Performance Based Novel Techniques for Semantic Web Mining

Mahendra Thakur[1] , Geetika S. Pandey[2]

[1] MTECH Scholar, Computer Science Department
Samrat Ashok Technological Institute, Vidisha, M.P, INDIA

[2]Asst. Professor, Computer Science Department
Samrat Ashok Technological Institute, Vidisha, M.P, INDIA

Abstract

The explosive growth in the size and use of the World Wide Web continuously creates new great challenges and needs. The need for predicting the users' preferences in order to expedite and improve the browsing though a site can be achieved through personalizing of the websites. Most of the research efforts in web personalization correspond to the evolution of extensive research in web usage mining, i.e. the exploitation of the navigational patterns of the web site's visitors. When a personalization system relies solely on usage-based results, however, valuable information conceptually related to what is finally recommended may be missed. Moreover, the structural properties of the web site are often disregarded. In this paper, we propose novel techniques that use the content semantics and the structural properties of a web site in order to improve the effectiveness of web personalization. In the first part of our work we present standing for Semantic Web Personalization, a personalization system that integrates usage data with content semantics, expressed in ontology terms, in order to compute semantically enhanced navigational patterns and effectively generate useful recommendations. To the best of our knowledge, our proposed technique is the only semantic web personalization system that may be used by non-semantic web sites. In the second part of our work, we present a novel approach for enhancing the quality of recommendations based on the underlying structure of a web site. We introduce UPR (Usage-based PageRank), a PageRank-style algorithm that relies on the recorded usage data and link analysis techniques. Overall, we demonstrate that our proposed hybrid personalization framework results in more objective and representative predictions than existing techniques.

Keywords- *Web personalization, Semantic web and Recommender systems.*

1.Introduction

During the past few years the World Wide Web has become the biggest and most popular way of communication and information dissemination. It serves as a platform for exchanging various kinds of information, ranging from research papers, and educational content, to multimedia content, software and personal logs. Every day, the web grows by roughly a million electronic pages, adding to the hundreds of millions pages already on-line. Because of its rapid and chaotic growth, the resulting network of information lacks of organization and structure. Users often feel disoriented and get lost in that information overload that continues to expand. On the other hand, the e-business sector is rapidly evolving and the need for web market places that anticipate the needs of their customers is more than ever evident. Therefore, the ultimate need nowadays is that of predicting the user needs in order to improve the usability and user retention of a web site. This paper presents novel methods and techniques that address this requirement. In brief, web personalization can be defined as any action that customizes the information or services provided by a web site to an individual user, or a set of users, based on knowledge acquired by their navigational behavior, recorded in the web site's logs, in other words, its usage. This information is often combined with the content and the structure of the web site, as well as the interests/preferences of the user, if they are available. The web personalization process is illustrated in Figure 1. Using the four aforementioned sources of information as input to pattern discovery techniques, the system tailors the provided content to the needs of each visitor of the web site. The personalization process can result in the dynamic generation of recommendations, the creation of index pages, the highlighting of existing hyperlinks, the publishing of targeted advertisements or emails, etc. In this paper we focus on personalization systems that aim at providing personalized recommendations to the web site's visitors. Furthermore, since the personalization algorithms we propose in this work are generic and applicable to any web site, we assume that no explicit knowledge involving the

users' profiles, such as ratings or demographic information is available [1] and [2] and [5].

