Conceptual Classification to Improve a Web Site Content

Sebastián A. Ríos[1], Juan D. Velásquez[2], Hiroshi Yasuda[1], and Terumasa Aoki[1]

[1] Applied Information Engineering Laboratory, University of Tokyo, Japan
{srios, yasuda, aoki}@mpeg.rcast.u-tokyo.ac.jp
[2] Department of Industrial Engineering,
University of Chile, Chile
jvelasqu@dii.uchile.cl

Abstract. This paper presents a conceptual based approach for improving a Web site content. Usually Web Usage Mining (WUM) techniques study the visitors' browsing behavior to obtain interesting knowledge. However, most of the work in the area leave behind the semantic information of web pages. We propose to combine the Concept-Based Knowledge Discovery in Text with the visitors sessions to perform the personalization task. This way, it is possible to obtain information about which are the users' goals when browsing a web site. Moreover, it is possible to give better browsing recomendations and help managers improving the content of their Web site. We test this idea on a real Web site to show its effectiveness.

1 Introduction

The World Wide Web has became an important way to reach information on almost any topic rapidly and effortlessly. Also, the Web has opened a new way of doing businesses, i.e. amazon.com that is one of the most used examples.

One important issue is how to deal with an overwhelming amount of documents. Most of the better search engines like google, yahoo, altavista, among others use algorithms based on keywords. However, when the web users perform a searching task have some questions, ideas or goals in mind [4]. Similarly, when a user finally reach a web site, she/he need to read the content in order to find if this information is suitable to her/his needs or goals. These problems get worse with the fast growing of the Web and forces to a new way of designing and developing web sites [3] to give a better browsing experience to the visitors.

Improving the web site usability, structure and content to keep the visitors interested on it is a challenging task [7]. Many techniques like Web Text Mining (WTM), Web Structure Mining (WSM), Web Usage Mining (WUM), Web Personalization (WP), etc. are used to help managers and web masters improve a web site [1] or automatically giving an on-line recommendation directly to the visitors. Many times when applying such techniques combined with keyword based algorithms the semantics of the web pages is lost. We define concepts trying to give a simple solution which consider this semantic factor. We show that the resulting process better fit users' needs and goals.

E. Corchado et al. (Eds.): IDEAL 2006, LNCS 4224, pp. 869–877, 2006.

The paper is organized as follows. Section 2, we show a short review about related research work. Section 3 explains how to identify and define concepts and how to apply the conceptual approach for web pages classification. Afterwards, in Section 4, we show how to improve the WUM using the conceptual information. Section 5, an experiment in a real-world case is presented. Finally, Section 6 presents the main conclusions and future work.

2 Related Work

There are more than one approach for improving the visitors' browsing experience in a web site. Many researchers focus on text content improvements [9,10], to do so they use a text preprocessing stage, sometimes a stemming process is applied to reduce the number of features and obtain better results in the generalization process which will applied later. Finally, the expert's collaboration to validate the results is always desirable.

Other researchers argue that in order to improve a Web Site we need to focus on how the users browse on it. This is called Web Usage Mining (WUM) and several works have been developed in this area, one example is Mobasher et al [5,12]. However, other researchers realize that better web sites' improvements recomendations can be obtained using a combination of visitors browsing behavior plus the textual content of the pages visited, some examples are [8,11,13].

All these techniques probe their effectiveness to help improve sites' usability. However, non of them take into consideration the semantic information of the web pages. Some authors realize this issue, and developed approaches aiming to consider the semantics of documents when performing the mining technique. A very good example is the Semantic Web Personalization System (SEWeP) [3], this system use concepts defined on a domain taxonomy to obtain the semantics of documents, afterwards, enhance the Web personalization process.

Other interesting work related with semantics but not with personalization is the one developed by Chau et al. in [2]. She is focused in the semantics from multilingual documents written in Chinese and English. She uses fuzzy logic to define concepts and afterwards she run a Fuzzy K-Means algorithm to filter the multilingual documents in topics regardless of the language. Afterwards, a Self Organizing Map (SOM) is used to obtain a topic-oriented multilingual text classification.

