Web Site Off-line Structure Reconfiguration: A Web User Browsing Analysis

Sebastían A. Ríos[1], Juan D. Velásquez[2,3], 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] Center for Collaborative Research, University of Tokyo, Japan
[3] on leave from Department of Industrial Engineering, University of Chile, Chile, jvelasqu@dii.uchile.cl

Abstract. The correct web site text content must be help to the visitors to find what they are looking for. However, the reality is quite different, many times the web page text content is ambiguous, without meaning and worst, it don't have relation with the topic that is shown as the main theme. One reason to this problem is the lack of contents with concept meaning in the web page, i.e., the utilization of words and sentences that show concepts, which finally is the visitor goal. In this paper, we introduce a new approach for improving the web site text content by extracting Concept-Based Knowledge from data originated in the web site itself. By using the concepts, a web page can be rewrite for showing more relevant information to the eventual visitor. This approach was tested in a real web site, showing its effectiveness

1 Introduction

The Web has became a immense collection of documents which contain valuable information in almost every imaginable topics. However, the problem of searching in this huge ocean of data is neither easy nor fast. These problems get worse with the fast growing of the Web and forces to a new way of designing and developing web sites [3]. Moreover, improving the web site usability, structure and content to keep the visitors interested on it is a challenging task [8].

In order to improve the web site structure and content, the web mining techniques have shown a reasonable effectiveness [4]. Specify, the Web Content Mining (WCM) and Web Usage Mining (WUM) techniques have being used for analyzing the web page content (mainly text) and the visitor browsing behaviour respectively [6, 15]. From the patterns extracted by using these techniques, recommendations about hyperlinks and content modifications are obtained.

This paper aims to provide a suitable content recommendation by applying a WCM technique for discovering Content-Based Knowledge, which is used for reducing the vocabulary problem present in a web page, i.e., badly utilization of irrelevant and redundant words for creating web page text content without meaning for the visitor.

This paper is organized as follows. In Section 2, a short review about related research work is done. Section 3 shows the concept discovery task in text contents. The proposed methodology is introduced in Section 4, and in Section 5 its application to a real-world case is presented. Finally, Section 6 presents the main conclusions and future work.

2 Related Work

Web site personalization according to Eirinaki et al. [3] is "*any action that adapts the information or services provided by a Web site to the needs of a user or set of users, taking advantage of the knowledge gained from the users' navigational behavior and individual interest, in combination with the content and the structure of the Web site*". Besides, web personalizations' objective according to Mulvenna et al. [7] is to "*provide users with the information they want or need, without expecting from them to ask for it explicitly*".

Many different approaches have being developed in order to perform a Web site personalization in the best way. The majority of the efforts correspond to those which only take into consideration the data of usage [6, 9, 13]; however, other researchers improved the personalization process incorporating the knowledge that is underlied in the textual content [1, 9, 10, 15] or structure [10] the site.

As a natural evolution to those approaches, the need for a solution that take into account the semantical information of the web site have been developed lately. Eirinaki et al. have developed the Semantic Web Personalization System (SEWeP) [3]. This work consist in combining web usage with content Knowledge. Then, they developed an enhanced version of the web logs registers which are called C-Logs (concept logs). These C-Logs consist of web sites' semantic information that is added to the traditional usage logs in the way of keywords. Afterwards, these C-Logs are used in the mining process to obtain better and broader recommendations.

On other hand, Knowledge Discovery in Text (KDT) concerns to the application of Knowledge Discovery in Databases (KDD) techniques over free text. Loh et al. use KDT process for developing a Concept-Based Knowledge Discovery process for texts [5]. In that work, the KDT techniques are applied over concepts rather than on attribute values, terms or keywords labeling texts. Then statistical analysis are performed to obtain interesting patterns. One of Lohs' objectives was to allow the user to search ideas, ideologies, trends and intentions presents on text.

3 Concept-Discovery in Text for Personalization

3.1 Concept Representation Model

The concept representation is not an easy task, inclusive the meaning of the word concept also is quite ambiguous. A concept from dictionary is an "idea,

opinion or thought". We can understand from this definition that the concepts have relation with ambiguity and also with the subjectivity that is commonly used by humans for performing different tasks, like, for instance, browsing a web site to obtain information about some topic.