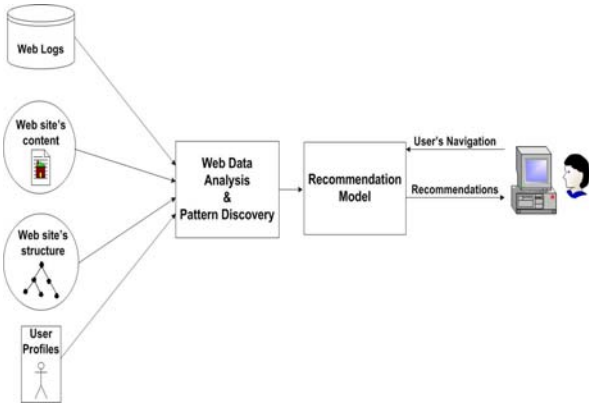


Figure: 1 the web personalization process

The problem of providing recommendations to the visitors of a web site has received a significant amount of attention in the related literature. Most of the research efforts in web personalization correspond to the evolution of extensive research in web usage mining, taking into consideration only the navigational behavior of the (anonymous or registered) visitors of the web site. Pure usage-based personalization, however, presents certain shortcomings. This may happen when, for instance, there is not enough usage data available in order to extract patterns related to certain navigational actions, or when the web site's content changes and new pages are added but are not yet included in the web logs. Moreover, taking into consideration the temporal characteristics of the web in terms of its usage, such systems are very vulnerable to the training data used to construct the predictive model. As a result, a number of research approaches integrate other sources of information, such as the web content or the web structure in order to enhance the web personalization process. As already implied, the users' navigation is largely driven by semantics. In other words, in each visit, the user usually aims at finding information concerning a particular subject. Therefore, the underlying content semantics should be a dominant factor in the process of web personalization. The web site's content characterization process involves the feature extraction from the web pages. Usually these features are keywords subsequently used to retrieve similarly characterized content. Several methods for extracting keywords that characterize web content have been proposed. The similarity between documents is usually based on exact matching between these terms. This way, however, only a binary matching between documents is achieved, whereas no actual semantic similarity is taken into consideration. The need for a more abstract representation that will enable a uniform and more flexible document matching process imposes the use of semantic web structures, such as ontology's. By mapping the keywords to the concepts of an ontology, or topic hierarchy, the problem of binary matching can be surpassed through the use of the hierarchical relationships and/or the semantic similarities among the ontology terms, and therefore, the documents. Finally, we should take into consideration that the web is not just a collection of documents browsed by its users. The web is a directed labeled graph, including a plethora of hyperlinks that interconnect its web pages. Both the structural characteristics of the web graph, as well as the web pages' and hyperlinks' underlying semantics are important and determinative factors in the users' navigational process. The main contribution of this paper is a set of novel techniques and algorithms aimed at improving the overall effectiveness of the web personalization process through the integration of the content and the structure of the web site with the users' navigational patterns. In the first part of our work we present the semantic web personalization system standing for Semantic Web Personalization that integrates usage data with content semantics in order to compute semantically enhanced navigational patterns and effectively generate useful recommendations. Similar to previously proposed approaches, the proposed personalization framework uses ontology terms to annotate the web content and the users' navigational patterns. The key departure from earlier approaches, however, is that standing for Semantic Web Personalization is the only web personalization framework that employs automated keyword-to-ontology mapping techniques, while exploiting the underlying semantic similarities between ontology terms. Apart from the novel recommendation algorithms we propose, we also emphasize on a hybrid structure-enhanced method for annotating web content. To the best of our knowledge, standing for Semantic Web Personalization is the only semantic web personalization system that can be used by any web site, given only its web usage logs and a domain-specific ontology [1], [2], [4] and [6].

2. Background

The main data source in the web usage mining and personalization process is the information residing on

the web site's logs. Web logs record every visit to a page of the web server hosting it. The entries of a web log file consist of several fields which represent the date and the time of the request, the IP number of the visitor's computer (client), the URI requested, the HTTP status code returned to the client, and so on. The web logs' file format is based on the so called "extended" log format. Prior to processing the usage data using web mining or personalization algorithms, the information residing in the web logs should be preprocessed. The web log data preprocessing is an essential phase in the web usage mining and personalization process. An extensive description of this process can be found. In the sequel, we provide a brief overview of the most important pre-processing techniques, providing in parallel the related terminology. The first issue in the pre-processing phase is data preparation. Depending on the application, the web log data may need to be cleaned from entries involving page accesses that returned, for example, an error or graphics file accesses. Furthermore, crawler activity usually should be filtered out, because such entries do not provide useful information about the site's usability. A very common problem to be dealt with has to do with web pages' caching. When a web client accesses an already cached page, this access is not recorded in the web site's log. Therefore, important information concerning web path visits is missed. Caching is heavily dependent on the client-side technologies used and therefore cannot be dealt with easily. In such cases, cached pages can usually be inferred using the referring information from the logs and certain heuristics, in order to re-construct the user paths, filling out the missing pages. After all page accesses are identified, the pageview identification should be performed. A pageview is defined as "the visual rendering of a web page in a specific environment at a specific point in time". In other words, a pageview consists of several items, such as frames, text, graphics and scripts that construct a single web page. Therefore, the pageview identification process involves the determination of the distinct log file accesses that contribute to a single pageview. Again such a decision is application-oriented. In order to personalize a web site, the system should be able to distinguish between different users or groups of users. This process is called user profiling. In case no other information than what is recorded in the web logs is available,

this process results in the creation of aggregate, anonymous user profiles since it is not feasible to distinguish among individual visitors. However, if the user's registration is required by the web site, the information residing on the web log data can be combined with the users' demographic data, as well as with their individual ratings or purchases. The final stage of log data pre-processing is the partition of the web log into distinct user and server sessions. A user session is defined as "a delimited set of user clicks across one or more web servers", whereas a server session, also called a visit, is defined as "a collection of user clicks to a single web server during a user session". If no other means of session identification, such as cookies or session ids is used, session identification is performed using time heuristics, such as setting a minimum timeout and assumes that consecutive accesses within it belong to the same session, or a maximum timeout, assuming that two consecutive accesses that exceed it belong to different sessions [3] and [4].

