A hybrid system for concept-based web usage mining

Sebastián A. Ríos\textsuperscript{a,}\textsuperscript{*}, Juan D. Velásquez\textsuperscript{b}, Hiroshi Yasuda\textsuperscript{a} and Terumasa Aoki\textsuperscript{a} \\
\textsuperscript{a}Applied Information Engineering Laboratory, University of Tokyo, 4-6-1 Komaba Meguroku, Japan \\
\textsuperscript{b}Department of Industrial Engineering, University of Chile, República 701, Santiago, Chile

Abstract. A web site should be easy to browse by visitors. However, sometimes the reality is quite different. Situations like several unrelated topics in a single web page may lead to confusion and make harder to reach the information that the visitors are looking for. The design of the whole site (interface, content, structure, usability, etc.) is one of the most important aspects for any institution that wants to survive in the cyberspace. This work aims to extract new knowledge about how the visitors are using the web site and, in this way, to improve the site structure and content. We propose a hybrid system that combines the web usage mining (WUM) approach for analyzing the visitor browsing behavior and the conceptual classification of web pages for understanding the visitor text preferences in the site. The proposed system was tested in a real web site, which shows its effectiveness.

1. Introduction

The World Wide Web (WWW) is one of the most remarkable achievements for humanity. Nowadays it is impossible to image business or entertainment services without it. Moreover, every day many people are getting access to Internet and some of them are active part of the WWW by using Blogs,\textsuperscript{1} Vlogs,\textsuperscript{2} etc. Probably, all of them have used a search engine and some may have used the WWW to buy some products.

The WWW was created by the combination of Internet and HTML (Hyper Text Markup Language).\textsuperscript{3} In this way, anyone was able to publish his own documents easily. At first, programing in simple HTML code and uploading the files to the server; later, just using tools which help to create web documents (knowing nothing of HTML programming, File Transfer Protocol (FTP), among others). All this allowed the massive proliferation of web pages, web sites, portals, e-commerce systems, etc. Although, it also brought a problem: "The high amount of unlabeled and semistructured information which makes extremely hard to find out useful information" [41].

A challenging question for Managers and Organizations arises: how to develop a "good" web site?. In other words, how to build a site which gives valuable information to the visitors? How to create a visual design that allows a simple access to the information in the site without confusing the visitors? How to give the information, product or services that visitors need? These questions do not have simple answers.

Nielsen [21] discusses several web site usability problems and also establishes that effectiveness is strongly related with its usability. In addition to web interface it is possible to improve the visitors’ browsing experience enhancing the sites’ structure and content; but once more, how to do so? It is also a complex question.

Web mining area attempt to answer above questions or give some hints in order to perform such enhancements. At the beginning, the problem was tackled using text processing techniques on web documents in order to extract interesting patterns and eventually to deduce some conclusion about how to perform smart improvements in the sites’ textual content. This area was called web text mining (WTM) [22] and sometimes

\textsuperscript{*}Corresponding author. E-mail: srios@npseg.rcast.u-tokyo.ac.jp.
\textsuperscript{1}http://en.wikipedia.org/wiki/Blog.
\textsuperscript{2}http://en.wikipedia.org/wiki/Vlog.
\textsuperscript{3}http://www.w3.org/MarkUp/.
when it involves also web objects (images, videos, whole paragraphs, etc) it is called web content mining (WCM) [35]. The problem using these approaches is that visitors’ opinion is not considered, i.e. “what is interesting to a visitor” is not included in the analysis, only the textual or multimedia information inside documents.

Then, web usage mining (WUM) arose as a useful tool. In WUM, we analyze visitors’ browsing behavior stored in web logs. At the beginning, simple techniques to analyze browsing sequences were developed [17, 33]; then, more complex ones like in [3, 35]. Other researchers developed techniques which include the textual content to the sequence analysis to give results closer to visitors’ preferences [29, 44].

All of the above techniques have shown to be a real and effective aid for web site content and structure enhancements. However, standard WUM processes do not consider semantic information from web documents. This may lead to poor, erroneous, hard to interpret results. In the case of off-line improvements of a web site, the recommendations extracted using a standard WUM process may also be far from real visitors’ preferences. Semantic web mining have been arising in the last few years as a solution to these problems.

We propose an alternative way to perform a mining process which considers the semantic meaning of the documents using a concept-based knowledge discovery process [15] in addition to the visitors’ browsing sequences analysis. This work is focused in the off-line web site reconfiguration of content and structure to keep visitors using the site because they find the information they need easily.

We developed a hybrid process which combines fuzzy reasoning model [20] for web documents classification and a Self Organizing Feature Map (SOFM) [11] for sessions classification. In this way, we extract better clusters (closer to visitors’ preferences). We evaluate the conceptual classification of web documents and also the visitors’ sessions classification results using precision and recall as shown in [10]. We probe that our proposal gives better results: closer in meaning and closer to visitors’ preferences.

This paper is organized as follow. In Section 2, we introduced related works on two areas relevant for this work: Web Mining and document classification. In Section 3, the hybrid system proposal is presented and the evaluation methodology used. The application of the proposed system to a real world case is shown in Section 4. Finally, the main conclusions and future work are shown in Section 5.

2. Related work

The problem that we face is how to improve the web site textual content and structure in order to provide a better browsing experience to the visitor, to allow the maintenance of both, new and old visitors.

We aim to discover relations among visitors pages browsing sequences. In this way, we can study how many clicks a visitor needs to do to reach the information he wants. Using this information we can change the link structure of the site to allow faster access (less clicks) to these pages.

We also aim to find out relations among similar documents to those most browsed ones. In this way, we are also able to place in a more visible place other set of pages with similar information. We have done this using processes described in [29]. In this work, we aim to include semantic information to find out more meaningful relations among web pages. Then, we can perform a better enhancement on the structure of the site by placing these closer related pages in an easier to access place.

Also, it is possible to create new pages depending on the visitors’ browsing behavior or create new sections in the structure of the web site to hold a set of related pages. The new section(s) can be labeled easily with the conceptual information extracted from the proposed process [27, 31].

We can modify the content of the pages to be more interesting for the visitors based on the browsing and the content relatedness analysis.