Concepts are represented by a coherent combination of words [3, 5]. However, in order to express an idea or event, we need to understand that not all words represent a concept in the same level, degree or context. For example, in Spanish the word "cancelar" (cancel) means "to stop doing something" however, it also means "to pay a bill". In this example we have two different concepts represented by one word depending on the context. Another example are the phrases "i bought a dog" and "i didn't buy a dog" the concept represented in one case is buying a pet but in the other phrase is the negation of that. For the explained reason, we need to represent the words with a weight that show the degree of relation that a term has to represent a specific concept.

From several approaches to represent concepts, we chose the Vector Space Model [12], which consist in transforming the text of the web site in a vectors of words. Each document is represented by a N dimensional vector, where N is the number of different words the whole site.

Chakrabarti [2] said that is better to use weighted vector instead of a binary model because the precision is higher. This is why we use a simple weight calculated using TF*IDF for each term and then we normalize the weight vector for each web page.

However, this weight only give some hints about the relative importance of the terms on the specific document. We still need to define how these terms represent a concept.

3.2 Definition and identification of Concepts

For defining concepts, we used a dictionary of synonyms and antonyms for Spanish. It is important to differentiate between words in the text of the Web pages and terms. We use terms to refer those words that are used to represent a concept.

We based our work in the one proposed by Loh [5], his idea is to use a *fuzzy reasoning* model to decide wether a concept is expressed by a web document or not. Computing the possibility for a concept to be on a text based on the weights obtained before and the membership values from the terms that represent a concept. The existence of Necessary Conditions (NC) and Sufficient Conditions (SC) allows to perform such task. If a SC is present then the presence of a concept is mandatory ($TERM \Rightarrow CONCEPT$). While NC are the consequence of the presence of a concept ($CONCEPT \Rightarrow NC$). In the case of this work we will use only the NC. It means that if a term that defines a concept appears in a web page text, then there is a high possibility this concept belong to the document. According to this we generate a list of terms that define a concept based on the definition extracted from the dictionary. In this first proposal we set up a simple binary weigh system: 1 if the term is a synonym or 0 if antonym

or any other word that it is not related with the concept. This system is very simple and does not take into account quasi synonyms or context of terms.

When having these two vectors, one for the words of the Web pages and other for the concepts, we need to apply our fuzzy reasoning model. We used Eq. (1) from [5] for such purpose.

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

In the Eq.(1) "$\circ$" represent a combination between two fuzzy relations and the symbols "[]" represent a fuzzy relation (can be a matrix).

After running the process that compares all the word vectors with the terms vectors we write a file for each web page that contains the concepts for the page.

4 Concept-based session analysis

4.1 Important concept-based web pages vector

Assuming that the degree of importance in some web page content is correlated with the time spent on it by the visitors, we can state that those pages where a visitor spends more time are those more interesting to him. To represent this idea Velásquez et al. in [14] defined the ι-Most important pages vector, however, now we need to slightly change that definition to incorporate the notion of concepts. Thus we redefined the ι-Most important pages vector as follows:

Def. 1 (ι-Most important concept-based pages vector) *We define a vector of two components $\vartheta_\iota = [(\kappa_1, \tau_1), \ldots, (\kappa_\iota, \tau_\iota)]$ where the pair $(\kappa_\iota, \tau_\iota)$ represents the ι^{th} most important page based on the time spent. Then κ is the vector of concepts that represent a page and τ is a scalar value to represent the percentage of time spent in it within a visitor session.*

For applying successfully the Def. 1 we need to develop a similarity measure that allows to compare vectors. In a previous work, the *Important Visited Pages Similarity* (IVS) [11, 14] was introduced. However, this similarity is not suitable to be used in the present work because it do not use the ι-Most important concept-based pages vector. Therefore, IVP similarity uses a combination of the relative time spent in two web pages and the textual content of those pages.

