ISSN No. 0976-5697

Volume 2, No. 5, Sept-Oct 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Efficient personalization and Recommendation Methods for Web Mining

Mahendra Pratap Singh Dohare*
M.Tech (Software System)
Samrat Ashok Technological Institute
Vidisha, M.P., India
mps.mt32@gmail.com

Premnarayan Arya
M.Tech (Software System)
Samrat Ashok Technological Institute
Vidisha, M.P., India
premarya_scs@yahoo.com

Vinod Kumar
M.Tech (Software System)
Samrat Ashok Technological Institute
Vidisha, M.P., India
ishere.vinod@gmail.com

Abstract: With the fast development of World Wide Web, Web-based applications and services should allow user to get the right personalized information quickly and effectively. Collaborative Filtering acts a very important role in web service personalization and Recommender System. 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 thesis, 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. Our experimental results show more efficient than existing works.

Key words: Web personalization, Semantic web and Recommender systems.

I. 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 thesis 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. 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 thesis 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].

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].

II. BACKGROUND

A. 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].

B. Integrating Content Semantics in Web Personalization:

Several frameworks supporting the claim that the incorporation of information related to the web site's content enhances the web personalization process have been proposed prior or subsequent to our work. In this Section we overview in detail the ones that are more similar to ours, in terms of using a domain-ontology to represent the web site's content.

Dai and Mobasher proposed a web personalization framework that uses ontologies to characterize the usage profiles used by a collaborative filtering system. These profiles are transformed to "domain-level" aggregate profiles by representing each page with a set of related ontology objects. In this work, the mapping of content features to ontology terms is assumed to be performed either manually, or using supervised learning methods. The defined ontology includes classes and their instances therefore the aggregation is performed by grouping together different instances that belong to the same class. The recommendations generated by the proposed collaborative system are in turn derived by binary matching of the current user visit, expressed as ontology instances, to the derived domain-level aggregate profiles, and no semantic similarity measure is used. The idea of semantically enhancing the web logs using ontology concepts is independently described in recent. This framework is based on a semantic web site built on an underlying ontology. The authors present a general framework where data mining can then be performed on these semantic web logs to extract knowledge about groups of users, users' preferences, and rules. Since the proposed framework is built on a semantic web knowledge portal, the web content is already semantically annotated (through the existing RDF annotations), and no further automation is provided. Moreover, the proposed framework focuses solely on web mining and thus does not perform any further processing in order to support web personalization.

In recent work also propose a general personalization framework based on the conceptual modeling of the users' navigational behavior. The proposed methodology involves mapping each visited page to a topic or concept, imposing a concept hierarchy (taxonomy) on these topics, and then estimating the parameters of a semi-Markov process defined on this tree based on the observed user paths. In this Markov models-based work, the semantic characterization of the content is performed manually. Moreover, no semantic similarity measure is exploited for enhancing the prediction process, except for generalizations/specializations of the ontology terms. Finally, in a subsequent work, explore the use of ontologies in the user profiling process within collaborative filtering systems. This work focuses on recommending academic research papers to academic staff of a University. The authors represent the acquired user profiles using terms of research paper ontology (is-a hierarchy). Research papers are also classified using ontological classes. In this hybrid recommender system which is based on collaborative and content-based recommendation techniques, the content is characterized with ontology terms, using document classifiers (therefore a manual labeling of the training set is needed) and the ontology is again used for making generalizations/specializations of the user profiles [8] and [9].

C. 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].

III. PROPOSED 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.

A. 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.

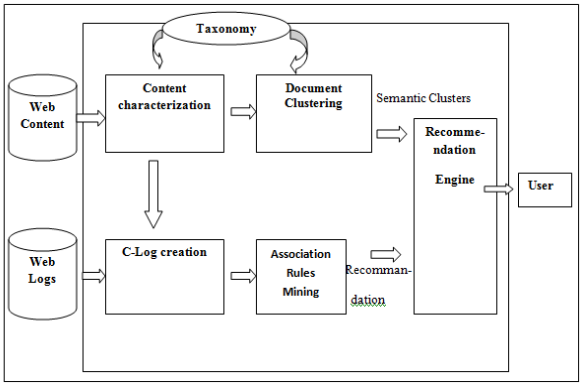

Figure 1 Standing for Semantic Enhancement for Web Personalization and Recommendation architecture

As illustrated in Figure 1, Standing for Semantic Enhancement for Web Personalization consists of the following components:

- ***a. 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.
- ***b. 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. 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.
- ***d. 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.

B. Proposed Methodology:

- ***a. 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).

b. ***Keyword Extraction:*** 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.

c. ***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.

d. ***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.

C. 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 thesis, 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:

a. 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.

b. 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.

IV. 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, 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.

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.

V. 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 thesis, 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.

VI. REFERENCES

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

[2]. 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.

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

[4]. 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.

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

[6]. 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.

[7]. 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).

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

[9]. 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.

[10]. 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.

[11]. 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.