2.1. Web mining techniques used for improving web site structure and content

Several authors [5, 18, 22, 34, 36] agree that web mining techniques can be divided in four basic subprocesses, which are:

1. Data selection: we choose the data sources, number of sites’ versions to use, time threshold for sessionization, etc. Also, we need to identify and define the relevant concepts which will be used in the concept-based web usage mining process.
2. Preprocessing: it is a very important process in order to gather good results. In these phase we clean the data. For web text cleaning usually stop-word lists, stemming algorithms, spell checking, etc. are used. As for web logs cleaning, we eliminate access errors (error 404, 403, etc.). Besides, every time a page is requested by the visi-
tor, the inner content (images, sound, videos) are also registered on logs, thus, they also must be erased. Afterwards, we perform a sessionalization process over cleaned logs. There are basically two nonintrusive sessionization approaches: time based and navigation based [35]. We used a time based heuristic, with threshold of 30 minutes.

3. Generalization: this subprocess consists of the application of one (or more) classification techniques to obtain patterns. In our case, since we do not have a model to fit the data we decided to use an unsupervised learning algorithm, thus, a Self Organizing Feature Map (SOFM).

4. Analysis/Evaluation of the results: We can compute cluster error rate and usually using the expert’s help [4] we establish if the patterns are useful or not. However, most of the works on web mining really mention the validity by using an expert, they relay in inner-clusters error (MAD), precision and recall, etc. We perform precision and recall and also we used a human expert to validate the web usage mining results. Besides, we also used precision and recall to evaluate the results of the concept-based classification process.

One of the most used techniques was WTM, where the main idea was the use of the vector space model (VSM) [32] for representing web documents and the use of some term frequency methods. A good example is shown in [24,38–40]. A natural evolution from WTM was the development of WCM, which also include web objects to the data used in the process [5]. Another good application of these techniques is the extraction of keywords for a better labeling of web documents [41]. In both approaches, it is commonly used the vector space model (VSM) [32] to represent documents’ text.

In the VSM, a web page is transformed into a feature vector, where each term in the document is a feature and it has a value associated. This value may be computed in several ways, however, one of the most popular way is to apply $TF*IDF$ (Term Frequency times Inverse Document Frequency), as shown in Eq. (1).

$$\psi_{ij} = f_{ij} \times \log(P/n_i)$$  \hspace{1cm} (1)

In Eq. (1), $f_{ij}$ is the number of times that the $i^{th}$ word in the $j^{th}$ page and $n_i$ is the number of documents containing the $i^{th}$ word. Therefore, the feature vector of a page is represented as: $p_j \rightarrow (\psi_{1j}, \ldots, \psi_{Wj})$, where $W$ is the number of different words in the entire collection of documents and $P$ is the number of different documents in the whole web site.

Afterwards, we need a method to measure the similarity among two documents represented by Eq. (1). Several different expressions exist for this purpose depending on the type of data to classify. A commonly used distance to compare two vectors is the cosine distance, which can be calculated as shown in Eq. (2), where $P_i$ and $P_j$ are two different feature vectors, $W$ is the number of components of the vectors and $\psi_{kj}$ is the $k$ feature of $P_j$.

$$PD(P_i, P_j) = \frac{\sum_{k=1}^{W} \psi_{ki} \psi_{kj}}{\sum_{k=1}^{W} (\psi_{ki})^2 \sum_{k=1}^{W} (\psi_{kj})^2}$$  \hspace{1cm} (2)

WTM and WCM don’t answer the question about how visitors are using our web site and how to make it more interesting for them. It just gives some hints about most common text and objects in the documents. This is why web usage mining (WUM) emerged as a way to discover interesting knowledge about visitors browsing behavior. Very good contributions can be read in [3, 18,19,33]. For example, Spiliopoulou propose the use of G-sequences to study visitors’ behavior, and also proposes the previous classification of three types of users. Then he uses the knowledge extracted for the creation of sequence rules [33].

Although, the above methods are very simple and they have showed their effectiveness to discover interesting patterns. Several researchers [2,8,9,13] realize that frequency-based methods lack of semantic meaning when they do not take into account the web pages content meaning. Therefore, the results are not semantically related and they are quite difficult to analyze and interpret correctly. In our case, interpretation is very important, since the improvements on the sites’ content and structure will be performed by the analyst.

There are several attempts to develop systems which include semantics to perform the web usage mining process. One of the first systems developed was the adaptive web site [23]. The author developed Page-gather algorithm which performs cluster mining to find related pages at a site using web logs. The resulting pages were used to create index pages as a non-destructive transformation of the site. The problem is that this system requires that all pages have been labeled using metadata.

Other work was developed in [8,9], where the authors developed an enhanced version of web logs called C-Logs (conceptual logs). The main idea is to generate profiles for the visitors based on the pages they visited. To do so, they used an ad-hoc ontology to classify web pages. Then, by using this information they labeled the visitors’ sessions and finally they generate recom-
mendations of next page style. A problem of these approach is that the generation and maintenance of it is very expensive. Other problem is that at the end of the process, 500 patterns were discovered. The analysis of so huge amount of patterns make this process not suitable to our purposes; we need a few meaningful clusters in order to analyze them and understand them, then use this information in the off-line reconfiguration of the site.

[2] labeled the visitors’ sessions using predefined web site categories of one web site. As the above approaches, she used the categories to classify the documents of the web site, then using G-sequences they rebuild the visitors’ sessions. The main drawback of this approach is that it requires that all pages have been labeled in one category before beginning the mining process.

2.2. Semantic classification of documents

Other important area related with our work is the semantic classification of documents where we can mention several techniques.

Li et al. [12] are working on automatic building and maintenance of domain-ontology used for the classification of the documents. They used Dempster-Shafer (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 Pattern-Evolving to discover the final ontology. Afterwards, an evaluation of results compared with two techniques: rocchio and DS model was performed. The method lack of the use of visitors’ sessions to obtain visitors browsing preferences and we need to define which sessions are positive examples in advance. The main problem of this approach is that requires to identify in advance the positive documents. In a web site with several hundreds of pages may be very difficult to do. Besides, web site pages are always been updated so the status may change form positive to negative or from negative to positive.

