we were to use in the study, but to introduce randomness this information was not available to the users. Table 2 shows the analyzed results for the ten URLs. The arithmetic mean for the 2-tuple linguistic aggregation operators was used for calculation.

Table 1. Website Online Questionnaire Example.

		None	Very Little	Little	Medium	High	Very High	Perfect
1.	Accurate information							X
2.	Believable information						X	
3.	Relevant information							X
4.	The right detail of information							X
5.	Value for your time			X				
6.	Would you visit it again				X			
7.	Does it deliver							X

Table 2. Analyzed data from the questionnaires.

URL	**Accuracy**	**Believable**	**Relevant**	**Details**	**Value**	**Revisit**	**Deliverability**
1	8.72	8.33	8.88	7.89	8.33	8.55	8.49
2	2.39	2.78	3.11	1.45	1.40	1.62	2.17
3	8.00	8.44	8.22	8.44	8.22	7.88	8.33
4	2.72	2.78	3.17	3.67	3.50	4.00	2.33
5	8.27	7.44	7.22	7.22	7.22	8.11	8.16
6	3.22	2.95	4.56	2.89	2.33	3.33	3.83
7	8.55	8.49	6.72	8.16	7.50	9.32	7.33
8	6.78	8.11	8.33	7.22	8.11	7.77	7.22
9	3.50	3.67	2.67	4.33	2.83	2.83	3.67
10	2.89	4.06	2.72	3.50	3.56	3.17	4.17

5. Results And Analysis

The Bisecting K-means algorithm was used. Table 3, shows the results obtained from the clustering program and Figure 3, shows the clusters for the ten URLs. The websites used for this experiment URL1, URL2…, URL10 were well analyzed with respect to what the websites promised to deliver. Any new website that promises to deliver what these websites were delivering can now be clustered with respect to the created clusters.

Table 3. Clustering Using k-Means.

Cluster	Data Unit	Sum	Time(Seconds)
1	0	59.2	5.3
	2	57.5	5.2
	4	54.1	5.5
	6	56.1	6.0
	7	53.5	5.6
2	1	14.9	2.7
	3	22.2	3.0
	5	23.1	3.1
	8	23.5	2.9
	9	24.1	3.6
Clustered Data: K=2			

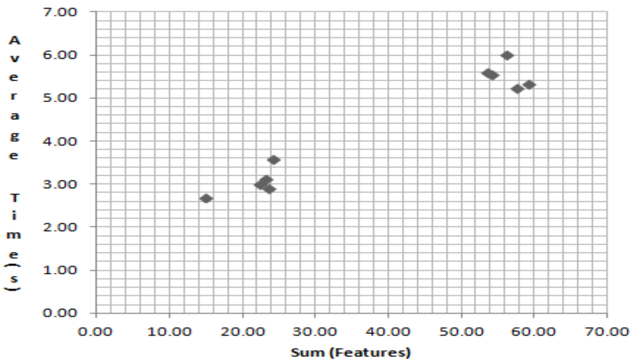

Figure 3. Clusters for the ten URLs.

6. Conclusion

In this paper, we proposed a new solution to automatic classification of websites based on the level of satisfaction of the previous users'. The experiment shows that clustering can be used as a way of telling how far a website is from the ideal users' website. The results so far point to an interesting direction in the sense that previous users' experience can be used to rank website and provide a metric for classification of websites. Combinations of these results are key point to developing future websites classification system.

REFERENCES

[1] E. Herrera-viedma, A. G. Lopez-herrera, and C. Porcel, “Evaluating the Information Quality of Web Sites : A Methodology Based on Fuzzy Computing With Words,” vol. 57, no. 4, pp. 538–549, 2006.

[2] R. Cooley, B. Mobasher, and J. Srivastava, “Web Mining : Information and Pattern Discovery on the World Wide Web * 1 Introduction,” pp. 558–567, 1997.

[3] C. Chen, W. Tai, and C. Chu, “A preference perception system on website by combining fuzzy set with data mining technology,” Int. J. Inf. …, vol. 19, no. 1, pp. 93–105, 2008.

[4] U. M. Fayyad, “Data Mining and Knowledge Applications in Astronomy Discovery in Databases : Science and Planetary,” pp. 1590–1592, 1996.

[5] R. Kosala, B.- Heverlee, and H. Blockeel, “Web Mining Research : A Survey,” vol. 2, no. 1, 2000.

[6] F. Herrera, E. Herrera-Viedma, and J. Verdegay, “Direct approach processes in group decision making using linguistic OWA operators,” Fuzzy Sets Syst., no. ii, 1996.

[7] Y.-J. Xu and Z.-J. Cai, “Method Based on Fuzzy Linguistic Judgement Matrix and Trapezoidal Fuzzy Induced Ordered Weighted Geometric (TFIOWG) Operator for Multi-Attribute Decision-Making Problems,” 2007 Int. Conf. Wirel. Commun. Netw. Mob. Comput., no. 1, pp. 5752–5755, Sep. 2007.

[8] L. Feng and T. S. Dillon, “to Provide Explanatory Semantics for Data Warehouses,” vol. 15, no. 1, pp. 86–102, 2003.

[9] M. Decision-making, F. Herrera, and L. Martínez, “A Model Based on Linguistic 2-Tuples for Dealing with Multigranular Hierarchical Linguistic Contexts,” vol. 31, no. 2, pp. 227–234, 2001.

[10] G. Bordogna and G. Pasi, “A Fuzzy Linguistic Approach Generalizing Boolean Information Retrieval : A Model and Its Evaluation,” vol. 44, no. 2, pp. 70–82, 1993.

[11] L. Hidalgo and F. J. C. J. L. G. E. Herrera-viedma, “Applying Fuzzy Linguistic Tools to Evaluate the Quality of Airline Web Sites,” 2007.

[12] L. a. Zadeh, “The concept of a linguistic variable and its application to approximate reasoning—I,” Inf. Sci. (Ny)., vol. 8, no. 3, pp. 199–249, Jan. 1975.

[13] W. Wang and Y. Zhang, “On fuzzy cluster validity indices,” Fuzzy Sets Syst., vol. 158, no. 19, pp. 2095–2117, Oct. 2007.

[14] Y. Sharon, J. Wright, and Y. Ma, “Minimum sum of distances estimator: robustness and stability,” Am. Control Conf. 2009. …, pp. 524–530, 2009.

[15] H. Kriegel and M. Schubert, “Classification of Websites as Sets of Feature Vectors.,” Databases Appl., pp. 127–132, 2004.

[16] W. Hung and R. J. Mcqueen, “Developing an Evaluation Instrument for e-Commerce Web Sites from the First-Time Buyer ’ s Viewpoint,” pp. 31–42, 2003.

[17] S. J. Barnes and R. T. Vidgen, “AN INTEGRATIVE APPROACH TO THE ASSESSMENT OF E-COMMERCE QUALITY,” no. August 1998, pp. 114–127, 2000.

[18] F. Herrera, “A 2-tuple fuzzy linguistic representation model for computing with words - Fuzzy Systems, IEEE Transactions on,” vol. 8, no. 6, pp. 746–752, 2000.