A Novel Approach towards Integration of Semantic Web Mining with Link Analysis to Improve the Effectiveness of the Personalized Web

Chanchala Joshi
Institute of Computer Science
Vikram University, Ujjain, M.P. India

Umesh Kumar Singh
Institute of Computer Science
Vikram University Ujjain, M.P. India

ABSTRACT
During the past few years World Wide Web has become a main source of information acquisition. The existence of such abundance of information, in combination with the dynamic and heterogeneous nature of the web, makes web site exploration a difficult process for the user. Websites personalization is the effective way to meet the requirement of efficient web navigation. This paper proposed novel technique that uses the content semantics and the structural properties of a web site in order to improve the effectiveness of web personalization. This paper presents a personalization framework CUMPW (Content & Web Usage Mining for Personalized Web) that integrates web content and web usage data with the user's navigational patterns and represents the correlation between contents and the usage of the website. In the second part of proposed method, this paper presents a novel approach for enhancing the quality of recommendations based on the underlying structure of a web site. This paper proposed Navigational PageRank (NPR) Algorithm that suggests link analysis in effective manner for web personalization. NPR is applied to navigational graph of user session in order to determine the importance of a web page. The proposed hybrid (CUMPW + NPR) framework provides more representative predictions results than existing techniques that rely solely on usage data.

Keywords
Web usage mining, navigational pattern, link analysis, personalized web

1. INTRODUCTION
World Wide Web has become the biggest and most popular way of communication and information retrieval. It serves as a platform for exchanging various kinds of information, ranging from research papers and educational content to multimedia content, software and personal blogs. 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.

The web personalization is the customization of information or services provided by a website to users based on knowledge acquired by their navigational behavior, recorded in website's log [1]. This information is combined with the content and the structure of the web site, as well as the interests of the user. The personalization process can generate dynamic results, which depends on behavior of user. This paper proposed a prominent technique that uses the content semantics and the structural properties of a web site in order to improve the effectiveness of web personalization. The proposed methodology is broadly divided into two parts. The first part provides a novel framework CUMPW (Content & Web Usage Mining for Personalized Web) for website personalization that uses association rules and sequential patterns on user's navigational data extracted from Web server logs to predict user visit patterns. The second part of proposed method, presents an approach for enhancing the quality of recommendations based on the underlying structure of a web site. This part proposed Navigational PageRank (NPR) Algorithm that suggests link analysis in effective manner for web personalization. NPR is applied to navigational graph of user session in order to determine the importance of a web page. The proposed hybrid framework provides more representative predictions results than existing techniques that rely solely on usage data.

2. METHOD
A website can be personalized in various ways, such as the creation of new index pages, personalized search services, or dynamic recommendations generation. This paper, deal with the link prediction that might be distinct for each specific visitor according to their interest. The link prediction followed two steps method, first web usage mining that obtained more relevant pages to be visited by user according to user's interest, then link analysis that result the probability of visiting the page by navigation behavior obtained by web log.

2.1 Web Usage Mining
In a website user mainly navigate through website's content, which means user's navigation is typically content-driven. The users usually search for information concerning a particular topic. Therefore, the content should be a prominent factor in the process of website personalization. This paper proposed a framework CUMPW (Content & Web Usage Mining for Personalized Web) that integrates web usage data and web content with the user's navigational patterns and represents the correlation between contents and the usage of the website.

The purpose of Content mining is to fetch the knowledge hidden in the log files of a Web server. By applying statistical and data mining methods to the Web log data, interesting patterns related to the user's navigational behavior can be identified. That represents the possible correlations between Web pages and user's groups.

Assume a user navigates through the pages of web portal, specializing in computer trends and technology. This user is

interested in latest technologies and looking for a notebook. Therefore he searches to find any information available. Since the amount of information available on web are very large and are not necessarily properly organized. Based on user's navigation, however, in combination with previous user's visits focusing on same object, the system makes recommendations to the users.

Suppose, there exists a web portal, a specialized space for latest IT trend and technology, called "thetechnoworld". This portal contains information about various resources about technologies ranging from latest technology news to advertisements on latest model and technical events. Now suppose, for example many users in the past have seen the page www.thetechnoworld.com/computing/notebooks then follows
www.thetechnoworld.com/computing/notebooks_accessoies.

If the current user visits the first page the system can recommended the second one, based on assumption that people with similar interest have similar navigational behavior.

The proposed framework follows Association Rule Mining technique. The personalization system of "thetechnoworld" applies association rules mining on its web logs in order to generate recommendations to its visitors, based on the assumption that users with similar interests have similar navigational behavior. Assume that one of the discovered patterns is the following:
www.thetechnoworld.com/computing/notebooks
www.thetechnoworld.com/computing/notebooks_accessoies
This pattern represents that "people that are interested in notebook and search for it, will probably be interested in purchasing notebook accessories". Based on the assumption that user is interested in finding a notebook and using personalization, next time when navigates through "thetechnoworld" and visit first web page, the personalized site will dynamically recommend to the next page.