2.1 Web Usage Mining and Personalization

Web usage mining is the process of identifying representative trends and browsing patterns describing the activity in the web site, by analyzing the users' behaviour. Web site administrators can then use this information to redesign or customize the web site according to the interests and behavior of its visitors, or improve the performance of their systems. Moreover, the managers of e-commerce sites can acquire valuable business intelligence, creating consumer profiles and achieving market segmentation. There exist various methods for analyzing the web log data. Some research studies use well known data mining techniques such as association rules discovery, sequential pattern analysis, clustering, probabilistic models, or a combination of them. Since web usage mining analysis was initially strongly correlated to data warehousing, there also exist some research studies based on OLAP cube models. Finally some proposed web usage mining approaches that require registered user profiles, or combine the usage data with semantic meta-tags incorporated in the web site's content. Furthermore, this knowledge can be used to automatically or semi-automatically adjust the content of the site to the needs of specific groups of users, i.e. to personalize the site. As already mentioned, web personalization may include the provision of recommendations to the users, the creation of new index pages, or the generation of targeted advertisements or product promotions. The usage-based personalization systems use association rules and sequential pattern discovery, clustering, Markov models, machine learning algorithms, or are based on collaborative filtering in order to generate recommendations. Some research studies also

combine two or more of the aforementioned techniques [3] and [7].

2.2 Integrating Structure in Web Personalization

Although the connectivity features of the web graph have been extensively used for personalizing web search results, only a few approaches exist that take them into consideration in the web site personalization process. To use citation and coupling network analysis techniques in order to conceptually cluster the pages of a web site. The proposed recommendation system is based on Markov models. In previous, use the degree of connectivity between the pages of a web site as the determinant factor for switching among recommendation models based on either frequent itemset mining or sequential pattern discovery. Nevertheless, none of the aforementioned approaches fully integrates link analysis techniques in the web personalization process by exploiting the notion of the authority or importance of a web page in the web graph [2] and [10].
In a very recent work, address the data sparsity problem of collaborative filtering systems by creating a bipartite graph and calculating linkage measures between unconnected pairs for selecting candidates and make recommendations. In this study the graph nodes represent both users and rated/purchased items. Finally, subsequent work, proposed independently two link analysis ranking methods, SiteRank and PopularityRank which are in essence very much like the proposed variations of our UPR algorithm (PR and SUPR respectively). This work focuses on the comparison of the distributions and the rankings of the two methods rather than proposing a web personalization algorithm [2], [3] and [11].

3. Proposed Personalization Techniques

In this paper we present standing for Semantic Enhancement for Web Personalization, a web personalization framework that integrates content semantics with the users' navigational patterns, using ontologies to represent both the content and the usage of the web site. In our proposed framework we employ web content mining techniques to derive semantics from the web site's pages. These semantics, expressed in ontology terms, are used to create semantically enhanced web logs, called C-logs (concept logs). Additionally, the site is organized into thematic document clusters. The C-logs and the document clusters are in turn used as input to the web mining process, resulting in the creation of a broader, semantically enhanced set of recommendations. The whole process bridges the gap between Semantic Web and Web Personalization areas, to create a Semantic Web Personalization system.

3.1 Standing for Semantic Enhancement for Web Personalization System Architecture

Standing for Semantic Enhancement for Web Personalization uses a combination of web mining techniques to personalize a web site. In short, the web site's content is processed and characterized by a set of ontology terms (categories). The visitors' navigational behavior is also updated with this semantic knowledge to create an enhanced version of web logs, C-logs, as well as semantic document clusters. C-Logs are in turn mined to generate both a set of URI and category-based association rules. Finally, the recommendation engine uses these rules, along with the semantic document clusters in order to provide the final, semantically enhanced set of recommendations to the end user.

