Exploiting User Interests to Characterize Navigational Patterns in Web Browsing Assistance

Daniela GODOY and Analía AMANDI
1 ISISTAN Research Institute, UNICEN University
Campus Universitario, Paraje Arroyo Seco
CP 7000, Tandil, Bs. As., Argentina
2 CONICET, Consejo Nacional de Investigaciones Científicas y Técnicas
CP 1033, Capital Federal, Bs. As., Argentina
{dgodoy,amandi}@exa.unicen.edu.ar

Abstract  In order to be capable of exploiting context for pro-active information recommendation, agents need to extract and understand user activities based on their knowledge of the user interests. In this paper, we propose a novel approach for context-aware recommendation in browsing assistants based on the integration of user profiles, navigational patterns and contextual elements. In this approach, user profiles built using an unsupervised Web page clustering algorithm are used to characterize user ongoing activities and behavior patterns. Experimental evidence show that using longer-term interests to explain active browsing goals user assistance is effectively enhanced.

Keywords  User profiling, context-awareness, browsing assistants.

§1  Introduction

In order to exhibit pro-activeness, agents need to identify the fraction of the user interests that is active at different times, which provides context underlying the user activities. Agents featuring context-awareness, the ability of capturing and using context to predict and anticipate user information needs, become able of focusing information discovery and, as a result of this, increasing the quality of the recommendations delivered to users.

In context-aware systems, user profiles are usually seen as a way to
disambiguate search topics \(^{19}\). In this case, profiles support interactive context-aware retrieval in which relevant documents are gathered upon a direct user request once the search keywords are contextualized with respect to the user interests. However, the lacking of knowledge about active user goals prevents profiles from supporting pro-active, context-aware retrieval in which relevant documents are automatically presented to users according to their activities \(^ {4}\).

In this paper we propose a method to derive semantically enhanced contexts that point out the information which is more likely to be relevant according to the user profile and the current browsing activity. Particularly, we developed this method for agents acting as browsing assistants which treat Web browsing as a cooperative search activity, learning user preferences by watching the user behavior on the Web and providing real-time display of recommendations \(^ {14}\).

Our approach aims to exploit hierarchical representations of user interests obtained by conceptual clustering \(^ {11}\) to characterize contexts which are in turn used to detect regularities and patterns of behavior regarding such interests \(^ {12}\).

The rest of the paper is organized as follows. Section 2 describes the proposed approach for recommendation and its integration with user profiling. Empirical evaluation of this approach is summarized in Section 3. Section 4 discusses related works on the use of context in personal agents. Finally, concluding remarks are stated in Section 5.

§2 Characterizing Browsing Patterns

The task of personal agents is not just choosing the right information for users, but also presenting this information at the right time. In consequence, in addition to knowing which the user preferences and interests are, personal agents require to know habits, routines and patterns of behavior regarding such interests. The understanding of behavior patterns contributes to effectively locate and provide the most appropriate information according to the user needs.

In a context-aware setting, agents recognize contexts and present information to users accordingly. Indeed, pro-activeness is usually considered context triggered, i.e. an agent receives some context information and decides whether to perform some action. Multiple contextual elements can be considered and combined to enhance assistance, including location, time and activities a user is carrying out. For browsing assistants acting in the Web, browsing is considered the main activity of users since it is the target of agent actions.

Browsing can be seen as a sequence of activities that are related to one
Exploiting User Interests to Characterize Navigational Patterns in Web Browsing Assistance

Information agents can take advantage of the knowledge gained from observing user browsing in conjunction with long-term user interests to retrieve context-relevant information. If an agent detects the user is browsing through certain interest categories, it can anticipate the categories in the same session the user is likely to be interested in. The goal of activity-awareness is, therefore, to proactively retrieve Web pages matching the active user interests to compute a set of recommendations for the current browsing session.

In this paper we proposed the integration of user profiles with the extraction of navigational patterns to characterize the behavior of users in terms of their interests. Hierarchical representations of long-term user interests provided by profiles support the categorization of Web pages the user browsed through into semantically meaningful concepts. Browsing sessions extended with this characterization of Web pages (i.e., the concepts or categories pages belong to) are used as input to an association mining process, resulting in a set of rules that relates categories in a profile according to the way the user access to them.

