Improving Web Site Content Using a Concept-based Knowledge Discovery Process

Sebastián A. Ríos*, Juan D. Velásquez†, Eduardo S. Vera‡§, Hiroshi Yasuda* and Terumasa Aoki*

* Applied Information Engineering Laboratory, University of Tokyo, Japan, Email: {srios,yasuda,aoki}@mpeg.rcast.u-tokyo.ac.jp
†Department of Industrial Engineering, University of Chile, Chile Email: jvelasqu@dii.uchile.cl
‡ Center for Collaborative Research University of Tokyo, Japan Email: esvera@vp.ccr.u-tokyo.ac.jp
§ On leave from Department of Electrical Engineering University of Chile, Chile

Abstract—Nowadays many organizations and enterprises have a web site to help accomplish several business goals from sales to customer support. A very important activity for today’s managers is how to improve the experience of visitors in their web site and make it more effective. Several techniques that use knowledge from visitors’ browsing behavior have shown good results to perform such task. However, how to introduce semantics into web usage mining techniques is still a challenge. This work presents a way to obtain a semantic classification of visitors’ sessions using a conceptual mining process. We performed experiments in a real web site.

I. INTRODUCTION

Internet has reached an unprecedented importance not only for visitors who can reach effortlessly information in almost any possible topic but also organizations and companies which have a new way to sell, promote, give customer support, among others.

The effectiveness of a commercial web site is a very important issue because it allows to retain old customers and gain new ones. Nielsen [7] discusses several web site usability problems and also establishes that effectiveness is strongly related with its usability.

The need of analyze in some way the visitors’ browsing behavior lead to the development of Web Usage Mining (WUM). Nevertheless, a new issue surfaced which is the semantic gap between documents on a visitor session.

This paper proposes a way to perform a mining process which considers the semantic meaning of the documents using a concept-based knowledge discovery process[5]. Afterwards, we improve the similarity measure proposed by Velasquez et al. in [14] using this conceptual classification. Then we use this similarity in a Self Organizing Feature Map (SOFM) to classify the user sessions. The process was tested in a real web site allowing us to obtain patterns that are broader in meaning.

II. RELATED WORK

The simplistic solution for mining the web text content is the use of term frequency methods e.g. TFIDF family. However, the problem of these methods is the poor interpretation of the results. A very good survey of these methods can be found in [1], [8].

On the other hand, many researchers are trying to add semantics into the WUM process [3], [4] to obtain results which contain more meaning for a better interpretation. The most popular way to do so is the use of an ontology-based approach to represent the content of the web site [3], [4]. The main idea of all this methods are the characterization of all words in a domain specific ontology which contains all the semantic relationships among words e.g. if a word is noun, verb, adverb, etc. and even more if some word is quasi-synonym or antonym of other words, of course the ontology should contain the formal definition of every term used.

One of the main drawbacks of these ontology-based methods is the construction and maintenance of the domain ontology. There are intense research in the area of automatically generate this domain ontology, a very good example could be found in [4]. In the mean while we need to manually generate the domain ontology in order to use these techniques. Other approaches which use WordNet or EuroWordNet have also been developed.

Eirinaki et al. [3] shows a very good application of ontolgy-based methods. They developed the Semantic Web Personalization System (SEWeP), this system use concepts defined on a domain taxonomy to obtain the semantics of documents, afterwards, they enhance the web logs by creating a C-logs (conceptual logs) to enhance the Web personalization process.

Other methods lesser popular are the use of fuzzy logic for the representation of concepts. Fuzzy logic is based on Fuzzy Sets theory, proposed by Zadeh [15], which allow us to
represent the degree of belonging of an element to a set. Using linguistic variables and membership functions we are able to represent concepts and the degree of this concept inside a web page. Using this information we can enhance the results of the mining process.

One interesting work related with fuzzy logic and semantics but not with personalization is the one developed by Chau et al. [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.