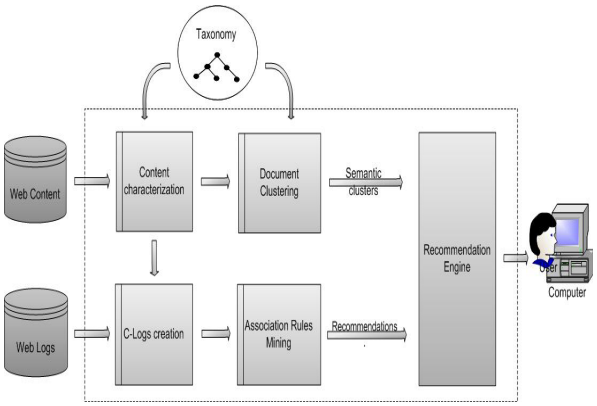


Figure 2 Standing for Semantic Enhancement for Web Personalization architecture

As illustrated in Figure 2, Standing for Semantic Enhancement for Web Personalization consists of the following components:

• **Content Characterization*:*** This module takes as input the content of the web site as well as a domain-specific ontology and outputs the semantically annotated content to the modules that are responsible for creating the C-Logs and the semantic document clusters. The content characterization process consists of the keyword extraction, keyword translation and semantic characterization sub-processes.
• **Semantic Document Clustering***:* The semantically annotated pages created by the previous component are grouped into thematic clusters. This categorization is achieved by clustering the web

documents based on the semantic similarity between the ontology terms that characterize them.

• **C-Logs Creation & Mining:** This module takes as input the web site's logs as well as the semantically annotated web site content. It outputs the semantically enhanced C-logs (concept logs) which are in turn used to generate both URI and category-based frequent itemsets and association rules. These rules are subsequently matched to the current user's visit by the recommendation engine.

• **Recommendation Engine:** This module takes as input the current user's path and matches it with the semantically annotated navigational patterns generated in the previous phases. The recommendation engine generates three different recommendation sets, namely, original, semantic and category-based ones, depending on the input patterns used.

The creation of the ontology as well as the semantic similarity measures used as input in the aforementioned web personalization process are orthogonal to the proposed framework. We assume that the ontology is descriptive of the web site's domain and is provided / created by a domain expert. In what follows we describe the key components of our architecture, starting by introducing the similarity measures we used in our work.

3.2 Content Characterization

A fundamental component of the Standing for Semantic Enhancement for Web Personalization architecture is the automated content characterization process. Our Personalization is the only web personalization framework enabling the automated annotation of web content with ontology terms without needing any human labeling or prior training of the system. The keywords' extraction is based both on the content of the web pages, as well as their connectivity features. What is more, our technique enables the annotation of multilingual content, since it incorporates a context-sensitive translation component which can be applied prior to the ontology mapping process. In the subsections that follow we describe in detail the aforementioned processes, namely, the keyword extraction, keyword translation and semantic characterization modules.

3.3 Keyword Extraction

There exists a wealth of methods for representing web documents, most of which have emerged from the area of searching and querying the web. The most straightforward approach is to perform text mining in the document itself following standard Information Retrieval (IR) techniques. This approach, however, has been shown insufficient for the web content, since it relies solely on the information included in the document ignoring semantics arising from the connectivity features of the web. It is difficult to extract keywords from web documents that contain images, programs etc. Additionally, many web pages do not include words that are the most descriptive ones for their content (for example rarely a portal web site includes the word "portal" in its home page). Therefore, in many approaches information contained in the links that point to the document and the text near them - defined as "anchor-window"- is used for characterizing a web document. This approach is based on the hypopaper that the text around the link to a page is descriptive of the page's contents and overcomes the problems of the content-based approach, since it takes into consideration the way others characterize a specific web page. In our work, we adopt and extend this approach, by also taking into consideration the content of the pages that are pointed by the page that is processed, based on the assumption that in most web pages the authors include links to topics that are of importance in the page's context.

More specifically, the keywords that characterize a web page p are extracted using:

1. raw term frequency of p
2. raw term frequency of a selected fraction (anchor-window) of the web pages that point to p (in links)
3. raw term frequency of the web pages that are pointed by p (out links) The three keyword extraction methods can be applied interchangeably or in combination. We should explain at this point the decision concerning term weighting phase, when the extracted keywords are given weights in order to use the most important ones. Term weighting, extensively used in the vector space model for document clustering, is carried out using several methods, such as raw term frequency. Raw term frequency is based on the term statistics within a document and is the simpler way of assigning weights to terms. The method used for collections of documents, i.e. documents that have similar content. In the case of a Web site however, this assumption is not always true since a Web site may contain documents that refer to different thematic categories (especially in the case of Web portals) and this was the reason for choosing raw term frequency as the term weighting method of our approach.

3.4 Keyword Translation

As already mentioned, the recommendation process is based on the characterization of all web documents using a common representation. Since many web sites contain content written in more than one language, this raises the issue of mapping keywords from different languages to the terms of a common

domain-ontology. For instance, our technique makes an implicit assumption of "one sense per discourse", i.e., that multiple appearances of the same word will have the same meaning within a document. This assumption might not hold in several cases, thus leading to erroneous translations. Our technique constitutes a first step toward the automated mapping of keywords to the terms of a common concept hierarchy; clearly, a more extensive study is required in order to provide a complete and more precise solution.