Fig. 1 outline the proposed approach. The user in this example seems to be interested in two broad categories economy and basketball, and within basketball in both NBA and Argentine league related information. Users exhibit some habits that are reflected in the associations among browsing activities. For instance, while visiting on-line newspapers the user in the example may tend to look at NBA and Argentine league news. In the proposed approach, browsing patterns referring to concepts in the user profile that are usually accessed together serve as the basis for recommendation and are mined starting from observation of the user browsing behavior on the Web.
Ultimately, user browsing habits are represented in the user profile by associations of the form $A \Rightarrow B$, where $A$ and $B$ are groups of categories in the profile and the rule indicates that, if the user current activities include visiting pages about the categories in $A$, the next activities are likely to include visiting pages about the categories in $B$. To illustrate the kind of patterns extracted, the following rule might be learned for the user in the example:

$$C_{NBA}(NBA, \text{games, ...}) \Rightarrow C_{Arg.\ league}(LNB, \text{argentine, ...})$$

If the user starts reading about NBA the application of this rule will lead to recommendations about the Argentine league. Categories in the user profile are used to expand and semantically focus recommendations so that any page matching the concept in the consequent can be recommended.

### 2.1 Learning User Interests

To provide effective personalized assistance, personal agents depend on the knowledge they have about individual users contained in user profiles, i.e. models of user interests, preferences and habits. In this work, we base the extraction and representation of interests on a user profiling technique that captures information preferences starting from observation of user behavior on the Web. This technique allows agents to acquire and adapt user profiles which are accurate in predicting the relevance of new pieces of information and, at the same time, provide a conceptual description of user interests.

User profile acquisition is based on a clustering algorithm, named WebDCC (Web Document Conceptual Clustering)\(^1\), with structures and procedures designed to support learning of user interests by information agents. This algorithm belongs to the conceptual paradigm\(^2\), which includes clustering and characterization, i.e. the formation of intentional concept descriptions for extensionally defined clusters. Non-intrusive observation of user behavior on the Web allows agents to capture experiences regarding user interests such as Web pages a user read or bookmarked for future reading. From the analysis of these experiences emerges the knowledge about users to be modeled in their profiles.

WebDCC incrementally creates and revises a concept hierarchy describing the categories user experiences belong to. Hierarchies produced by this algorithm are classification trees in which internal nodes represent concepts, denoted $C = \{c_1, c_2, \ldots, c_n\}$, and leaf nodes represent clusters of experiences. Each concept have an associated text-like description, $c_i = \langle(t_1, w_1), \ldots, (t_m, w_m)\rangle$, ob-
Exploiting User Interests to Characterize Navigational Patterns in Web Browsing Assistance


2.2 Linking Browsing Activities to User Interests

User browsing sessions are sets of page references that take place during one logical period. For example, a session can be defined as the sequence of page accesses from a log in to a log out of the Web browser application. The identification of session boundaries ensures that the information collected from a single session is within the same context, which provides a good foundation for inferring and applying context in recommendation.

In Web usage mining, data sources are primarily Web server logs. Then, a session contains the requests of a single user to a given Web site. Several problems arise when logs have to be processed for data analysis, including page-view identification (i.e., what a single browser display consists of), user identification, session identification, inference of missing references due to caching and other factors. Conversely, browsing assistants focus on extracting behavioral patterns of individual users based on observation of their behavior across the Web.
In order to collect Web usage data, browsing assistants have to record user actions by monitoring Web browsers (using remote agents such as Java applets or modifying the source code of existing browsers). As a result of data collection, assistants obtain logs registering the activities carried out by the observed user in one or more browser windows. Thus, the content of Web pages the user browsed through, access time, time spent on each page, actions performed in the browser (e.g., bookmarking, scrolling, etc.) and other data can be analyzed. Moreover, actions such as opening or closing the browser can be used to detect the begin and end of browsing sessions safely. Each session $S_j$ is considered to be a list of pages a user accessed to ordered by time-stamp:

$$S_j = \{(p_1, \text{time}_1), (p_2, \text{time}_2), \ldots, (p_n, \text{time}_n)\}$$  \hspace{1cm} (1)

where $\text{time}_i$ is the time the user accessed the page $p_i$ such that $\text{time}_i \leq \text{time}_j, \forall i \leq j$. Then, the user browsing activities are partitioned into a set of sessions $S = \{S_1, S_2, \ldots, S_k\}$ containing individual page references.