Other alternatives to the above approach are those based on fuzzy sets and fuzzy logic theory. Fuzzy sets theory was first proposed by Zadeh in [48] and allow us to represent the degree of belonging of an element to a set. Similarly, fuzzy logic can represent degrees of truth or falseness in contrast with traditional or crisp logic which is based in two values: true or false. By using linguistic variables and membership functions, we are able to represent concepts and the degree of these concepts inside a web page. Therefore, we can enhance the results of the mining process by including semantic relations between concepts and web pages.

Chau et al. in [6,7], developed an interesting work related with fuzzy logic and semantics for documents classification but not with sessions analysis. This research is focused in the semantics from multilingual documents written in Chinese and English. To classify the web documents she used fuzzy k-means algorithm for filter multilingual documents in topics regardless of the language. Documents are mapped using this topic classification and they are used to train a Self Organizing Map (SOM) to obtain a topic-oriented multilingual text classification. The problem of this approach is that require the application of two classification algorithms to obtain the topic classification of documents. This approach is good when working in information retrieval of the information, but it does not fit our needs. We need more precise information about visitors’ behavior. It is not useful label each document with hundreds of labels. Besides, this approach uses two clustering algorithms to obtain the classification of documents, which may take much time.

We based our work in Loh’s work presented in [15]. He proposed to use a fuzzy reasoning model to decide whether a concept is expressed by a web document or not. This way, after the application of the reasoning model, we have classified all documents by its concepts. To do so, we compute the possibility that a concept has to be on a text based on the weights obtained before and the terms’ membership values 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 ($CONCEPT \Rightarrow NC$) [16]. 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 represents the document.

We are going to use Loh’s approach to perform documents preprocessing phase, where each document will be automatically labeled with its concepts. Then, this information will be used to perform web usage mining. An advantage of this approach is its velocity, because, the reasoning model is reduced to the computation of the product of two matrices which can be computed very rapidly. Another advantage is that we are able to allow the expert to define which concepts he needs, or he wants to study, giving him more freedom and participation in the final results. The analyst may define
general or precise concepts as required. In a previous work [29] achieved a clustering of visitors’ sessions in about 24 h. This was due the huge number of features used (about 15,000) and the clustering method used (a SOFM) to perform the clustering. In this paper, we show that using the proposed approach we obtain a huge improvement in velocity (less than 50 minutes) and the quality of clusters disc overed was not diminished, in fact, it was greatly improved.

We have seen another approach called LSISOM for automatic labeling of documents [1], but it requires the application of two SOFM and the LSA decomposition at the beginning of the process. This technique is very computationally complex compared to the application of the fuzzy reasoning model. First, we need to perform the singular value decomposition over vectors of over 5000 words, then we need to apply a SOFM twice over shrunken vectors, which may lead to loss of information.

3. The hybrid system proposal

The hybrid system proposed for mining web data, combines two approaches: WUM and conceptual classification of documents. In this way, we are able to add semantics as concepts or topics in the patterns discovered from the visitors’ browsing sessions, and it is also possible to discover focused information about visitors.

An overview of the hybrid system proposed is shown in Fig. 1. The system uses web logs, web documents from our target web site and concepts which describe visitors’ browsing behavior.

The system has a “Web Logs Cleaner” module where errors are erased from logs, e.g. page not found and accesses to media files like pictures, sounds, etc. We eliminate those multimedia accesses because they don’t provide useful information.

Afterwards, the systems uses a “Logs Sessionizer” module, which transform web logs in visitor’s sessions. As explained before, we used a heuristic based on a time parameter of 30 minutes [44,46].

The proposed hybrid system for concept-based web usage mining

Similarly, we need to preprocess the text information. To do so, the system uses a “Web Document Cleaner” module which is an HTML parser and which is also capable to apply several text filters like: stop-word lists, stemming (English, French, Spanish and other 7 languages).

The sites’ expert(s) needs to identify which are the main concepts to focus the attention and the analysis. This task is performed in the “Concepts Identification” module. Next, we automatically extract the concepts definition using the “Concepts Definition” module which uses the help of a Dictionary and Thesauri. The concepts definition may be enhanced by the expert(s) also using this module. The definition consist in several lists: a set of synonyms, antonyms, quasi-synonyms and quasi-antonyms. In the end, these terms lists are stored in our “Concepts Base”, shown in Fig. 1.

The filtered text of each web page is classified using the “Fuzzy Reasoning” module, where we used Nakanishi’s fuzzy reasoning model [20] in Eq. (5) and the concepts previously stored in the “Concepts Base” (see Fig. 1). The main goal of the conceptual classification of documents is to automatically discover concepts or topics that describe each document. The process ends with the entire set of web pages classified by concepts. We are partially using the process proposed by Loh et al. in [15] to obtain patterns richer in meaning. In this way, we make easier their analysis and extraction of useful knowledge to improve the web site.

The conceptual classification is analyzed by the expert(s) to check the quality of the classification. If his/her criteria is not satisfied, the concepts base needs to be improved. We have two sources of miss classification: first and the most common error is that the experts forgot to identify a concept. The second reason, which is lesser frequent, is that we miss to include a term in the definition of the concept. If the conceptual classification is “ok” then we are able to apply the “SOFM” module for clustering.

Once we have the documents successfully classified by its concepts and the sessions reconstructed, we combine them in the “SOFM” module (see Fig. 1) and then we cluster this data.

Finally, the system provides all information needed by an analyst. We are able to show the SOFM as a 3D grid for easy visualization of clusters discovered. We provide the clusters centroids and associated documents in a vicinity using the reverse cluster analysis (RCA) [24–26,30], Voronoi regions and now we can also provide the conceptual relations among web pages. Moreover, we are able to automatically label similar visitors’ sessions to allow a better understanding of these sessions by the analyst even if the documents text is different or is in a different section of the web site. This information makes easier, for a human analyst, the process of understanding the patterns found and decide how to improve the site content and structure as mention in Section 2.
3.1. Fuzzy logic for conceptual classification

Linguistic variables (LV) values are not numbers but words or sentences in natural language. These variables are more complex but less precise. Let $u$ be a LV, we can obtain a set of terms $T(u)$ which cover its universe of discourse $U$, e.g. $T(temperature) = \{\text{cold, nice, hot}\}$ or $T(pressure) = \{\text{high, ok, low}\}$.