Procedure **translateW(Gr,En)**
1. K ← Ø ;
2. for all g ∈ Gr(D) do
3. for all s ∈ Sn(g) do
4. score[s] = 0;
5. for all w ∈ En(D)U Gr(D)-{g} do
6. sim = max(WPsim(s, Sn(w)));
7. score[s] += sim;
8. done
9. done
10. s_{max} = s';
(score[s'] = max(score[s]), s ∈ Sn(g))
11. K ← e, e ∈ En(g), e contains s_{max};
12. done

Figure 3 The keyword translation procedure

3.4 Semantic Characterization

In order to assist the remainder of the personalization process (C-logs cretion, semantic document clustering, semantic recommendations) the n most frequent (translated) keywords that where extracted in the previous phase, are mapped to the terms O = {c1, …, ck.} of a domain ontology (in our approach we need the concept hierarchy part of the ontology). This mapping is performed using a thesaurus. If the keyword belongs to the ontology, then it is included as it is. Otherwise, the system finds the "closest" (i.e. most similar) term (category) to the keyword through the mechanisms provided by the thesaurus. Since the keywords carry weights according to their frequency, the categories' weights are also updated.
We should stress here that the selection of the ontology influences the outcome of the mapping process. For this purpose, it should be semantically relevant to the content to be processed. In order to find the closest term in the ontology O for a keyword k that describes document, we compute the similarity between all senses of k, Sn(k) and all senses of all the categories c in O, Sn(ci). At the end of this process, each keyword is mapped to every category with a similarity s respectively. We select the (k,c) pair that gives the maximum similarity s. This process is shown in Figure 4.

Procedure **CategoryMapping(k, O)**
1. for all sns ∈ Sn(k) do
2. for all ci ∈ O do
3. $scsim_{max}$ ← $maxarg_{sc}$ ∈ Sn(ci)(WPsim(sns, sc));
4. done
5. $ssim_{max}$= max({$scsim_{max}$});
6. cmax = c ∈ O, for which ($scsim_{max}$ == $ssim_{max}$);
7. done
8. sim = max({$ssim_{max}$});
9. cat = c' ∈ {c_{max}}, for which ($ssim_{max}$ == sim);
10. return(cat, sim);
11. done

Figure 4 the semantic characterization process

3.5 Semantic Recommendations

Navigational patterns: We use the Apriori algorithm to discover frequent itemsets and/or association rules from the C-Logs, *CLg*. We consider that each distinct user session represents a different transaction. We will use $S = \{I_m\}$, to denote the final set of frequent itemsets/association rules, where $Im = \{(uri_i)\}$, $uri_i \in CLg$.
Recommendations: In brief, the recommendation method takes as input the user's current visit, expressed a set of URIs: $CV = \{(uri_j)\}$, $uri_j \in WS$, (*WS* is the set of the web site's URIs. Note that some of these may not be included in *CLg*). The method finds the itemset in *S* that is most similar to *CV*, and recommends the documents (labeled by related categories) belonging to the most similar document cluster $Cl_m \in Cl$ (*Cl* is the set of document clusters). In order to find the similarity between URIs, we perform binary matching (denoted as *SIM*). This procedure is shown in Figure 5.

Procedure **SemanticRec(CV)**
1. CM ← Ø ;
2. **Im** = maxarg**I**∈**S** SIM(I,CV);
3. for all d ∈ I_m do
4. for all c**j** ∈ d do
5. if c**j** ∈ CM then
6. r**j' +=** r**j;**
7. CM ← (c**j,**r**j'**);
8. else
9. CM ← (c**j,**r**j**);
10. done
11. done
12. return D = {d}, {d} ∈ Cl_m, maxarg**Cln**∈**Cl** WPsim(CL_m,CM);

Figure 5 the *semantic* recommendation method

The method finds the itemset in *C* that is most similar to *CV,* creates a generalization of it and recommends the documents (labeled by related categories) belonging to the most similar document cluster $Cl_n \in Cl$ (*Cl* is the set of document clusters). To find the similarity between categories we use the metric, whereas in order to find similarity between sets of categories, we The same procedure can be run by omitting the weights in one or all the phases of the algorithm. On the other hand, in case weights are used, an extension of the Apriori algorithm, which incorporates weights in the association rules mining process, can be used.