The segmentation of user activities into sessions is performed using a time-oriented heuristic in which a time-out establishes a period of inactivity that is interpreted as a signal that the session has ended. Frequently, a 30 minutes window is used as default time-out, although empirical studies have reported other values according to different analysis of Web server logs such as 25.5 minutes. If the user did not request any page for a period longer than $\max$ time (this is an input parameter for the method, 30 min. is used by default) subsequent requests are considered to be in another session. In addition, the active session is finished when the browser is closed and a new session is started when the browser is re-opened. The algorithm for identifying sessions which are in turn used to detect the activity context of users can be outlined as follows:

1. For each visited Web page in the browser do the following steps
2. If the page is an image file, it is discarded. Image files are detected by looking at their file name extensions (.gif, .jpg, etc.)
3. If the elapsed time from the last request is within $\max$ time the page is appended into the current session; otherwise, the session is closed and the page is the first entry of a new session
4. If the browser is closed, the current session is also closed and a new session is created when the browser is opened again by the user.

The notion of session can be further abstracted by selecting a subset of pages that are significant or relevant for analysis within a session. Each
Exploiting User Interests to Characterize Navigational Patterns in Web Browsing Assistance

A semantically meaningful subset of pages belonging to a user session is referred to as a transaction. A transaction differs from a session in that the size of a transaction can range from a single page to all of the pages in the session, depending on the criteria used to delimit transactions.

For transaction identification it is assumed that user sessions have already been determined. The input to this process consists of all of the page references for each user session $S_j$ in $S$. Unlike in traditional domains for data mining, such as market databases, in Web usage mining there is no convenient method of clustering page references into transactions smaller than an entire user session $8)$. Hence, the problem of identifying transactions consists in either dividing a large transaction into multiple smaller ones or merging small transactions into fewer larger ones $9)$. 

Maximal forward reference (MFR) $7)$ and reference length $8)$ are examples of divide approaches that attempt to identify semantically meaningful transactions. In MFR, a maximal forward reference is defined as the set of pages from the first page in a request sequence to the final page before a backward reference is made, where a backward reference is a page that has already occurred in the current session. Reference length approach is based on the assumption that the amount of time a user spends on a page correlates to whether the page should be classified as a content page or an auxiliary one, i.e. a page used for linkage in which a user would spend a relatively short time. In turn, transactions can be either auxiliary-content, composed of a sequence of auxiliary pages that ends with a content page, or content-only, formed by all the content pages in a session.

We consider content pages as those pages read by the user, detected using implicit interest indicators, which belong to one or more categories in the user profile, while pages that can not be related to any interest in the profile are considered auxiliary ones. Unlike content pages in the mentioned approaches, which are identified simply based on the time spent on them or backtracking, content pages in our approach are detected by comparing the actual content of the page with the user profile as the user is browsing. Web pages that can not be classified into any category are considered irrelevant for usage mining since do not possess information about the user habits regarding interests.

Content-only transactions, formed by all the content pages in a session, resulting from comparing the visited Web pages with the user profiles are further divided or merged to ensure that each transaction has some minimum overall length based on a time window approach $8)$. This approach assumes that
meaningful transactions have an overall average length associated with them and divides (or merge) browsing sessions into time intervals no longer than a specified threshold. For a large enough specified time window, each transaction will contain an entire user session. If \( W \) is the length of the time window, then two pages \( p_i \) and \( p_j \) are in the same session if \( p_i.time - p_j.time \leq W \). Then, the pages \( P = \{p_1, p_2, \ldots, p_n\} \), each with its associated \( time_i \), appearing in the sessions of \( S \) are partitioned into a set of \( m \) transactions \( T = \{t_1, t_2, \ldots, t_m\} \) where each \( t_i \in T \) is a subset of \( P \) as result of transaction identification. The problem of mining association rules is defined over this collection of subsets from the item space where an item refers to an individual page reference.