Similar to fuzzy logic we can also mention Dempster-Shafer (DS) model that is “...an elaborate formalism for representing and revising degrees of support rendered by multiple sources of evidence to a common set of propositions” [12]. This model is a probabilistic vision designed to quantify and revise the degrees of association between documents or between ontology terms as one may see in [4]. In this work, Li et al developed an algorithm for automatically discover an ontology which help them to understand the visitors’ needs. They use DS model to represent the support of each ontology class pattern based on a set of positive documents (called \( D^+ \)) used for training. Then they developed an ontology evolution algorithm called PatternEvolving to discover the final ontology. This process is very well presented with an evaluation of results compared with a simple rocchio and a traditional DS model however they lack of the use of visitors’ sessions to obtain visitors browsing preferences remaining a WTM method.

We are going to use the process proposed by Loh et al. in [5] to obtain patterns which are richer in meaning and this way easier to analyze. However, we change the statistical analysis used by Loh; by the use of a SOFM.

III. CONCEPTUAL DOCUMENTS CLASSIFICATION PROCESS

The main goal of this process to us is the obtention of a conceptual classification of the documents in a web site. To do so, we have used a fuzzy reasoning model which allow us to represent the documents in a suitable way to be used later by a classification algorithm like a SOFM, k-means, etc.

A. Fuzzy reasoning model

A document can be represented by its concepts using linguistic variables. The linguistic variable (LV) values are not numbers but words or sentences in natural language. These variables are more complex but less precise. If \( u \) is a LV we can obtain a set of terms \( T(u) \) which cover its universe of discourse \( U \) and represent \( u \). e.g. \( T(temperature) = \{cold, nice, hot\} \).

Then if we assume that a document can be represented as a fuzzy relation among LV we can write this as a composition of fuzzy relations. This is shown in Eq.(1) where \( WP \) represent the Web Pages and each [.] term is a matrix. This approach was proposed by Loh et al. [5].

\[
[Concepts \times WP] = \text{[Concepts \times Terms]} \otimes \text{[Terms \times WP]} \tag{1}
\]

If \( P \) is the total number of different web pages on the web site and \( W \) the total amount of different words, as defined before, and \( K \) the total amount of concepts identified for the web site. Then we are interested in the fuzzy composition matrix in Eq.(2). Where \( \mu_{C \times WP}(x, z) = \mu_{C \times T \otimes T \times WP} \) represents the membership function of the composition in Eq.(1). The membership values are between 0 and 1.

\[
\mu_{C \times WP}(x, z) = \begin{pmatrix}
\mu_{1,1} & \mu_{1,2} & \cdots & \mu_{1,K} \\
\mu_{2,1} & \mu_{2,2} & \cdots & \mu_{2,K} \\
\vdots & \vdots & \ddots & \vdots \\
\mu_{K,1} & \mu_{K,2} & \cdots & \mu_{K,K}
\end{pmatrix} \tag{2}
\]

There are several alternatives to perform the fuzzy composition. An important issue is that even if some terms are not present in a web page, the degree of expressing a concept doesn’t suffer alterations. We decided to use the Nakanishi et al. compositional rule [6].

IV. THE METHODOLOGY PROPOSAL

A. Combining concepts and visitors sessions

In order to generate a sessions classification using a classification algorithm, we need to establish a suitable way to represent the visitors sessions and similarity measure. We based the present work on previous works on the area that can be seen in [9], [14].

If we assume that the degree of importance in some page content is correlated with the time spent on it by the visitors. With this assumption we are able to define the \( \iota \)-Most important pages vector [14]. Which represent each visitors’ session as a vector of visited pages and each page is represented as a pair of content and time spent on it.

Then we are able to use the Velasquez et al. similarity measure [13] in Eq.(3). In this expression we compare the \( \iota \)-most important pages into the sessions of two different visitors \( V^{i} \) and \( V^{j} \). The function \( PD() \) is the dot product between vectors and compute the similarity between the content of the pages on the sessions. The \( min(\iota, \iota) \) term is the relative time spent between the pages being compared.

\[
IVS(V^{i}, V^{j}) = 1/\iota \sum_{k=1}^{\iota} min(V^{i}(k), V^{j}(k)) * PD(V^{i}(k), V^{j}(k)) \tag{3}
\]

This is the way in which we combine the content of the site with the visitors browsing preferences. But we want to combine the semantic information of the concepts within the sessions.