Now, we need a new similarity that allows combine the relative time spent with the concepts web pages visited. Fortunately, if we use the new definition ι-Most important concept-based pages vector, the changes on IVS are minimal. Then it is possible to modify IVS in order to include the concepts, as we show in Eq.(2). We called this expression *Important Concepts-Based Visited Pages Similarity (ICVS).*

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

The expression shown in Eq.(2) compares the ι-most important concept-based pages vectors into the sessions of two different visitors S^i and S^j. On the other hand, the function $PD()$ is introduced in Eq.(3). This is the way in which we combine the content of the site with the visitors browsing preferences. The term $S^i_\tau(p)$ represent the time spent on page p for the visitor i. Similarly, $S^i_\kappa(p)$ are the concepts that represents the page p for the visitor i.

$$PD(p_i, p_j) = \frac{\sum_{k=1}^{W} p_{ki} p_{kj}}{\sum_{k=1}^{W} (p_{ki})^2 \sum_{k=1}^{W} (p_{kj})^2} \tag{3}$$

The *Page Distance* introduced in Eq.(3) is the dot product between two vectors p_i and p_j.

When two visitors sessions are similar in browsing time $min(\frac{S^i_\tau(p)}{S^j_\tau(p)}, \frac{S^j_\tau(p)}{S^i_\tau(p)}) \approx 1$ and if the sessions are similar in content $PD(S^i_\kappa(p), S^i_\kappa(p)) \approx 1$ furthermore $IVS(S^i, S^j) \approx 1$. In the opposite case, if the text contents are dissimilar then $PD(S^i_\kappa(p), S^i_\kappa(p)) \approx 0$ the expression $IVS(S^i, S^j) \approx 0$. On the other hand if the times spent are very different then $min(\frac{S^i_\tau(p)}{S^j_\tau(p)}, \frac{S^j_\tau(p)}{S^i_\tau(p)}) \approx 0$.

One important observation is that, even thought the expression for $ICVS$ is almost the same with the IVS, the results are totally different because, they use totally different processing vectors.

4.2 Analyzing the visitor behaviour in a web site

We used a *Self Organizing Feature Map* (SOFM) as clustering method to discover significant patterns from the combination of the web pages' concepts and the visitors spent time per page. We select a toroidal topology because it maintain the continuity of the space [11, 14, 15].

The SOFM is randomly initialized. This means that the neurons feature vectors, which are normalized, are created with random values between $[0, 1]$ in the epoch $t = 0$.

We use a Gaussian function that depends on the distance from the centroid to propagate the learning to the neighbor neurons. This function allows the centroid neuron to learn the pattern shown. Afterwards, the effect of the learning is passed to the neighborhood to a lesser degree, inversely proportional to the centroid distance.

We applied the $ICVS$ shown in Eq.(2) to compare the sessions examples with the documents concepts.

At the end of the process we can take the SOFM and we applied a technique that we called Reverse Clustering Analysis (RCA). The RCA technique for WUM is explained in detail in [11]. This technique allows discover which are real pages that are in the clusters of the SOFM and for this way, to create content recommendations for the web site.

5 Experiments in a real web site

The whole process explained before was applied to the web site of the School of Engineering and Sciences of the University of Chile.[4]. 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 December 2005 of the web site. Besides, we took approximately 2 months of web logs November and December 2005.

The length of the *ι-most important concept-based pages vector* was set in three pages. Therefore, we needed the sessions which contain at least three pages visited to create those vectors in order to apply the Definition 1. To do so, we sorted the sessions by time spent on each page and then we only kept the three pages where the visitor spent more time.

This experiment was performed using a combination of web pages content and the visitors sessions. First, we clean the Web pages text with a stop words list, then we applied a stemming process. The concepts used in this case were the concepts form the titles of the web pages and links. We discover that several times we had titles that have different words however, this pages use synonyms of the words. Afterwards, we compute the TF*IDF for all the concepts.

The next step was to apply a sessionization process on the web log registers, i.e., to reconstruct the original visitor session.

As final step, a visitor behaviour pattern extracting process was carry out by using a SOFM of 8×8. The results are shown in Figure 1. The results in this case were just four main clusters for the 64 documents used before.

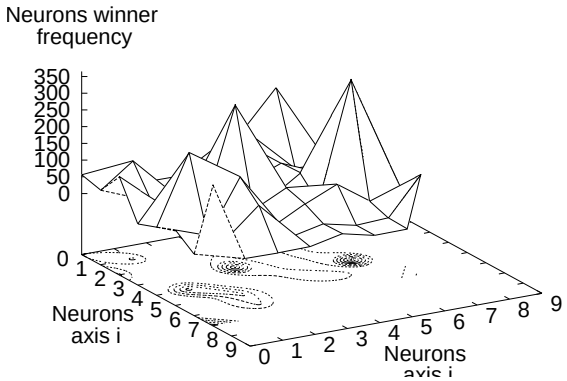

Fig. 1. Results using the concept-based approach.