Once transactions are identified page references are linked to concepts in the user profile in order to extract patterns at conceptual level. This enriched version of transactions leads to rules including categories instead of single Web pages. To integrate interests and browsing activity, each page \( p_i \) in a transaction \( t_j \) is associated to the set of categories \( C_i \) it belongs to, defined by all of the categories \( c_j \) in the path from the root of the hierarchy to the leaf cluster in which the page \( p_i \) can be classified into. This classification takes place while the user is browsing, when the distinction between content and auxiliary is made.

If only the cluster a page belongs to is used to describe sessions, the discovered association rules will relate clusters but not categories. Instead, the inclusion of ancestors in the path from the leaves to the root concept in the hierarchy makes it possible to find rules at different hierarchical levels. The result of transforming the elements of the transactions in \( T \) from page references to categories in the user profile is a set of transactions \( T' = \{t'_1, t'_2, \ldots, t'_m\} \) where each \( t'_i \in T' \) is a subset of \( C \), the categories in the user profile. The algorithm for transaction identification can be outlined as follows:

1. For each session \( S_i \in S \) create a new transaction \( t_i \) in \( T \)
2. For each page \( p_j \in S_i \), find the set \( C_j \) by classifying the page into the current user interest hierarchy
3. If \( C_j \neq \emptyset \), add \( p_j \) to the transaction \( t_i \) since the page is a content page
4. Repeat steps 2 and 3 until all page references have been either added to the transaction or discarded
5. Repeat steps 1 to 4 until all sessions in \( S \) have been processed
6. Use the time window approach to partition each \( t_i \in T \) into transactions smaller than \( W \)
7. For each resulting transaction \( t_i \in T \), create the transaction \( t'_i \) in \( T' \)
replacing each page $p_j \in t_i$ by the corresponding $C_j$.

### 2.3 Mining Associations in Browsing Activities

The formal statement of the association rule mining problem was firstly stated in $^2$. Let $\mathcal{I} = \{I_1, I_2, \ldots, I_m\}$ be a set of literals called items, a subset $X \subseteq \mathcal{I}$ is called an itemset and a $k$-itemset is an itemset that contains $k$ items. Let $\mathcal{D}$ be a database of transactions, where each transaction $T$ is a set of items such that $T \subseteq \mathcal{I}$. A transaction $T$ contains an itemset $X$ if $X \subseteq T$. Each itemset has a certain statistical significance called support such that an itemset $X$ has support $s$ in the transaction set $\mathcal{D}$ if $s\%$ of transactions in $\mathcal{D}$ contain $X$, i.e.

$$s(X) = \frac{|\{T \in \mathcal{D} | X \subseteq T\}|}{|\mathcal{D}|}$$

An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset \mathcal{I}$, $Y \subset \mathcal{I}$, and $X \cap Y = \emptyset$. In this rule, $X$ is called the antecedent and $Y$ is called the consequent of the rule. The rule $X \Rightarrow Y$ holds in the transaction set $\mathcal{D}$ with confidence $c$ if $c\%$ of the transactions in $\mathcal{D}$ that contain $X$ also contain $Y$, i.e.

$$c(X, Y) = \frac{s(X \cup Y)}{s(X)}$$

For the set of transactions $\mathcal{D}$, the problem of mining association rules consists in finding all rules $X \Rightarrow Y$ with support and confidence greater than some user-specified minimums, $\text{minsup}$ and $\text{minconf}$. For each rule, the support threshold describes the minimum percentage of transactions containing all items that appear in the rule, whereas the confidence threshold specifies the minimum probability for the rule consequent to be true if the antecedent is true.

The problem of finding categories that are usually visited together is similar to finding associations among itemsets in transaction databases. In analyzing sessions for Web usage mining the visit-coherence assumption is usually made $^{18}$, this is the pages a user visits during one interaction with the site tend to be conceptually related. It can not be assumed that all pages in a single visit are related, but real associations arise adding support and confidence across sessions. Likewise, interest associations are intended to find frequently occurring combinations of categories in user browsing sessions.

Finding associations rules for items belonging to a hierarchy requires placing rules at the correct level, denoting more general or specific associations...
according to the user habits. Mining multiple-level or generalized association rules assumes a hierarchy or taxonomy $T$ on the items instead of a flat itemset $I$. Two reasons make association rules across different levels of a taxonomy useful: (1) rules at lower levels may not have the minimum support, grouping items in categories allows discovering interesting rules which would otherwise not be found, due to the fine leaf-level granularity caused by the large number of items; (2) the hierarchy can be used to prune uninteresting or redundant rules.