In our case, we intended to use the concept-based knowledge discovery in text proposed by Loh et al. in [4] to improve the WUM process. In his proposal Fuzzy Logic is used in order to define concepts which express the semantics of documents. Then he applied a fuzzy reasoning model to classify the documents into its concepts. Finally the application of statistical techniques allow to discover interesting patterns in concept distribution.

3 Conceptual Approach for Web Pages Classification

To begin with the conceptual approach we need to understand the meaning of the word "concept". From a Spanish dictionary a "concept" is an "idea, opinion

or thought", from WordNet 2.0 is "an abstract or general idea inferred or derived from specific instances". Both definitions show the ambiguity and subjectivity of the word. Also, they show the inference capacity which humans have for performing different tasks.

3.1 Identification and Definition of Concepts

We worked with the web site of the Faculty of Sciences Physics and Mathematics of the University of Chile. Identification of concepts was performed with sites' expert help. This way was possible to establish a set of concepts which can be important for the visitors of the site as shown in Table 1.

Table 1. Small sample of concepts identified for the site in Spanish

CONCEPTOS	CONCEPTS
ACTIVIDADES EVALUATIVAS	EVALUATIVE ACTIVITIES
SERVICIOS GENERALES	GENERAL SERVICES
SERVICIOS PERSONALES	PERSONAL SERVICES
ACTIVIDADES EXTRAPROGRAMATICAS	EXTRACURRICULAR ACTIVITIES
CALENDARIO DE PRUEBAS	TEST SCHEDULE

Note that we don't use all possible concepts, just the most important to the visitors based on the expert criteria. Afterwards, we need to define these concepts by a coherent combination of words [3,4]. To do so, we used a synonyms dictionary to extract words to characterize each concept also we use quasi-synonyms. A quasi-synonym from the dictionary is "a term in a controlled vocabulary, such as a thesaurus, that is treated as if it means the same thing as another term". For example. In the case of "Personal services" we consider words like "agenda" or "u-agenda" which is the name of the agenda system for the students and professors of the Faculty. Other example is the incorporation of the word "U-Cursos" that is the name of the portal for all courses in the Faculty. This contains all the classes documents (.ppt, .doc, etc), bibliography and many other useful information.

3.2 Fuzzy Logic for Pages Classification

We decided to apply the Fuzzy Reasoning model proposed by Loh et al. [4]. To characterize the documents of the web site. This solution is based in the idea that we need to gather the relation between the concepts and the documents which can be represented as a fuzzy composition, shown in Eq.(1). In that expression the terms $[Concepts \times Terms]$ and $[Terms \times Words]$ are fuzzy relations, therefore matrices. The operator $\circ$ represent the compositional rule of inference according to Nakanishi et al. [6]. From this point we call "terms" to the special words that represent a concept and "words" to any word in a document like a web page.

$$[Concepts \times Words] = [Concepts \times Terms] \circ [Terms \times Words] \quad (1)$$

In order to apply the expression in Eq.(1) we defined a list of concepts and terms that represent these concepts in the previous section. However, we still

Table 2. A column extracted from the compositional matrix [$Concepts \times Words$]

URL: http://escuela.ing.uchile.cl/servicios.htm	
ACTIVIDADES EVALUATIVAS	=> 0
SERVICIOS GENERALES	=> 0.707106781187
SERVICIOS PERSONALES	=> 0.424264068712
REGLAMENTACION	=> 0
INFORMACION GENERAL ESCUELA INGENERIA	=> 0.707106781187
CLASES	=> 0
CALENDARIO DE PRUEBAS	=> 0
....	

need to set up the membership values for this relation. Once again we use the experts' knowledge to define this values (direct method with one expert).

We used a simple model that use relative words frequency on each document to deffine the second fuzzy relation [$Terms \times Words$]. The documents were preprocessed to eliminate the HTML and JavaScript code and using a stop word list we also erased words that are not important. We intended to keep nouns, adjectives and verbs. At this point we decided to not use stemming process to maintain the words intact and compare it with its synonyms and quasi-synonyms without problems of having one stem and more than one possible word.