In order to use LV for conceptual classification, we assume that a document can be represented as a fuzzy relation $[\text{Concepts} \times WP]$ also called $[C \times WP]$. Which is a matrix where each row is a concept and every column is a web page. To obtain such matrix we can rewrite this relation in a more convenient manner in Eq. (3) [15]. In this expression we call “Terms” the words that can be used to define a concept and we write “WP” to refer any word inside a Web Page. In Eq. (3) the symbols “$\times$” and “$\otimes$” represent the fuzzy relation and fuzzy composition respectively.

$$[\text{Concepts} \times WP] = [\text{Concepts} \times \text{Terms}] \otimes [\text{Terms} \times WP] \quad (3)$$

As defined above, let $P$ the total amount of web documents in the whole Web site, $W$ the total number of different words among all documents and $K$ the total number of concepts defined for the web site. Then we can characterize the matrix $[\text{Concepts} \times WP]$ by its membership function shown in Eq. (4), where $\mu_{C\times WP} = \mu_{C\times T\otimes T\times WP}$ represents the membership function of the fuzzy composition in Eq. (3). The membership values are between 0 and 1.

$$\mu_{C\times WP}(x,z) = \begin{pmatrix} \mu_{1,1}, \mu_{1,2}, ..., \mu_{1,P} \\ \mu_{2,1}, \mu_{2,2}, ..., \mu_{2,P} \\ \vdots & \vdots \\ \mu_{K,1}, \mu_{K,2}, ..., \mu_{K,P} \end{pmatrix} \quad (4)$$

There are several alternatives to perform the fuzzy composition; Nakanishi et al. performed a study between six different reasoning models [20]. One important issue that must be considered is that even if some terms are not present in a web page, the degree of expressing a concept should not suffer alterations. This is a reason to use Nakanishi et al. compositional rule Eq. (5).

$$\mu_{Q\circ Z} = \bigvee \{\mu_{Q}(x,r) \land \mu_{Z}(r,y)\} \quad (5)$$

Let $Q(U,V)$ and $Z(V,W)$ be two fuzzy relations which share a common set $V$. Let $\mu_{Q}(x,r)$ with $x \in U \land r \in V$ and $\mu_{z}(r,y)$ with $r \in V \land y \in W$ membership functions for $Q$ and $Z$ respectively then we can write the compositional rule as shown in Eq. (5). Where $\bigvee$ is the limited Sum $= \min(1, x + r) \land \land$ is the algebraic product $(x \ast r)$. 

---

**Fig. 1.** The proposed hybrid system for concept-based web usage mining.
3.1.1. Identification and definition of concepts

In order to apply the above proposal, we need to begin identifying the relevant concepts for the study. It is important to remark that we do not look for a conceptual classification for information retrieval, which may include thousands of concepts and terms in order to retrieve all relevant documents regardless of the keywords used in the user’s query. We are looking for a concepts to describe the browsing behavior of the visitors. To do so, we make use of experts’ knowledge whom identify which are the most interesting concepts to describe visitors’ behavior in the web site.

Subsequently, we use the help of a thesaurus and dictionaries to extract terms to define the relevant concepts i.e. to express every concept like a list of terms (assuming that a concept is a LV). We used synonyms, quasi-synonyms, antonyms, etc.

We realize that several important terms were not present in dictionaries nor thesaurus. In our particular case, the site has a system called “U-Cursos”, which provides important information of the classes to the students or another system called “U-Agenda”, which is the students’ agenda. Both are of course “Personal Services”, however they are not present in dictionaries or thesauri; nevertheless the expert can include it as a term that defines a concept. In our experiments we used “Personal Services” as concept, thus, these words “u-cursos” and “u-agenda” and other similar systems were added. Other example occurs when defining the concept “Organizations”. Organizations in the faculty usually use an abbreviated name i.e. Department of Industrial Engineering (DII – in Spanish), Department of Computer Science (DCC – in Spanish), among others. We are able to include these terms in the definition of the concept.

Afterwards, we need to define the membership values for the fuzzy relations \([Concepts \times Terms]\) and \([Terms \times WP]\). We used relative frequency of words in a web page to represent the membership values of matrix \([Terms \times WP]\).

More complex is the definition of \([Concepts \times Terms]\) values. We performed this operation by asking the expert to assign the degree of a term to represent a concept. To do so, he compared two terms each time and gave a value between 0 and 1. For example, a synonym can receive a value near 1; a quasi-synonym, may receive a value near between 0.65 and 1; an antonym can be set to 0, etc. This method is an indirect method with one expert.

Finally, we obtained the fuzzy relation \(\mu_{C \times WP}(x, z)\) by applying Eq. (5). In Table 1 we present a column of matrix \(\mu_{C \times WP}(x, z)\), which represents the conceptual classification of a services page. From this Table we can say that the page “services.htm” have a strong relation with the concept “General Services” and “Personal Services” but almost nothing with the other concepts.

On the other hand, the work required to define hundreds of concepts may be tremendous, similar to generate the metadata for each document. However, we need to establish the difference between web sites and web portals. A web site is a collection of web pages related by an specific concept or topic. For example, the web site of the Department of Industrial Engineering, or the web site of the chess club, or the site of the cinema. This way, there are not many interesting or important concepts to work with in order to enhance the web site. On the contrary, a web portal is a collection of web pages which are not related by a particular theme, e.g. Yahoo, Terra, Lycos, AOL, etc. mining of portals’ data may be harder because there are hundreds of concepts which should be included to perform the analysis. Therefore, our approach is feasible to be applied to study web sites. In this way, we are able to incorporate the expert criteria in the clustering phase, obtaining better results as shown later.

On the other hand, we are able to apply an automatic approach for automatic fuzzification. The problem of these approaches is that usually they are designed to include all possible concepts. They take all words or the most repeated words in the corpus as concepts. Then they go to a thesaurus and use an algorithm for automatic fuzzification. Then the algorithms incorporate all possible concepts which produce results that are very complex to understand. Most of these approaches are used for information retrieval. This way, they should use hundreds of concepts with hundreds or thousands of terms of retrieve all relevant documents when the user enters a query. This is not our case, our main goal is to perform web site enhancements, not retrieve documents based on a query.