The problem of mining generalized associated rules is stated as follows. If $D$ is a set of transactions where each transaction $T$ is a set of items such that $T \subseteq I$, a transaction $T$ supports an item $x \subseteq I$ if $x$ is in $T$ or $x$ is an ancestor of some item in $T$. Hence, a transaction $T$ supports $X \subseteq I$ if $T$ supports every item in $X$. A generalized association rule is an implication of the form $X \Rightarrow Y$, where $X \subseteq I, Y \subseteq I, X \cap Y = \emptyset$ and no item in $Y$ is an ancestor of any item in $X$ as this would be a trivially valid association.

The rule $X \Rightarrow Y$ holds in the transaction set $D$ with confidence $c$ if $c\%$ of transactions in $D$ that support $X$ also support $Y$. The rule $X \Rightarrow Y$ holds in the transaction set $D$ if $s\%$ of transactions in $D$ support $X \cup Y$. These are called generalized association rules because both $X$ and $Y$ can contain items from any level of the taxonomy $T$. For determining if a transaction $T$ supports an itemset $X$, it is necessary to check for each item $x \in X$ whether $x$ or some descendant of $x$ is present in the transaction. To simplify this task, all the ancestors of each item in $T$ are added to this transaction to form an extended transaction $T'$.

A straightforward method to find generalized association rules is to run any association rule algorithm on the extended transactions since $T$ supports $X$ if and only if $T'$ is a superset of $X$. Thus, *Apriori* algorithm is directly applicable to describe browsing patterns since extended transactions are obtained when pages are classified in the concept hierarchy to determine whether they are content pages during transaction identification. For empirical evaluation of the proposed approach, we mined association rules using an efficient implementation of *Apriori* over the set of extended transactions.

### 2.4 Browsing Assistance and Recommendation

From user browsing sessions, patterns representing the user navigational behavior regarding long-term user interests are extracted in the form of association rules. Browsing assistants, therefore, become able to proactively retrieve

---

*http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html*
relevant information to generate recommendations for a user by matching the current user browsing activity against the discovered patterns. By expressing rules at a conceptual level, the set of Web pages to be recommended is broadened to include those pages an agent might discover on the Web that can also be characterized by the involved categories.

In contrast to association rule mining, recommendation is an on-line process in which agents need to determine a set of candidate recommendations before rule matching or trigger a Web search. To gather potential recommendations in advance, an agent might perform a Web search to retrieve pages belonging to the concepts the user is interested in during idle computer time. For example, agents can retrieve pages from some fixed sites periodically (e.g., newspapers early in the morning) or find nearest neighbor documents by using some experiences in the profile as query (e.g., "more documents like this one").

Once a set of candidate recommendations is available, a fixed-size sliding window is used over the active session to capture the current user activity so that recommendations can be delivered in the precise moment they are needed. For a sliding window of size $n$, the active session ensures that only the last $n$ visited pages influence recommendation. The use of a window is important in discovering context since most users go back and forth while browsing to find the desired information so that earlier portions of the browsing history may refer to no longer valid information needs.

In the recommendation phase, a semantically enriched representation of the active session denoting the active user interests is compared with the previously discovered rules. If the active session matches the antecedent of one or more association rules, recommendations are generated by matching the rule consequents against the candidate set of recommendations gathered by the agent beforehand or resulting from a search triggered by the context of activities.

Fig. 2 shows a screenshot of a browsing assistant generating recommendations during Web browsing. The user browsing history for the current session is displayed and recommendations stemming from proactive agent actions based on the browsing context are presented to the user. The expected level of interest on each recommendation is calculated based on the confidence of the association rule whose consequent contains the recommended category.

§3 Empirical Evaluation

To evaluate the activity-based recommendation approach, a client-side
Fig. 2  Browsing assistant generating recommendation during Web browsing

log of visited Web pages in a number of topics the user is interested in as well as the location of these pages on the concept hierarchy is needed. Unfortunately, available log datasets record the accesses of multiple users to a single Web site since they are mostly used to improve the usability of Web sites by finding associations between pages according to how users browsed the site.