Procedure **CategoryRec(CV)**
1. I**k** = maxarg**I**∈**S** THEsim(I,CV);
2. for all c**j** ∈ CV do
3. c**i** = maxarg**c**∈**Ik** WPsim(c,c**j**);
4. c**n** = least_common_ancestor(c**i**,c**j**), r**n** = max(r**i**,r**j**);
5. CI ← (c**n**, r**n**);
6. done
7. return D = {d}, {d} ∈ Cl**n**, maxarg**Cln**∈**Cl** WPsim(CL**n**,CI);

Figure 6 the category-based recommendation method

Let us also stress that even though this description of the method focuses on sets' representation (derive frequent itemsets and use them in the recommendation method), it can also be applied (with no further modification) to the association rules that can be derived by those sets. If association rules are derived, then the user's activity is matched to the LHS of the rule (step 2), and recommendations are generated using the RHS of the rule (step 7).

So far, we have described the framework for enhancing the recommendation process through content semantics. Our claim, that the process of semantically annotating web content using terms derived from a domain-specific taxonomy prior to the recommendation process enhances the results of web personalization, is intuitive. Since the objective of the system is to provide useful recommendations to the end users, we performed an experimental study, based on blind testing with 15 real users, in order to validate the effectiveness of our approach. The results indicate that the effectiveness of each recommendation set (namely, *Original, Semantic, Category*), depends on the model, incorporating all three types of recommendations, generates the most effective results.

4. Methodology

Data Set: The two key advantages of using this data set are that the web site contains web pages in several formats (such as pdf, html, ppt, doc, etc.), written both in Greek and English and a domain-specific concept hierarchy is available (the web administrator created a concept-hierarchy of 150 categories that describe the site's content). On the other hand, its context is rather narrow, as opposed to web portals, and its visitors are divided into two main groups: students and researchers. Therefore, the subsequent analysis (e.g. association rules) uncovers these trends: visits to course material, or visits to publications and researcher details. It is essential to point out that the need for processing online (up-to-date) content, made it impossible for us to use other publicly available web log sets, since all of them were collected many years ago and the relevant sites' content is no longer available. Moreover, the web logs of popular web sites or portals, which would be ideal for our experiments, are considered to be personal data and are not disclosed by their owners. To overcome these problems, we collected web logs over a 1-year period (01/01/10 – 31/12/10). After preprocessing, the total web logs' size was approximately 105 hits including a set of over 67.700 distinct anonymous user sessions on a total of 360 web pages. The sessionizing was performed using distinct IP & time limit considerations (setting 20 minutes as the maximum time between consecutive hits from the same user).

Keyword Extraction – Category Mapping: We extracted up to 7 keywords from each web page using a combination of all three methods (raw term frequency, inlinks, outlinks). We then mapped these keywords to ontology categories and kept at most 5 for each page.

Document Clustering*:* We used the clustering scheme described in recent, i.e. the DBSCAN clustering algorithm and the similarity measure for sets of keywords. However, other web document clustering schemes (algorithm & similarity measure) may be employed as well.

Association Rules Mining*:* We created both URI-based and category-based frequent itemsets and association rules. We subsequently used the ones over a 40% confidence threshold.

4.1 Link Analysis for Web Personalization

The connectivity features of the web graph play important role in the process of web searching and navigating. Several link analysis techniques, based on

the popular PageRank algorithm [BP98], have been largely used in the context of web search engines. The underlying intuition of these techniques is that the importance of each page in a web graph is defined by the number and the importance of the pages linking to it. In this paper, we introduce link analysis in a new context, that of web personalization. Motivated by the fact that in the context of navigating a web site, a page/path is important if many users have visited it before, we propose a new algorithm UPR (Usage-based PageRank). UPR is based on a personalized version of PageRank, "favoring" pages and paths previously visited by many web site users. We apply UPR to a representation of the web site's user sessions, termed Navigational Graph in order to rank the web site's pages. This ranking may then be used in several contexts:

➔Use it as a "global ranking" of the web site's pages. The computed rank probabilities can serve as the prior probabilities of the pages when recommendations are generated using probabilistic predictive models such as Markov Chains, higher-order Markov models, tree synopses etc.