After applying the compositional rule of inference we obtain the [$Concepts \times Words$] matrix were each row is a concept and each column is a web page and each value in the matrix represent the degree of possibility of a concept to be represented in a web page (i.e. the membership value of the composition shown in Eq.(1)). Therefore, we achieved a Conceptual based classification for each web page on the site. An example is shown in Table 2, where we have the column that represent "http://escuela.ing.uchile.cl/servicios.htm" (services.htm in English) and then we have the concepts and its membership values. In this case this page contains information of services for students, professors and links to them. We can see that the concept "SERVICIOS GENERALES" (General Services) has a membership value of 0.707106781187 and "SERVICIOS PERSONALES" (Personal Services) has a membership value of 0.424264068712. As we mention before this can be interpreted as "the degree of the concept X to be represented in the page servicios.htm". In this particular example we can observe for example that the page is more semantically related to the concept "SERVICIOS GENERALES" than "SERVICIOS PERSONALES" this is because most of the services are not only for students or professors but also for the whole comunity. Similarly others concepts in Table 2 have a membership value of 0 which means that the page doesn't talk about these concepts.

4 Mining User Sessions

Several techniques to perform WUM over a site exists. We chose to perform a WUM that combine the text of the documents and the visitors' sessions according to [11,13] and then compare its results with the concept-based approach.

To begin with WUM process we need to transform the web pages in a more useful way. We use the Vector Space Model to represent each document as a vector of words combined with the $TF * IDF$ to establish the wight of every word (frequency) on each document. On the other hand, we need to pre-process the web servers' logs to figure out which are the visitors' sessions. We perform a time heuristic sessionization over the web logs. After this step, we have the pages and time spend on each document visited for the visitors.

4.1 Visitors' Session Classification

After the pre-processing stage we have to apply the generalization process. We selected to use a Self Organizing Feature Map (SOFM) because it is an unsupervised algorithm which means we don't need to know before hand how many clusters are. We used a process similar to the one published in a previous works [13,11].

We based our work in a similarity measure proposed by Velasquez et al in [13] that combines the visitors sessions and the content of the web pages. However, we intended to generate more semantically related results rather than textual contents related. This is why we modified this similarity in order to apply the conceptual classification obtained before. The similarity measure used is shown in Eq.(2) were $CS(S^i, S^j)$ is the similarity between the session S^i from visitors i and session S^j from visitor j. The sessions S should be of the same length so we sort the pages per time spent on each. Then we use the first ι pages where the visitor spend most of his time. The formula uses the $S^j_\tau(k)$ as the time in seconds spent by visitor j in page k and similarly for visitor i. The term $S^j_\rho(k)$ is a vector which contains the concepts and the degree of membership for page k of visitor j session (Section 3.2) . This way, the equation is unaltered from its original form but we are using concepts instead of terms frequency to compare the sessions.

$$CS(S^i, S^j) = \frac{1}{\iota} \sum_{k=1}^{\iota} min(\frac{S^i_\tau(k)}{S^j_\tau(k)}, \frac{S^j_\tau(k)}{S^i_\tau(k)}) * PD(S^i_\rho(k), S^j_\rho(k)) \quad (2)$$

Finally, we obtain clusters that are related by its conceptual meaning. Therefore, we obtain a conceptual classification of the documents on each visitors' session.

5 Experiments in a Real Web Site

We applied this process to the web site of the School of Engineering and Sciences of the University of Chile [1]. This web site has 182 web pages and it is almost static throw the year (only the news page change continuously), thus we used the version of March 2006 of the web site. Besides, we took approximately 2 months of web logs 2006.

[1] http://escuela.ing.uchile.cl

We detect 12 important concepts, we defined them using a dictionary, a thesaurus and the experts' help. Afterwards, each document was classified into its concepts as explained before (see Table 2).