We used the content and logs of the *Music Machines* Web site\(^2\) for experimentation. This site contains information about electronic musical equipment, primarily grouped by manufacturers. For each manufacturer there are multiple entries for the different instrument models and for each model there are pictures, reviews, user manuals, etc. In total, the site has approximately 4582 distinct pages including HTML pages, plain text, images and audio samples.

In the logs\(^3\), containing accesses to the site from 02/12/97 through 04/30/99, users are anonymized with respect to the originating machines, i.e. all hits from one machine on a particular day have the same label. Thus, the browsing behavior of individual users can be interpreted respecting of their interest categories within the content of the Web site. Each access log consists of a user label, request method (*Get*, *Post*, etc.), URL of the page accessed, data transmission protocol, access time and browser used to access the site. In server logs, a request of a user to view a particular page usually results in several log entries since images, sound and other files are downloaded from the server in addition to the requested page. For this reason, the site logs contain several entries that are irrelevant for analysis and have to be discarded.

*Music Machines* logs were filtered to remove entries resulting in errors, using a request method other than *Get* or recording accesses to image and audio files. Even though the basic structure and most of the site files remained un-

\(^2\) http://machines.hyperreal.org/
\(^3\) Available at http://www.cs.washington.edu/ai/adaptive-data/
Exploiting User Interests to Characterize Navigational Patterns in Web Browsing Assistance. (D. Godoy, A. Amandi).

The final publication is available at Springer via http://dx.doi.org/10.1007/s00354-008-0044-x

Table 1  Summary of data about the users involved in the experiments

<table>
<thead>
<tr>
<th>ID</th>
<th>label</th>
<th>date</th>
<th>begin time</th>
<th>end time</th>
<th>login duration</th>
<th># entries</th>
</tr>
</thead>
</table>

Exploiting User Interests to Characterize Navigational Patterns in Web Browsing Assistance

Touched during the period the logs were collected, some URLs refer to pages that do not exist in the available site copy. Entries corresponding to these pages were also removed from the logs. From all users who entered the site after 08/20/98, the date on which the copy of the site content was made, the five users having the longest sessions in terms of the time they spend browsing the site were selected for experimentation. Table 1 shows the label and number of accesses of these users as well as the initial and end time of their log entries.

For each user a profile was built based on the content of Web pages from the site. The experimental procedure simulates users that are browsing the Music Machines site and obtaining recommendations as follows:

1. Identify the user entries in the log files
2. Extract the URLs of the visited pages and run WebDCC algorithm over these pages using the available copy of the Music Machines site for extracting the content of Web pages and, thus, obtain a user profile
3. Identify user sessions in the logs using max_time=30 minutes
4. Divide the resulting set of transactions into a training (approx. 70%) and a testing set (approx. 30%) to perform the experiments
5. Use the training set to mine association rules regarding categories in the user profile
6. Use the testing set to simulate active session windows and make recommendations
7. Evaluate the effectiveness of recommendations

In order to assess quantitative and qualitative values of recommendation performance, we used the adaptations of the standard measures precision and coverage proposed by 16). In this setting, precision measures the degree to which the method produces accurate recommendations for the active session and coverage measures its ability to recommend all the items that are likely to be visited by the
user in the active session. Low precision indicates recommendations which are not interesting to the user in a given context of activities, whereas low coverage indicates the absence of some relevant recommendations in certain contexts.

Given a transaction $t$ and a set of recommendations $R$ produced using a window $w$ such that $w \subseteq t$, the precision of $R$ with respect to $t$ is defined as:

$$\text{precision}(R, t) = \frac{|R \cap (t - w)|}{|R|} \quad (4)$$

and the coverage of $R$ with respect to $t$ is defined as:

$$\text{coverage}(R, t) = \frac{|R \cap (t - w)|}{|t - w|} \quad (5)$$

F-Measure is the weighted harmonic mean of precision and coverage defined as follows:

$$F_\alpha(R, t) = \frac{(1 + \alpha) \cdot \text{precision}(R, t) \cdot \text{coverage}(R, t)}{\alpha \cdot \text{precision}(R, t) + \text{coverage}(R, t)} \quad (6)$$

as $\alpha$ value increases, the weight of coverage also increases. For instance, $F_2$ measure weights coverage twice as much as precision.