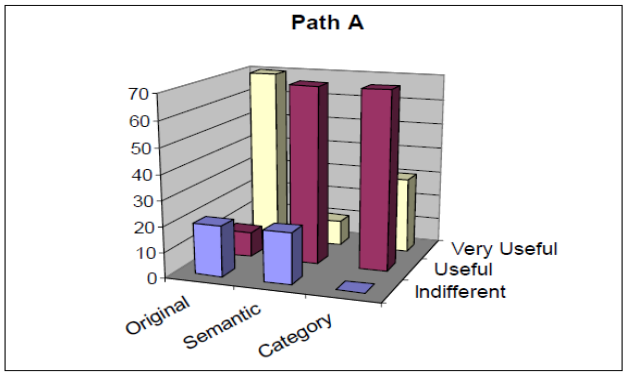
➔Apply UPR to small subsets of the web site's navigational graph (or its approximations), which are generated based on each current user's visit. This localized version of UPR (named l-UPR) provides localized personalized rankings of the pages most likely to be visited by each individual user. In what follows we illustrate our approach through a motivating example. We then provide the required theoretical background on link analysis before presenting the proposed algorithm. We prove that this hybrid algorithm can be applied to any web site's navigational graph as long as the graph satisfies certain properties. We then proceed with describing the two proposed frameworks in which UPR can be applied, namely, the localized personalized recommendations with l-UPR and the hybrid probabilistic predictive models (h-PPM). We conclude with an extensive experimental evaluation we performed on both frameworks (l-UPR and h-PPM), proving our claim that the underlying link structure of the web sites should be taken into consideration in the web personalization process, and details on the system prototype we used.

5. Results

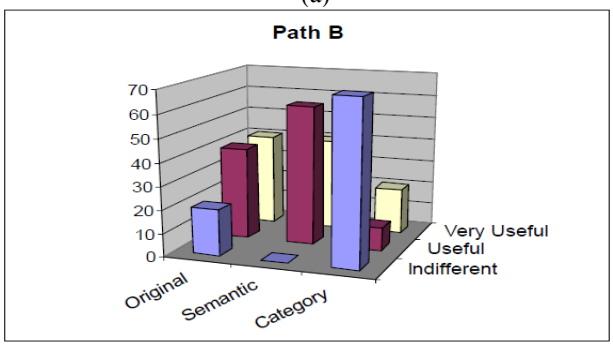
We created three different sets of recommendations named Original, Semantic, and Category (the sets are named after the respective recommendation methods). We presented the users with the paths and the three sets (unlabeled) in random order and asked them to rate them as "indifferent", "useful" or "very useful". The outcome is shown in Figure a, b, and c.

The results of the first experiment revealed the fact that depending on the context and purpose of the visit the users profit from different source of recommendations. More specifically, in visit A, both Semantic and Category sets are mostly evaluated as useful/very useful. The Category recommendation set performs better, and this can be explained by the fact that it's the one that recommends "hub" pages, which seems to be the best after a "random walk" on the site.

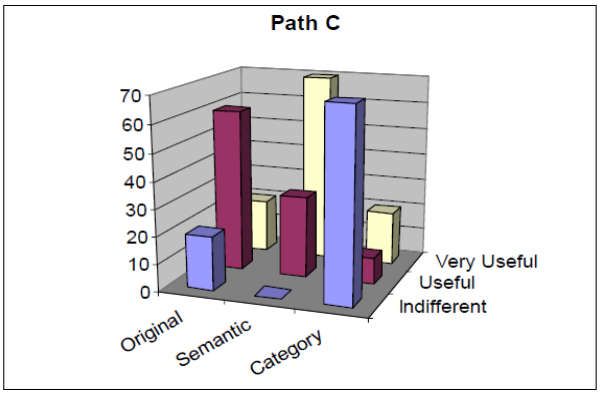
On the other hand, in visits B and C, Semantic performs better. In visit B, the path was focused to specific pages and the same held for the recommendations' preferences. In visit C the recommendations that were more relevant to the topics previously visited were preferred.



(a)



(b)



(c)

Figure 7 a, b, c: Recommendation sets' evaluation

Users had to select between the Category-based recommendation set and the Hybrid one. The outcome is shown in Figure d.

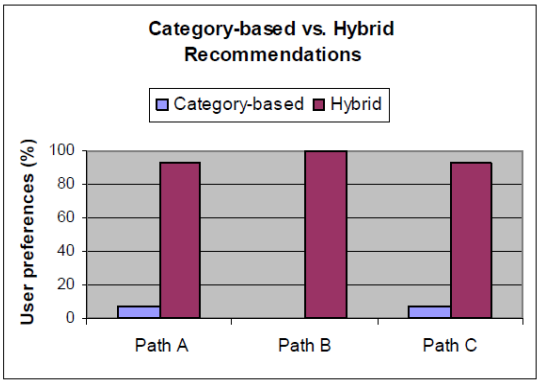


Figure7 d: Category-based vs. Hybrid Recommendations

The results of this experiment demonstrate the dominance of the Hybrid recommendation set over the Category-based one. One explanation for this would be that in the second case, important information may be lost during the generalization (convert user's current path to categories) back to specialization (convert categories to URIs) process.
Based on these experimental results, we observe that what is characterized as useful by the users depends on the objective of each visit. Out of the three possible recommendation sets, the Semantic recommendation set, generated after the semantic expansion of the most popular association rule performs better. Comparing all three recommendation sets with the Hybrid one, we observe that it dominates the other three, since the hybrid recommendations are preferred by the users in most cases. Therefore, we conclude that Standing for Semantic Enhancement for Web Personalization semantic enhancement of the personalization process improves the quality of the recommendations in terms of complying with the users' needs.