We decided to use a SOFM of 7×7 neurons. Then we kept the 3 pages per session were the visitors spend most of its time. Afterwards, we compute the similarity measure in Eq.(2) for the SOFM through 50 epochs. In order to compare the results obtained with the conceptual approach. We also compute the traditional approach that uses $TF * IDF$ words vector instead of a vector of concepts in Eq.(2). Using the same network size, epochs, and session length.

Our results using the traditional approach are shown in the Table 3. We can observe that we obtain three clusters for the users sessions $\{0, 1, 2\}$. Each cluster found, is represented for several visitors' sessions, e.g for cluster $ID = 0$ we have a list of nine sessions $\{1, 2, 50, 58, 64, 66, 78, 127, 262\}$. Each number in this list is the ID of a visitor session. For example, the session ID=1, has 3 pages which are $\{mail2.htm, barraizquierda.htm, maincalendarios.htm\}$ where the visitor spent most of his/her time (See Section 4.1). The problem in this case is we have the pages that the visitors browsed and we now that he/she was interested in the text inside that pages. Usually the recommendation is to link those pages. However, with the conceptual approach we can discover what topic he/she is interested in and we can give a family of pages that satisfy his/her needs, even if the pages are not present in the cluster sessions.

Table 3. Results of traditional approach

CLUSTER ID	SESSIONS [$IDs.$]
0	$\{1, 2, 50, 58, 64, 66, 78, 127, 262\}$
1	$\{54, 57, 66, 67, 69, 74, 90, 112, 146\}$
2	$\{0, 66, 93, 123, 184, 224, 256, 267\}$

The results of the conceptual approach proposed are shown in Table 4. This time we discovered five clusters (two more clusters than using the traditional approach). Afterwards, we extracted the concepts on the real sessions representative of each cluster. To do so, we computed the common concepts to all documents on the cluster. The result of this is shown on Table 5. If there is no concept common to all of the documents on the cluster sessions, then the value "N/A" is given to that cluster.

Observing this results on Table 5 we can notice that Cluster $ID = 1$ is about people who is looking for contact information, telephone numbers, the name and e-mail of people in charge of the administrative office or certificate office or other area. Other interesting information is obtained from cluster $ID = 0$ where the concepts are { *Vacations Schedule, News/Advertisements, Seminars/Extention, Extracurricular Activities, Organizations* }. This means that the students look for activities to do on vacations. Other interpretation is that when a visitor to the site (most of them students) search for vacations schedule, also is looking for activities to do on his/her free time. This is way he/she also browse for

Table 4. Results of conceptual approach

CLUSTER ID	SESSIONS [$IDs.$]
0	$\{57, 66, 93, 102, 224, 230, 268\}$
1	$\{56, 66, 90, 93, 269\}$
2	$\{66, 78, 224, 256, 268\}$
3	$\{4, 57, 66, 78, 239, 267\}$
4	$\{2, 66, 78, 223, 265\}$

Table 5. Concepts obtained from clusters

CLUSTER ID	CONCEPTS
0	{ Vacations Schedule, News/Advertisements, Seminars/Extention, Extracurricular Activities, Organizations }
1	{ Location Information / Physical Addresses, Contact Information }
2	{ Classes Material, Tests Schedule, Classes Inforation. }
3	N/A
4	{ Reglaments, Classes Information, Tests schedule }

extracurricular activities information, Organizations Information (Photography Club, Role Gamimg Club, etc).

If we study the structure of the pages on the cluster sessions, we notice that many of the page in the session, are not close one to the other. In the case of Cluster $ID = 1$, the schedules are in a section different from the section of the Organizations and this both in a section different from the section of extracurricular activities. The visitors must browse all the sections doing several clicks to reach the calendars then exit this section and browse down in the next section to find the information about Role Gaming Club. Then, the recommendation in this case is to link the relevant pages on this cluster. Many alternative exist, from creating one single page with all this information, to link the existing pages among them. Moreover, since we have all web pages classified by its conceptual meaning we do not need to limit the recommendation to those pages in the cluster. We can also generate a single page where all the pages with information about extra curricular activities or organizations we have. When we talk about a concept we talk about a family of web pages that express the concept in a certain degree.