A good property of the application of a fuzzy reasoning approach is that the information is still in a suitable form to
be used by our clustering algorithm using the expression on Eq.(3) with minimum changes. An ontology based approach require the use of different similarities e.g compute proximity of terms on the ontology tree.

We redefine the \(t\)-Most important pages vector to use concepts and in this way we can continue using the IVS similarity. We are going to represent each visitors’ session as a vector of visited pages and for each page we will use a pair of concepts and time spent on it. Where the concepts are a list of membership values. This way instead of computing the dot product between two content vectors we use two concept vectors.

V. REAL SITE EXPERIMENTS

We performed the experiments into the site of the School of Engineering and Science of the University of Chile. This is a medium size web site with 210 web pages which contains information about the different specialties on engineering and several services for students, professors and general public.

We choose about one month of web logs and after sessionization process we obtained about 1023 sessions. We setup the \(t\)-Most important pages vector in \(t = 3\), which resulted in 141 training vectors.

A. Applying a Traditional Approach

We first applied a traditional WUM process approach for mining the web site i.e we represent the content by using frequencies. We chose a SOFM as our clustering algorithm although it is possible to apply any other like k-means.

We setup our SOFM in 36 neurons which are composed by an array that represents the \(t\)-Most important pages vectors. These means that one component of the array is a list of terms and for each term a random value in \([0, 1]\) is assigned. The second component is the time spent which it is also randomly assigned in the same interval. We need to be sure of the length of these random vectors is still 1 so we also normalize both components (terms vector and times). In previous works [10], [9], [11] we have tuned the traditional processes empirically in 50 epochs.

Several manners to extract clusters’ centroid form a SOFM exist for example: square vicinity, circular vicinity, hexagonal vicinity, etc. After about 28 hours of processing we successfully achieve a classification of the sessions. We discover only two main clusters which are shown in Table I.

<table>
<thead>
<tr>
<th>CLUSTER #</th>
<th>SESSION ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>{ 231 , 982 , 951 , 932 , 757 , 325 , 929 , 867 }</td>
</tr>
<tr>
<td>1</td>
<td>{ 965 , 503 , 979 , 878 , 962 , 365 , 765 , 786 , 966 }</td>
</tr>
</tbody>
</table>

After analyzing the web sessions inside the clusters with the sites’ expert help. We cannot obtain any useful knowledge about the visitors behavior. One of the reasons for this situation is the high amount of web pages in the cluster not topically related. There are pages that are about organizations inside the faculty, department of sports. General information about the school of engineering, academic schedule, extracurricular activities, social benefits for students, among others. This information doesn’t give a clue about the main goals or preferences of the visitors of the site.

B. Applying the Concept-based Approach

To begin with this approach we need to define the main concepts to be used in the process. Using the sites’ expert to perform this task we decided to use 12 general concepts instead of having hundreds of concepts. This allow us to better analyze the results of the method.

After having the concepts identified we need to define them using a dictionary, thesaurus and the expert criteria. Afterwards, we are able to setup the values of the membership function comparing two terms in a concept definition and assigning a value between 0 and 1. One limitation of our implementation is that we only are able to use single words not sentences or paragraphs which can improve greatly the results of the method.

Continuing the processes, we created a SOFM of similar characteristics to the one used above \((6 \times 6\) neurons, toroidal topology and 50 epochs) and also, keeping the same length of \(t = 3\) for the \(t\)-Most important pages vectors. However, we replaced the content vector \((\text{terms , tfidf})\) by the concept vector \((\text{concept , membership value})\) when constructing the SOFM.

The web site has about 5000 different non-stemmed words, which means that in the traditional approach the feature vectors are about this dimension. Therefore, a \(3\)-Most important pages vector has 15,000 terms. On the other hand, in the conceptual approach we only work with 12 concepts, which means that in this approach the feature vectors are only 12 features and a \(3\)-Most important pages vector has 36 features.