<table>
<thead>
<tr>
<th>Concepts</th>
<th>(HC \times WP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Services</td>
<td>0.88</td>
</tr>
<tr>
<td>Personal Services</td>
<td>0.72</td>
</tr>
<tr>
<td>Regulations</td>
<td>0</td>
</tr>
<tr>
<td>Test Schedule</td>
<td>0.12</td>
</tr>
<tr>
<td>Vacation Schedule</td>
<td>0.01</td>
</tr>
<tr>
<td>Classes</td>
<td>0</td>
</tr>
<tr>
<td>Evaluative Activities</td>
<td>0</td>
</tr>
</tbody>
</table>
We wish to discover a documents’ classification that make sense to the visitors. Therefore, we used few concepts which are really important (based on expert criteria) to perform studies on the visitors behavior and extract precise (or general) information, easy to understand, and meaningful to enhance the web site. When working with less than 30 concepts, the task of fuzzification is not so complex and 30 concepts represents a large number if we consider how a visitor uses a web site.

3.2. Self Organizing Feature Map (SOFM)

The generalization process consists in the application of pattern recognition techniques, machine learning or Artificial Intelligence [37]. There exists two main classes of these which are: supervised and unsupervised algorithms. Unsupervised algorithms do not require to establish an output set a priori, however, this kind of algorithms are more complex to solve than supervised ones. Unsupervised algorithms are also called clustering algorithms.

We chose a SOFM of the Kohonen type [11] which is an unsupervised algorithm to extract significant patterns from the visitors sessions. A toroidal topology was set up to maintain space continuity [30,41]. Maintaining the continuity of the space when clustering is very important to obtain good results. When a visitor is browsing a Web Site, he/she is constrained to web site pages and structure of the site, visitors are not randomly jumping from one page to another; a link is needed. It is also required that the target page has content related and relevant to visitors’ browsing goals in a specific session.

In order to propagate the learning to the network, we used a Gaussian function that depends on the distance between the centroid neuron and the rest of the neurons, as shown in Eq. (6).

$$
\Theta_{ci}(t) = \alpha(t) \cdot \exp \left( \frac{\|x(t) - c_i\|^2}{2 \sigma^2(t)} \right) \tag{6}
$$

The neurons’ learning is described by Eq. (7), where $\xi(t)$ is the example shown to the network on the epoch $t$. If we are classifying documents, then $\xi(t)$ is the web page feature vector. If we need to classify sessions, then $\xi(t)$ are the visitors sessions sequences. Therefore, $\chi_i(t+1)$ represents the neuron $i$ on the epoch $t+1$, which is obtained based on $\chi_i(t)$ (the neuron’s state on the past epoch $t$).

$$
\chi_i(t+1) = \chi_i(t) + \Theta_{ci}(t) \cdot (\xi(t) - \chi_i(t)) \tag{7}
$$

3.2.1. Similarity measure for concept-based usage mining

It is important to remark that to classify a set of web pages $P$, we need to present all of these pages to the algorithm. If we need to process a new version of the site which contains updated, deleted or new pages, then we need to apply the algorithm again. This is called the Domain Restriction.

In the above expressions, Eqs (6) and (7), we are implicitly using a similarity measure to compare the examples with the neurons. However, we cannot use the expression in Eq. (2), because, it is designed for documents comparison not for visitors’ sessions comparison.

Based on Eq. (2) from [47], we developed a similarity measure to compare visitors sessions as in Eq. (8). The main idea of this expression is the combination of the visited pages sequences (these sequences are ordered but by time spent not by access) with the content and time spent on each page.

We assume that the visitor spend more time on a web page that it is more interesting to him, although, this statement may be of huge controversy. The main problem of this assumption is the fact that many visitors when using the site for first time, get lost incurring in high browsing times of pages which actually are not interesting to them. In the case of Velasquez work [42, 43,45], he applied this assumption to bank’s sites to explain the behavior of casual visitors which probably is an error if we think about the lost visitors. On the other hand, he also used the above assumption to study a university site [40,41], in this case the visitors browsing the site for first time is rather small compared to the visitors (not registered) who belong to the University and who are familiar with the site. In this case, it is reasonable to use this assumption.

Based on the pages preferences assumption we are able to define the $i$-Most important pages vector in Definition 1 to explain the visitor browsing behavior.

Definition 1. ($i$-Most important pages vector) Let

$$
\sigma^v = [(\rho^v_1, \tau^v_1), \ldots, (\rho^v_i, \tau^v_i)]
$$

be the $i$-Most important pages vector, where the pair $(\rho^v_i, \tau^v_i)$ represents the $i^{th}$ most important page (text content) and the percentage of time spent in it, within a visitor session $v$ [40,41].

The internet visitors similarity (IVS) in Eq. (8) compares two $i$-most important pages vectors $V^i$ and $V^j$ for two different sessions $i$ and $j$. A $i$-most important pages
vector is a sequence of the i-pages where the visitor spent most of his time during the session. If \( t = 5 \) then \( V^i \) and \( V^j \) have 5 component each one. Furthermore, for a session \( i \) each component \( k \) from \( V^i \) is a pair of (Text Content, Time Spent) also called \((V^i_p(k), V^i_\tau(k))\) that represent the k-web page, where text content \( V^i_p(k) \) is represented as a vector \((word_i, weight_i)\). Besides, Eq. (8) also makes use of function \( PD() \) which compares the content of the pages on the sessions (\( \rho \) component from \( V \)). The \( \min(.) \) term is the relative time spent between the pages being compared (\( \tau \) component from \( V \)).

\[
IVS(V^i, V^j) = \frac{1}{t} \cdot \sum_{k=1}^{t} \min\left(\frac{V^i_\tau(k)}{V^j_\tau(k)}, \frac{V^i_p(k)}{V^j_p(k)}\right) \cdot PD(V^i_p(k), V^j_p(k)) \tag{8}
\]

A problem of this similarity is that it was designed to be used in traditional visitors sessions classification, which means that it is based on comparison of TFIDF values. Giving poor results when trying to establish interesting recommendations from the visitors’ point of view.