For a given transaction $t$ in the testing set and an active session window of size $n$, we randomly chose $|t| - n + 1$ groups of items, each having size $n$, from the transaction as the surrogate active session windows. For each active session, recommendations are produced based on the extracted rules and compared to the remaining items in the transactions, i.e. $t - w$, to compute performance measures. For each measure, the final score of the transaction $t$ is the average over all of the $|t| - n + 1$ surrogate active sessions associated with this transaction.

Table 2 summarizes the number of remaining entries in the logs after removing images and other useless files as well as the number of sessions and unique pages that each user accessed in the site. WebDCC algorithm was run.
Exploiting User Interests to Characterize Navigational Patterns in Web Browsing Assistance

The final publication is available at Springer via http://dx.doi.org/10.1007/s00354-008-0044-x

over the documents each user accessed to identify user interest categories. Even though no concepts were extracted by the clustering algorithm, the pages each user read were partitioned into several clusters. Table 2 shows the number of clusters after filtering singleton clusters. The impossibility to create concepts summarizing pages was mainly due to the site content and structure. It contains few pages referred to numerous manufacturers, so that different clusters are created for each of them and no generalization is feasible. In turn, user sessions were partitioned into transactions and pages in each transaction were mapped to clusters obtaining a set of content-enhance transactions. Apriori algorithm was run over transactions to obtain association rules. The following rule exemplifies the kind of relations obtained in these experiments:

\[ \text{Cluster}\#5(\text{mixer, dual, space,...}) \Rightarrow \text{Cluster}\#3(\text{yamaha,...}) \]

This rule implies that the user tend to look for information of the mentioned manufacturer when starts reading about mixers.

To evaluate recommendation, the impact of several parameters have to be analyzed conjunctively. We first analyzed the impact of \(W\) on the quality of recommendations by testing the values 3, 5, 10 and 15 minutes. This parameter affects primarily the number of transactions obtained from a given session and, consequently, the number of rules and the quality of the corresponding recommendations. For the previous values, association rules having a support greater than 2\% were extracted, whereas a fixed active window of size 3 and a minimum confidence of 90\% were used to obtain recommendations.

Fig. 3 summarizes the average and standard deviation of results obtained for the five users regarding to number of transactions, proportion of extracted
rules given the number of recommendations and precision of recommendations. For each value of $W$, the results for rules having 1-itemsets and 2-itemsets are shown. Fig. 3(a) depicts how the number of transactions decreases as the time window $W$ is expanded. Fewer transactions lead to more association rules since there are more rules supported by the data, but their quality is inferior as is denoted by the decrease in precision in Fig. 3(b). In further experiments we considered a time window $W = 3$ min. since the dropping in precision for the immediately next value is significant ($\simeq 9\%$). For 2-itemsets not only precision decreases, but also the number of rules rises drastically. Since recommending Web pages implies rule matching against the current browsing history, the online analysis of such high number of rules becomes computationally expensive.

The second parameter to analyze is the size of the active window session used to produce recommendations, experiments were performed using sizes from $n = 1$ to 6. Fig. 4(a) summarizes the average and standard deviation of $F_2$ scores achieved for the five users varying the confidence threshold. We considered that precision can be sacrificed for the sake of an increase in coverage, since a lower precision means that the agent is recommending pages which are not contextually relevant but are still content relevant. There is a trade-off between enlarging the window enough to recommend most of the pages that are relevant to the current context but not those that were relevant to the previous one. It can be observed that the longer the window the higher the scores, whereas the poorest value was obtained with a window of size one (i.e., only the last visited page is considered). As the window is expanded more rules matches the pages inside the window, increasing the number of recommendations and their coverage.

Fig. 4(b) presents a comparison of the proposed approach with a commonly used method for extracting context which considers profiles built based on the keywords present on recently consulted documents. The user context in this method is represented by a vector of the most frequent words (10 words in the experiment) in documents within the sliding window. Pages are recommended if their similarity with this vector exceeds a given threshold ($0.7$ was used). We used the same data to evaluate both approaches, resulting in poorer $F_2$ scores for keyword-based profiles due to the decrease of precision in recommendation.