This huge difference affects directly the processing time of the classification process, in the concept-based approach the algorithm just take about 50 minutes to finish using the above settings.

As a results of the generalization stage, five sessions clusters were obtained. The sessions on these clusters are closest by conceptual meaning rather than text content. Also, the resulting clusters can be seen in detail on the Table II.

TABLE II

<table>
<thead>
<tr>
<th>CLUSTER #</th>
<th>SESSION ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>{ 515 , 757 , 549 }</td>
</tr>
<tr>
<td>1</td>
<td>{ 290 }</td>
</tr>
<tr>
<td>2</td>
<td>{ 878 , 256 , 567 }</td>
</tr>
<tr>
<td>3</td>
<td>{ 765 , 929 , 390 , 503 , 978 , 979 }</td>
</tr>
<tr>
<td>4</td>
<td>{ 425 , 982 , 220 , 600 }</td>
</tr>
</tbody>
</table>

We are able to see that there are more than just two classes in the visitor browsing behavior as common reasoning would suggest. Analyzing cluster 0 from traditional approach (Table I), we see that it have been split into at least three conceptual clusters. Now the session 757 is in cluster 0, session 929 is in cluster 3 and session 982 is in cluster 4. Besides, sessions
have been eliminated from the conceptual clusters. We find out that these erased sessions are not conceptually related with any of the clusters discovered in the conceptual approach. The same analysis is valid for cluster 1 of traditional approach.

Studying the internal pages from each conceptual cluster, we have smaller amount of web pages and pages that, as we expected, are more conceptually-related. A small example is shown in Table III where only three (from session 290) pages belong cluster 1, however, they are strongly related, these pages are about the list of distinguished students per year and we can see that students go to index page and then to these two pages. The analysis of the path that the student must follow to successfully reach the information on these session is not as straight forward as we show here. Several recommendations to modify the site structure can be derived from this analysis e.g. we can place a link to the section with the distinguished students in a top page.

### Table III

#### Real Pages Inside Conceptual Cluster 1

<table>
<thead>
<tr>
<th>Cluster # 1</th>
<th>Clusters inside this cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>index.html</td>
<td></td>
</tr>
<tr>
<td>escuela/LISTA_2004.html</td>
<td></td>
</tr>
<tr>
<td>escuela/LISTA_2003.html</td>
<td></td>
</tr>
</tbody>
</table>

Other example is the one presented by cluster 0, whose internal pages structure is shown in Table IV. The first session in cluster 0 (ID=515), contains 3 pages sorted out by maximum time spent (i.e the time spent on “baseorganizaciones.htm” is the highest of the 3 pages on the session). This page also contains links to information about organizations of the University of Chile, such as the library, The Moises Mellado Foundation, the Faculty of Mathematical and Physical Sciences, etc. The page “escuela/sobrelasescuela.htm”, contains information about the School of Engineering and Science. The “index.html” page is the main document where the visitor can reach any of this session pages. Similarly, studying the pages in session 757 we conclude that page “organizaciones/estudiantes.htm”, contains information about several different organizations that are related with extracurricular activities as well as contact information about each organization or group. For example, we have the rol-play gaming club, the students chorus, the yoga group, the karate group, among others. The other two pages in session 757 are frames to access this page. However, “infocalendarios.htm” is the main frame for the schedules (probably this page has been misplaced because the sessionization process is an heuristic). Although the meaning of the result is not altered. Finally, session 549 is the last in cluster 0. The “departamentos/index.htm” page links to all the Departments of the Faculty of Mathematical and Physical Sciences, such as Astronomy, Industrial Engineering, Computer Science, etc. As we can see, the results are strongly related. All the sessions in the cluster talk about different types of organizations within the Faculty.

Several specific recomendations that can be derived from this analysis are those studying the link structure of pages inside session 0 in Table IV and the Web site itself:

- A new page which contains the information about all organizations is needed: Every time the students want to obtain information about Departments or Faculty Organizations or Organizations like clubs or groups the visitor need to return to index.html where exist two different main sections to obtain information and not linked.