One example, we have many pages with title: “Faculty of Science Physics and Mathematics”, “School of Engineering”, “Faculty of Engineering”, etc. The vector that represent the words are very different one to each other. This is why

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

in the traditional approach we obtained several clusters although in the concept based approach this clusters are all one, because all of them are referring to the School of Engineering.

The concepts reduce the amount of words to process in the creation of a vectorial representation for a web page. An example of this is what happen in the particular case of the analyzed web site. The compound words "Faculty of Science Physics and Mathematics", "School of Engineering" and "Faculty of Engineering" have the same semantic meaning, so they are similar when the Eq. (3) is applied.

From the clusters extracted, it is possible to extrapolate that the visitors are interested in:

- Test calendar, which is expressed for the concepts "prueba" (test), "control"(a monthly examination), and "examen" (the semester final examination). All of these words appear in different pages, then a recommendation is to create a unique page with the whole information.
- Educational material, which is expressed for the concepts "cátedra" (main lecture), "clase auxiliar" (lecture for resolving problems and exercises), "tareas" (homework) and "laboratorios" (laboratories). These concepts appear in different pages, which is no bad, but alternatively, could be necessary a unique page that concentrate the whole information

6 Conclusions

We show our first attempt to obtain concept-based mining technique which allows to obtain patterns that have more relation with the visitors goals. The process then can be used for the off-line personalization of a web site, in order provide text content and structure recommendations for modifying the web site.

This web mining approach allows to analyze a web site from the concept point of view, i.e., now the question about what the visitor is looking for change from "which words?" to "which idea?".

Because before to apply our concept-base web mining algorithm it is necessary to reduce the page text content to concepts, there is a manual previous stage, i.e., a human being must to read the page text and to specify which concepts are inside.

As future work, we want to develop a semi-automatic preprocessing algorithm for extracting concepts from a web page text content.

Acknowledgmentt

This work has been funded partially by theMillenniumm Scientific Nucleus on Complexes Engineering Systems, Chile.

References

1. B. Berendt and M. Spiliopoulou. Analysis of navigation behavior in web sites integrating multiple information systems. *The VLDB journal*, 9:27–75, 2001.
2. S. Chakrabarti. Data Mining for Hypertext: A Tutorial Survey. *SIGKDD Explorations*, 1, 2000.
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. M. Eirinaki and M. Vazirgiannis. Web mining for web personalization. *ACM Trans. Inter. Tech.*, 3(1):1–27, 2003.
5. S. Loh, L. K. Wives, and J. P. M. de Oliveira. Concept-based knowledge discovery in texts extracted from the web. *SIGKDD Explor. Newsl.*, 2(1):29–39, 2000.
6. B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining. *Commun. ACM*, 43(8):142–151, 2000.
7. M. D. Mulvenna, S. S. Anand, and A. G. Buchner. Personalization on the net using web mining: introduction. *Commun. ACM*, 43(8):122–125, 2000.
8. J. Nielsen. User Interface directions for the web. *Communications of ACM*, 42(1):65–72, 1999.
9. M. Perkowitz. *Adaptative Web Sites: Cluster Mining and Conceptual Clustering for Index Page Synthesis*. PhD thesis, Univerity of Washingtong, 2001.
10. M. Perkowitz and O. Etzioni. Adaptive web sites. *Commun. ACM*, 43(8):152–158, 2000.
11. S. A. Ríos, J. D. Velásquez, H. Yasuda, and T. Aoki. Web Site Improvements Based on Representative Pages Identification. In S. Zhang and R. Jarvis, editors, *AI 2005: Advances in Artificial Intelligence: 18th Australian Joint Conference on Artificial Intelligence*, volume 3809, pages 1162–1166, Sydney, Australia, November 2005. Lecture Notes in Computer Science.
12. G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM archive*, 18(11):613–620, November 1975.
13. 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.
14. 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.
15. J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. A new similarity measure to understand visitor behavior in a web site. *IEICE Transactions on Information and Systems, Special Issues in Information Processing Technology for web utilization*, E87-D(2):389–396", February 2004.