6. Conclusion

Most of the research efforts in web personalization correspond to the evolution of extensive research in web usage mining, i.e. the exploitation of the navigational patterns of the web site's visitors. When a personalization system relies only on usage-based results, however, valuable information conceptually related to what is finally recommended may be missed. Moreover, the structural properties of the web site are often disregarded. In this paper, we present novel techniques that incorporate the content semantics and the structural properties of a web site in the web personalization process. In the first part of our work we present a semantic web personalization system. Motivated by the fact that if a personalization system is only based on the recorded navigational patterns, important information that is semantically similar to what is recommended might be missed, we propose a web personalization framework that integrates usage data with content semantics, expressed in ontology terms, in order to compute semantically enhanced navigational patterns and effectively generate useful recommendations. The diversity of our specializations verifies the potential of our approach in providing an integrated framework for applications of link analysis to web personalization.

ACKNOWLEDGMENT

This research is supported by the Computer Science and Engineering department, SATI, Vidish

References

[1] Dario Vuljani, Lidia Rovan, and Mirta Baranovi, "Semantically Enhanced Web Personalization Approaches and Techniques", IEEE Proceedings of the *ITI 2010 32nd Int. Conf. on Information Technology Interfaces,* June 21-24, 2010, Cavtat, Croatia.

[2] Raymond Y.K. Lau, "Inferential Language Modeling for Selective Web Search Personalization and Contextualization", IEEE 2010 3rd International Conference on Advanced Computer Theory and Engineering(1CACTE).

[3] Esteban Robles Luna, Irene Garrigos, and Gustavo Rossi, "Capturing and Validating Personalization Requirements in Web Applications", IEEE 2010.

[4] Annappa B, K Chandrasekaran, K C Shet, "Meta-Level Constructs in Content Personalization of a Web Application", IEEE *Int'l Conf. on Computer & Communication Technology*-ICCCT'10.

[5] Pedro J. Muñoz-Merino, Carlos Delgado Kloos and Martin Wolpers, and Martin Friedrich, "An Approach for the Personalization of Exercises based on Contextualized Attention Metadata and Semantic Web technologies", 2010 10th IEEE International Conference on Advanced Learning Technologies.

[6] Xiaogang Wang and Yan Bai and Yue Li, "An Information Retrieval Method Based On Sequential Access Patterns", IEEE 2010 Asia-Pacific Conference on Wearable Computing Systems.

[7] Xiangwei Mu, Yan Chen and Taoying Li, "User-Based Collaborative Filtering Based on Improved Similarity Algorithm", IEEE 2010.

[8] Dimitris Antoniou, Mersini Paschou, Efrosini Sourla, Athanasios Tsakalidis, "A Semantic Web Personalizing Technique The case of bursts in web visits", 2010 IEEE Fourth International Conference on Semantic Computing.

[9] Rong Shan and Zhibin Ren, "Research on Personalized Recommendation System in E-Iearning", IEEE 2010 2nd International Conference on Education Technology and Computer (ICETC).

[10] Yan Gao, Bin Zhang, Shao-wei Shi, Hong-ning Zhu, Jun Na, Fu-cai Zhou, "A User Requirement-driven Service Dynamic Personalized QoS Model", IEEE 2010 Third International Conference on Dependability.

[11] Muhammad Shoaib, Amna Basharat, "Ontology based Knowledge Represeenation and Sematnic Profiling In Personalized Semantic Social Networking Framework", IEEE 2010.

AUTHORS PROFILE

Geetika S. Pandey presently working as Asst. Professor in Computer Science & Engineering at Samrat Ashok Technological Institute Vidisha M.P India. The degree of B.E. (Hons) secured in Computer Science & engineering. She secured M.Tech in Computer science and Engineering from Banasthali University. She is currently pursuing PHD in Computer science and engineering.

Mr. Mahendra Thakur is a research scholar pursuing M.Tech in Computer Science & Engineering from Samrat Ashok Technological Institute Vidisha M.P India. He secured degree of B.E. in IT from Rajiv Gandhi Technical University, Bhopal (M.P.) India in 2007.