- The new page should contain all the information at once and classified under three categories: “Departments”, “Faculty Organizations” and “Extracurricular Activities Organizations”.

- Link all the related pages on each category to allow an easier visitor browsing experience.

With these recommendations it is possible to reach any page which contains information about organizations in less than 3 clicks. Also, we improve the presentation of the Web site’s index page fusing two main sections in one.

1) Conceptual Analysis of the Sessions: Until now we have successfully achieved a better classification of visitors sessions and we are able to analyze and recommend several content and structure changes to managers or web masters. Nevertheless, we still haven't performed a conceptual analysis over the results using software.

To extract the concepts inside the clustered sessions we compute the intersection of the concepts inside the documents on the sessions. We setup a threshold of 0.99, over it the concept was considered and below it was not as shown in Table V.

### Table V

#### Concepts Extracted from the Clustered Sessions

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td>1</td>
<td>N/A</td>
</tr>
<tr>
<td>2</td>
<td>NEWS/ADVERTISEMENTS, SEMINARS/EXTENTION, CLASSES MATERIAL, ADDRESS/CONTACT INFO.</td>
</tr>
<tr>
<td>3</td>
<td>REGULATION, CLASSES, TEST SCHEDULE</td>
</tr>
<tr>
<td>4</td>
<td>VACATION SCHEDULE, NEWS/ADVERTISEMENTS, SEMINARS/EXTENTION, EXTRACURRICULAR ACTIVITIES</td>
</tr>
</tbody>
</table>

On the resulting clusters 0 and 1 we could not find repre-
sentative concepts. This is probably due to a hard constrain such as the intersection of concepts in all documents and the high threshold established. More experiments are required for threshold tuning. Also other models like union of concepts in documents can be used.

An interesting case is cluster 0 from conceptual approach. After the experts’ analysis of this cluster its main concept is “Organizations”. However, there is not such concept defined in among the concepts used in the experiment. Nevertheless, Other concepts have a non zero value for example, SEMINARS/EXTENSION have a degree of 0.76 and EXTRACURRICULAR ACTIVITIES = 0.83, ADDRESS/CONTACTINFO = 0.86. However, non of these concepts are over the threshold of 0.99. This is one interpretation of the N/A value obtained in Table V. Similarily, for cluster 1 (which is specifically about distinguished students) we do not have any concept defined to fit the main topic of this cluster and the other concepts have a membership value lesser than 0.99.

The concepts on the cluster 2 show that visitors are interested in looking for news and advertisements probably about the the classes and classes material. If we add to this information the information about the web pages inside the class rooms assignation page, which is classified under the general concept of ADDRESS/CONTACTINFO on Table V. This way we improve and facilitate the analysis of the web usage mining results.

VI. CONCLUSION

The need of reducing the semantic gap from traditional web usage mining methods, motivated us to propose a concept-based usage mining process, using a fuzzy reasoning model to classify the visitors sessions.

The method was successfully applied to a real web site, showing better sessions classification and more meaningful clusters. Moreover, we are able to perform a semantic analysis of the sessions clustered to enrich the analysis.

Other interesting result is that many of the representative sessions in the traditional approach are missing and replaced by other ones. Analyzing these new sessions, we found that sessions in the conceptual approach are more representative in content and in meaning.

The proposed process also reduced the time spent in the generalization stage from more than 24 hours to only 50 minutes.

In the above examples we can also show a problem derived from concepts definition, which is the use of general concepts like ADDRESS/CONTACTINFO. This concept covers physical addresses, name of buildings, class room names, e-mails, etc. In the future we intend to refine the concepts used and to extend the concept base in order to obtain more meaningful results.

The development and testing of different ways of extracting the concepts from the clustered sessions also will be included in the future.

Moreover, we need to test the method in other web sites, preferably a general purpose web site (not a specialized one) and to develop new conceptual definitions to compare the results with the traditional approach.

Finally, we are presently working on an evaluation of the method, comparing it with other semantic usage mining methods. We need to further study performance, maintenance and results compared with these other algorithms.

REFERENCES