Another problem with Eq. (8) is how to represent word importance. In fact, it depends on the word frequency on a document. Thus, if a word appears several times in the text it is an important word and if only appears few times its importance will be very low. If an important word does not appear several times we obtain erroneous results. For example, a word belonging to document’s title which does not appear a reasonable amount of times in the remaining text could be considered not important. Hence, the sessions clustering process will produce low quality clusters with mixed information, which is very difficult to analyze and which is probably erroneous.

Moreover, similarity measure showed in Eq. (8), assumes that relatedness is based on word frequency. However, a good writer probably will try to change words in order to be grammatically correct, i.e. synonyms and quasi-synonyms will be used.

Now, let us explain the worst problem that may occur when using Eq. (8). If a document is about concept “B” which is not related to other concept “A”, then in document, “B” will appear “B” or as synonym several times and also may appear “A” with some frequency (usually when comparing two or more nouns). Using this similarity, these documents may be very similar to other documents that talk about “A” and “B”, which is not correct. This also happens in sites like universities ones where words like “students”, “professors”, “courses”, etc. appear in many pages of unrelated main topics and the web pages have few lines of text to be able to successfully differentiate topics by term frequency.

We have used this similarity in previous works [30, 31]. However, the above reasons forced us to redefine the bases of Velasquez’s similarity to be able to perform a conceptual classification of visitors sessions. The redefinition is shown in Definition 2.

The conceptual classification of sessions is much more complex than the simple process using the words in the free text of the web pages. We still need the textual information and time spent. However, we also need the conceptual classification of each document (represented by \( \varphi \) component of Conceptual \( \iota \)-Most important pages vector in Definition 2).

**Definition 2.** (Conceptual \( \iota \)-Most important pages vector) Let

\[
\Omega'_\iota = [(\rho^i_1, \tau^i_1, \varphi^i_1), \ldots, (\rho^i_{\iota}, \tau^i_{\iota}, \varphi^i_{\iota})]
\]

be the Conceptual \( \iota \)-Most important pages vector, where three components \((\rho^i_{\iota}, \tau^i_{\iota}, \varphi^i_{\iota})\) represent the textual content, time spent and concept classification (a list of concepts) for the \( i^{th} \) most important pages within the visitor session \( i \).

Performing this basic change we are able to maintain the form of Eq. (8), however, instead of using raw text, we are processing concepts and membership values which express the degree that each concept has to express the specific document (see Table 1).

Let \( S^i \) and \( S^j \), two conceptual \( \iota \)-Most important pages vectors, where the concepts component is \( S^i_\varphi \). In this way, we can rewrite Eq. (8), to include the semantics. We defined the Concept-based Internet Visitors Similarity (C-IVS) as shown in Eq. (9).

<table>
<thead>
<tr>
<th>Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nine relevant concepts were identified</td>
</tr>
<tr>
<td>GENERAL SERVICES</td>
</tr>
<tr>
<td>SCHEDULES</td>
</tr>
<tr>
<td>NEWS/ADVERTISEMENTS</td>
</tr>
<tr>
<td>ADDRESS/CONTACT INFO.</td>
</tr>
<tr>
<td>DISTINGUISHED STUDENTS/AWARDS</td>
</tr>
</tbody>
</table>
Fig. 2. Precision and Recall for Concepts Classification.

\[ C - IV S(S^i, S^j) = \frac{1}{l} \sum_{k=1}^{l} \min \left( \frac{S^i \tau(k)}{S^j \tau(k)} \right) \cdot PD(S^i \phi(k), S^j \phi(k)) \]  

In previous works, we have tested that the new similarity improved greatly the results of the mining process and we have compared it with standard processes based on the Velasquez similarity in [27,28,31]. In the next Section we show the methodological evaluation of the proposed approach, which has not been presented.
in previous works.

3.3. Analysis and process evaluation

The evaluation of this kind of systems is very complex [10] established an evaluation framework to compare two recommender systems. According to them, to evaluate the process we need to decompose the system in subprocesses and then, each subprocess can be studied. He proposed five evaluation methods for different purposes and he mapped them to each part of the KDD process. For example, to evaluate the learning process we can separate the data in two groups. Use one group for training and the other for evaluation. However, this only evaluates how well the patterns were learned. We still do not know anything about the utility of the patterns discovered.

We are going to perform an evaluation using precision in Eq. (10) and recall in Eq. (11) for the conceptual classification of documents [10,14]. Similarly, we applied precision for the evaluation of the patterns discovered. In addition, we performed a traditional web mining technique to compare its results with the conceptual approach.

\[
\text{Precision} = \frac{\text{Items Correctly Classified}}{\text{Total Classified Items}} \quad (10)
\]

\[
\text{Recall} = \frac{\text{Items Correctly Classified}}{\text{Total Correct Items in the Class}} \quad (11)
\]

However, we still need to perform a utility evaluation based on a survey performed to visitors of the site. In [10], the author explains how the utility test is the only way to evaluate the whole process of recommendation by measuring if discovered patterns are useful or not.

Another important requirement is to measure the quality of the results by performing a study of the discovered patterns by the expert. The expert must study each cluster; if it contains many unrelated concepts then the cluster should be discarded, otherwise, it is accepted. Then, we can calculate the percentage of accepted clusters.

4. Experimental results

We performed the experiments onto the web site of the School of Engineering and Science of the University of Chile. The version of the site used this time contained 187 web pages.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Avr. Precision</th>
<th>Avr. Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>32.76%</td>
<td>100.00%</td>
</tr>
<tr>
<td>0.100</td>
<td>78.78%</td>
<td>78.52%</td>
</tr>
<tr>
<td>0.200</td>
<td>90.58%</td>
<td>75.34%</td>
</tr>
<tr>
<td>0.300</td>
<td>100.00%</td>
<td>62.56%</td>
</tr>
<tr>
<td>0.400</td>
<td>100.00%</td>
<td>45.70%</td>
</tr>
<tr>
<td>0.500</td>
<td>100.00%</td>
<td>34.51%</td>
</tr>
<tr>
<td>0.600</td>
<td>100.00%</td>
<td>29.96%</td>
</tr>
<tr>
<td>0.700</td>
<td>100.00%</td>
<td>18.14%</td>
</tr>
<tr>
<td>0.800</td>
<td>100.00%</td>
<td>9.61%</td>
</tr>
<tr>
<td>0.900</td>
<td>100.00%</td>
<td>6.00%</td>
</tr>
<tr>
<td>1.000</td>
<td>100.00%</td>
<td>4.15%</td>
</tr>
</tbody>
</table>

4.1. Applying the conceptual classification approach

First, we identified the relevant concepts as mentioned in Section 3.1.1. We discovered that important concepts were omitted and some concepts used were not important (based on some experiments performed before [27,31]). In this way, we enhanced our concepts’ list, that resulted in a new nine concepts’ list shown in Table 2. In this work, we are using 9 concepts, while in [27,31] we used 12 concepts.