Even though further experiments need to be carried out to evaluate the accurateness of the method when rules are extracted at different hierarchical levels, the extraction of generalized association rules is supported by the approach. Finally, the effectiveness of association rule mining relies on having
enough data to collect evidence about potential associations which requires the user to produce enough personal Web usage information.

§4 Related Works

User profiles are frequently used as a method of gathering contextual information since they persist across retrieval sessions and can be added to queries. However, profiles are passive representation of user interests which do not provide support to pro-active, context-aware information retrieval. Browsing assistants have no means to anticipate information needs based solely on the knowledge about user interests. In our approach browsing activities characterized according to long-term interests act like trigger for retrieval of information matching profiles, then allowing just-in-time access to relevant information.

In contrast to systems like WordSieve 3) or Watson 5), which retrieve documents based on words extracted from recently consulted documents, our approach extracts information about how users tend to access interests in order to determine what kind of documents is likely to be interesting in a certain context. Other works deals with the integration of Web usage and content mining for the extraction of navigational patterns. For example, a framework for enhancing Web usage records with formal semantics from an ontology underlying a particular Web site is presented in 17). To mine interesting patterns, user transactions are semantically enriched with concept labels assuming that semantic annotation of Web content is performed a priori. Likewise, SEWeP 10) system uses both usage logs and Web site semantics for personalization. This system uses a domain-specific taxonomy for Web content characterization, whereas association rule mining is used to extract navigational patterns.
Works that use profiles to describe context usually assumed the existence of a pre-defined, general-purpose taxonomy to explain interests. For example, in ARCH profiles generated applying an unsupervised clustering technique over interesting documents are used to automatically learn the semantic context of user needs based on a domain-specific concept hierarchy such as Yahoo!. Users interact with this hierarchy to derive an enhance query. Other work that assumes the existence of a conceptual hierarchy for modeling users whose surfing behavior is dynamically governed by their current interest topics is presented in [1]. This approach involves mapping visited pages to hierarchical topics and then estimating the parameters of a semi-Markov process defined on the tree based on the observed transitions among the visited pages.

The approach presented in this paper differs from previous works in several aspects. First, server-side Web usage mining extracts information about the behavior of multiple users in a specific Web site extracting group behavior patterns. By contrast, our approach tries to extract associations among interests of a single user browsing the Web. Second, most usage mining approaches obtain association rules that relate URLs within a site in order to reorganize it. By capturing navigational patterns at conceptual level our approach provides flexibility in both mining and retrieval phases since navigation patterns refer to broad user interests. Third, conceptual hierarchies in the mentioned approaches are either domain-dependant, manually constructed ontologies or automatically learned hierarchies representing a given Web site. In our approach taxonomies describing user interests are learned from observation of user behavior on the Web, so that assistants can deal with diverse and unpredictable interests.

§5 Conclusions

The context of the user activities is an important issue to take into account during recommendation which, however, has received little attention in browsing assistants. Most agents are concerned with estimating the user interest regarding new pieces of information, instead of trying to place the interesting information in the right moments. In this paper, we have described a novel approach to characterize the activity context of users according to their interests and extract navigational patterns in order to enable context-aware information retrieval and recommendation. By taking advantage of contextual data, autonomous agents can proactively search, filter, and present information which is not only relevant, but useful to the user current activities and browsing goals.
Empirical evaluation of the proposed approach showed that the extraction of association rules describing browsing patterns at a conceptual level helps to predict part of the interests which are relevant to the user in a browsing session. By relating browsing activities to the information preferences described in the user profiles, more coherent, semantically meaningful contexts can be recognized. Thus, browsing assistants acquire the ability of focusing retrieval on those interests that are in the semantic scope of the ongoing user activity as well as detect out-of-context interests.

Acknowledgment

This research has been partially supported by ANPCyT PICT No. 34917.

References


Daniela Godoy is a professor in the Computer Science Department at UNICEN University and researcher at ISISTAN Research Institute. Her research interests include intelligent agents and user profiling. She received her PhD in computer science from the UNICEN University in 2005.

Anaíla Amandi is a professor in the Computer Science Department at UNICEN University, where she leads the ISISTAN Research Institute’s Intelligent Agents Group. She received her PhD in 1997. Her research interests include personal assistants and knowledge management.