5.1 Discussion

After extracting the concepts form each real page on the session, we discover that cluster $ID = 3$ it is not valid from concepts perspective. If we use other manner to extract concepts from the cluster we can obtain results. However, we need more work to do in this area, because it is not simple and the results can change greatly depending on the technique used to extract the concepts form the sessions.

We need more work about the concepts base used. Our experiment only take into consideration a small amount of concepts (12) however, it is possible to define a wider concepts base to obtain more information about our visitors.

6 Conclusions

Many different techniques and methodologies exists to help managers or web masters improve web sites' usability. However, most of them do not consider the semantic information from the web documents. We propose a simple process to achieve a conceptual classification of documents using a fuzzy reasoning model. Then a Self Organizing Feature Map for the Generalization stage.

We use this process to improve the Web Usage Mining process results, to obtain patterns that have more relation with the visitors goals and then recommend managers changes in the web sites' content.

We perform two experiments using a traditional WUM approach and then we used our proposal in a real Web site.

We are working in more experimental results as well as in the evaluation of the whole concept-based usage minig process proposed in this work.

References

1. S. Chakrabarti. Data Mining for Hypertext: A Tutorial Survey. *SIGKDD Explorations*, 1, 2000.
2. R. Chau and C.-H. Yeh. Filtering multilingual web content using fuzzy logic and self-organizing maps. *Neural Comput. Appl.*, 13(2):140–148, 2004.
3. M. Eirinaki, C. Lampos, S. Paulakis, and M. Vazirgiannis. Web personalization integrating content semantics and navigational patterns. In *WIDM '04: Proceedings of the 6th annual ACM international workshop on Web information and data management*, pages 72–79, New York, NY, USA, 2004. ACM Press.
4. S. Loh, J. P. M. D. Oliveira, and M. A. Gameiro. Knowledge discovery in texts for constructing decision support systems. *Applied Intelligence*, 18(3):357–366, 2003.
5. B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining. *Commun. ACM*, 43(8):142–151, 2000.
6. H. Nakanishi, I. B. Turksen, and M. Sugeno. A review and comparison of six reasoning methods. *Fuzzy Sets and Systems*, 57(3):257–294, Aug. 1993.
7. J. Nielsen. User Interface directions for the web. *Communications of ACM*, 42(1):65–72, 1999.
8. M. Perkowitz and O. Etzioni. Adaptive web sites. *Commun. ACM*, 43(8):152–158, 2000.
9. S. A. Ríos, J. D. Velásquez, E. S. Vera, H. Yasuda, and T. Aoki. Establishing guidelines on how to improve the web site content based on the identification of representative pages. In *IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology*, pages 284–288, Compiegne, France, September 2005. IEEE Computer Scociety.
10. S. A. Ríos, J. D. Velásquez, E. S. Vera, H. Yasuda, and T. Aoki. Using SOFM to Improve Web Site Text Content. In *Advances in Natural Computation:* 1^{st} *Int.l Conf., ICNC 2005*, volume 3611 of *Lecture Notes in Computer Science*, pages 622–626, Changsha, China, August 2005. Springer-Verlag GmbH.

11. S. A. Ríos, J. D. Velásquez, H. Yasuda, and T. Aoki. Web Site Improvements Based on Representative Pages Identification. In *AI 2005: Advances in Artificial Intelligence: 18th Australian Joint Conference on Artificial Intelligence*, volume 3809 of *Lecture Notes in Computer Science*, pages 1162–1166, Sydney, Australia, November 2005.
12. M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa. A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis. *INFORMS J. on Computing*, 15(2):171–190, 2003.
13. J. D. Velásquez, S. A. Ríos, A. Bassi, H. Yasuda, and T. Aoki. Towards the identification of keywords in the web site text content: A methodological approach. *International Journal of Web Information Systems*, 1(1):11–15, March 2005.