Afterwards, we used a Spanish dictionary, thesaurus and the expert’s help to define the terms to characterize each concept in Table 2. Our program download, parse the html and write the terms to a text file.

We performed the evaluation of the conceptual classification using precision and recall, in Eqs (10) and (11). These results are shown in Figs 2 and 3. We can see that we achieved a very good precision for all concepts as shown in Table 3. For thresholds over 0.3, we obtained 100% precision. In our process, it is more important precision than recall, since the session of a visitor determines which pages are being used. However, with a threshold of 0.3, we obtained 62.56% of recall. These results show that the concept based approach achieved an excellent classification of the web site pages.

4.2. Sessions reconstruction

Once the documents have been classified, we can rebuild the sessions stored in web logs as shown in Section 3. We set up a time of 30 minutes as the maximum time for each session.

We choose about one month of web logs and after cleaning and sessionization process we obtained about 1023 sessions. However, when \( i = 3 \) most important pages vector was setup to \( i = 3 \), only 141 training vectors remained. The reason for this is that only 141 sessions
have more or equal than 3 visited pages (as explained above). Most visits to our target web site are from students who already know the site structure.

4.3. Semantic visitor’s sessions classification

The next step is the utilization of the SOFM from Section 3.2 to classify the visitors’ sessions. To do so, we used a SOFM of the Kohonen type [11] with toroidal topology in order to maintain the continuity of sessions space [29,41].

We have finally tuned our SOFM in $6 \times 6$ neurons and 50 epochs based on previous experiments [27,31].

In a traditional web mining approach, each neuron is a list where each element is a vector with two components: the time spent on a visited page and a vector of words with the frequency of these words on a page. However, we have replaced this vector of words by a vector of concepts and membership values which allowed the generation of conceptual web usage mining.

The results of the process are shown in Table 4. We obtained four clusters applying the concept-based similarity in Eq. (9). Every session ID has a corresponding list of three real pages.

In order to be able to evaluate the process we used precision in two ways: first, to the whole session classified ($Prec_s$); second, to all pages in clustered sessions for each cluster ($Prec_d$), as shown in Table 5.

We can observe that results of $Prec_s$ are higher than $Prec_d$, since, in $Prec_s$, we used the whole session, i.e. we do not care if the pages inside the session were the home page or other non-relevant pages or not re-
4.4. Traditional frequency-based approach

To compare the results gathered from the concept-based approach, we performed a traditional web usage mining using the same parameters used before (6 \times 6 SOFM, toroidal topology and 50 epochs). We used the Velasquez similarity in Eq. (8) without modification. These means that we used TFIDF in Eq. (1) to setup the frequency of each word on the documents. We also used a stemming process and stop word list in order to obtain better results. After this process, we obtained a vector of near 5,000 words.

The classification process took about 28 hours of processing until we successfully achieved a classification of visitors’ sessions. We discovered only two main clusters which are shown in Table 6.

By undertaking this process, only two clusters were discovered. To evaluate the precision of the clusters discovered we also used $P_{rec_c}$ and $P_{rec_d}$ as shown in Table 7.

Comparing Tables 5 and 7 it is possible to realize that we achieved a huge improvement in precision of concept-based usage mining compared with the traditional approach. Besides, we discovered more clusters which may give better information to an analyst to improve the web site.

4.5. Analysis of the discovered patterns

The analysis of clustered sessions is one of the most important tasks. The main idea is to discover interesting usage patterns in order to improve the site structure, give on-line recommendations, etc.

In the case of the traditional technique, we discovered only two clusters, although, the experts’ knowledge was expecting more that two clusters in the browsing behavior. When studying, in deep, each session on each cluster, we realize that extracting some useful knowledge was very hard. One of the reasons is that there are many web pages which are partially related or not related at all with the main trend of pages in a cluster. In cluster $ID = 0$ from Table 6, we have a main trend about “Organizations inside Faculty”, although, many pages are not about this topic. There are pages that contains information about other topics like: General description about the school of engineering, academic schedule, extracurricular activities, social benefits for students, among others. These makes difficult to say with confidence what are the most important topics for the visitors that are being expressed by this cluster.

This fact can be also reflected in the values for precision obtained in the traditional mining in Table 7. Similarly, for cluster $ID = 1$, the main trend of this cluster is “distinguished students and awards”, although, many other sessions that have no relation with this one are included in the results.

On the other hand, if we analyze in detail the resulting sessions on each cluster from concept-based usage mining we discovered a very high degree of relationship among all sessions on the clusters and all pages inside those sessions. This is why we obtained a very high value for precision in Table 5. As an example, we are going to take cluster $ID = 3$ from Table 4, and we will expand it as shown in Table 8. It is possible to observe that the pages inside session $ID = 6$ are talking about the distinguished students (pages “escuela/LISTA_2001.html” and “escuela/LISTA_1996.html”). The page “servicios/bienestar.htm” is about scholarships information.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>$P_{rec_c}$</th>
<th>$P_{rec_d}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.75</td>
<td>0.58</td>
</tr>
<tr>
<td>1</td>
<td>0.67</td>
<td>0.89</td>
</tr>
<tr>
<td>2</td>
<td>0.80</td>
<td>0.73</td>
</tr>
<tr>
<td>3</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 6

<table>
<thead>
<tr>
<th>Cluster</th>
<th>$P_{rec_c}$</th>
<th>$P_{rec_d}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.63</td>
<td>0.46</td>
</tr>
<tr>
<td>1</td>
<td>0.33</td>
<td>0.41</td>
</tr>
</tbody>
</table>

Table 7

By clustering the home page reduce the value of sessions pages. In this case, for example, the access to if every page is related or not with the other clustered sessions we decompose each session on a cluster and we studied others sessions or not. On the other hand, for the whole session in the cluster is very related to the main concept of the cluster; we only consider the analysis of the session as a whole, i.e. if this page is not related at all with the main concept or there is no link to a page that is related to the main concept. On Table 5, we can also see that precision $P_{rec_c}$ and $P_{rec_d}$ are both very high for all clusters discovered.

### Table 5

<table>
<thead>
<tr>
<th>Cluster</th>
<th>$P_{rec_c}$</th>
<th>$P_{rec_d}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.75</td>
<td>0.58</td>
</tr>
<tr>
<td>1</td>
<td>0.67</td>
<td>0.89</td>
</tr>
<tr>
<td>2</td>
<td>0.80</td>
<td>0.73</td>
</tr>
<tr>
<td>3</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

### Table 6

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

### Table 7
Next, session $ID = 326$ is the sequence of index pages which lead to the main page of distinguished students ("escuela/a_destacados.htm"). We may observe that sessions 3 and 326 are strongly related and they are talking about "distinguished students". If we see Table 5, the values for precision are 1.

Other example is Cluster $ID = 1$ from concept-based approach shown in Table 9. The pages inside session $ID = 390$ are "index.html" of the main site and "reglam.htm" which is the index page for the faculty's rules. On session $ID = 390$ we have "escuela/sobrelaescuela.htm", which talks about general information about the faculty which is linked to "reglam.htm"; the page "index.html" that is the index of the site which is also linked to "reglam.htm" and the last page on this session is "infocalendarios.htm", which is a frame for faculty's schedules. The last session to study is session $ID = 978$, where the pages "reglamento2991.html", "reglamento_de_alcoholes.html" and "reglaconducta.html" are all about specific reglaments (general rules, alcohol reglaments, studies rules). We can see that two out of three sessions are strongly related (sessions $IDS = 503$ and 978) but session $ID = 390$ is not very related. This can be seen in the precision results for cluster $ID = 1$ in Table 5, where $Prec_a = 0.67$, although the value for $Prec_d$ is higher because in session $ID = 390$ two out of three pages are related with the main topic of the sessions which is "Reglaments".

The proposed system improves the quality of patterns found. We can observe that with the proposed system the patterns are more related than using a traditional approach. Moreover, the resulting patterns have more relation with visitors' interests. For example, in Table 9, we can extract the concepts of the cluster $ID = 1$ which is "Reglaments". As the reader may observe, the whole cluster can be reduced to a concept, which is that visitors are looking for reglaments more than the rest of 10 different sections that the web site has. In this way, the web manager now has a specific guideline to improve the web site, i.e., to easily reach reglaments from the top page instead of going five levels from the top page. Several improvements to the structure, as the one explained, or to the content of the site may be derived by using similar analysis.

Other improvement of concept-based approach is that now performing the analysis such as explained above is very intuitive. However, in a traditional frequency based approach, performing such kind of analysis is impossible. The reason is that in concept based approach every page is labeled by a list of concepts; therefore it is easy to extract them. In traditional approaches, we only have a long list of words and frequency (in our case, almost 5000 words), which makes impossible the simple extraction of the concepts (or main trend) from those clusters.

### 4.6. Discussion and future work

The previous sections explained the difference between web sites and web portals. Since we are working with a web site, we only needed few concepts to perform our analysis. These few concepts describe the way visitors use the web site. We have performed a utility survey to students in past work [30]. From this study, we obtained the first 12 concepts used in a previous concept-based classification work. Therefore, we can support the idea that we do not need hundreds of concepts to describe the behavior of most users in the site. However, if the analyst wants to include hundreds of concepts, our proposal may be difficult to be used because we need to set up the membership values by hand.

To overcome this problem, we are currently working on the implementation of a fuzzification based on Fu-
types of precision based classification results. Similarly, we applied two precision and recall for the conceptual classification. The results of this exploration study in a web portal, since we have hundreds of possible concepts to work with. Afterwards, our proposal can give better information for an off-line enhancement of the site.

We still need to apply the hybrid system to different types of web sites in order to test the proposal in different domains. Also, we need to perform an utility test to evaluate the results of the whole process.

5. Conclusions

Traditional web usage mining techniques are based on frequency-based methods which leads to a semantic gap in the results. This makes hard to extract any knowledge from these results. To reduce the semantic gap, we proposed a hybrid-system which combines a fuzzy reasoning model with a self organizing feature map (SOFM). In this way, we are able to classify a visitor session by its conceptual degree rather than the frequency of words.

We performed experiments onto a real-web site to show the effectiveness of the proposal. The results of the concept-based usage mining has shown better session classification results, which are easier to interpret by the business expert.

We have performed an evaluation of the results using precision and recall for the conceptual classification of web documents. We achieved very good concept-based classification results. Similarly, we applied two types of precision $\text{Prec}_c$ and $\text{Prec}_d$ to evaluate visitors sessions classification results. The results of this evaluation showed very good precision on the clusters discovered, in excess of 69%.

We compared our approach to a traditional terms based web usage mining. We proved that our proposal is better than the traditional approach for the usage mining of web sites. We obtained a 100% more clusters with our approach and all new clusters were accepted by the analyst.

The proposed approach also reduced the time spent in the generalization stage from more than 24 hours to only 15 minutes, by reducing the features used for training the SOFM. In addition, as explained above, the quality of the clusters was greatly improved.

Acknowledgments

We thank to Professor Eduardo S. Vera from the Electrical Engineering Department of the University of Chile and Pablo Roman, PhD(c) of Industrial Engineering Department of University of Chile, for their continuous support and useful comments while improving this paper. Also, we thank to the Millennium Scientific Nucleus on Complexes Engineering Systems (Chile), which has partially funded this work.

References


[43] J.D. Velásquez, H. Yasuda and T. Aoki, Combining the web content and usage mining to understand the visitor behavior in a web site, in: *Procs. 3th IEEE Int. Conf. on Data Mining*, Melbourne, Florida, USA, November 2003, 669–672.