The "thetechnoworld" content however, is continuously updated. Suppose that the notebook accessories department has just announced a sale on notebook arms and stands
www.thetechnoworld.com/computing/ notebooks_accessoies /sale_notebook-arms-stands
Since this is a new web page, it isn't included in the web logs, or is included in very low ratio, no one or only a few users have visited this page, therefore is definitely not included in the derived association rules comprising our navigational model. As a consequence, if the "traditional" usage-based personalization process is follows, it will never be recommended to user even though it is apparent that it is very similar to their search intentions. It shows that pure usage-based personalization is problematic in several cases. It is considerable that information conceptually related to the user's visit should not be "missed", and introduce the CUMPW personalization system that addresses the aforementioned shortcomings by generating semantically enhanced recommendations.

2.2 Proposed CUMPW Architecture

CUMPW uses a combination of web mining techniques to personalize a website. The website's contents are processed and then characterized by a set of mining rules. The visitor's navigational behavior is also updated with this semantic knowledge to create an enhanced version of web logs as well as semantic document clusters. Web logs are in turn mined to generate both a set of recommendations 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 personalized web to the end user.

Figure 2 shows the architecture of CUMPW framework

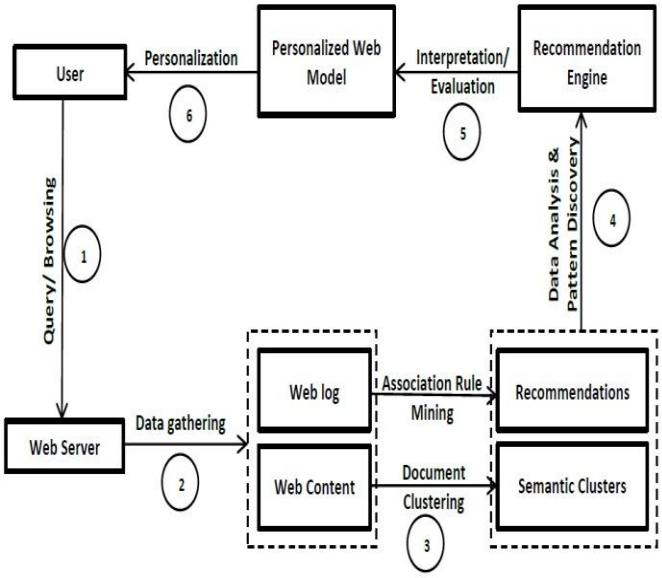

Figure2. CUMPW Architecture for Web Personalization

As illustrated in Figure 2, CUMPW consists of the following components:

1. Data gathering: Web logs, which record user activities on Web sites, provide the most comprehensive, detailed Web usage data. Web contents represent the user interest.
2. Content Characterization: This module takes the content of the web site as an input, on which clustering rules are applied to generate semantic clusters. The content characterization process consists of the keyword extraction, keyword translation and semantic characterization.
3. Semantic Document Clustering: The semantically related pages created by the previous component are grouped into clusters. This categorization is achieved by clustering the web documents based on the semantic similarity between the ontology terms that characterize them.
4. Web Logs Mining: This module takes as input the website's logs. After applying category-based frequent itemsets and association rules on the web log, it outputs the semantically enhanced log.
5. Recommendation Engine: This module takes as input the current user's path and matches it with the navigational patterns generated in the previous phases.

CUMPW is based on the integration of content semantics with the user's navigational behavior in order to generate recommendations. The web site's documents are mapped to ontology terms, enabling further processing (clustering, association rules mining, recommendations' generation) to be performed based on the semantic similarity between these terms. Using this representation, the final Web Model presented to the user is semantically enhanced.

2.3 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 PageRank algorithm [11] and have been largely used PageRank algorithm in the context of web search engines. The basic principle 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.

This paper proposed link analysis in a new context of web personalization. Based on the fact that in the context of navigating a web site, a page is important if many users have visited it before.Consider a web graph in Figure, here nodes represent the webpages and edges represent the link between them.

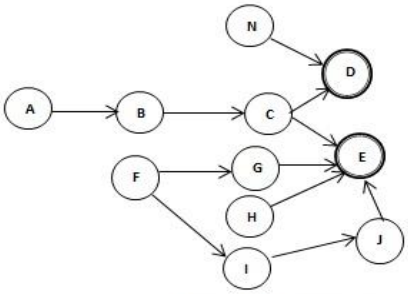

Figure2. Web Graph

Suppose that a user has already visited pages A, B and C. The aim is to predict the most probable path that user will follow next, either C→D or C→E. Here node E is linked by more pages than D. E has four inlink and no outlink, while D has one inlink and one outlink. According to Pagerank algorithm page E is more important, which means probability to visit page E is high. PageRank algorithm is based on the assumption that page which is pointed by more pages has higher priority to be visited. That means page having more incoming links has higher ranking. This ranking has been very useful in the context of web search.

Now revise the web graph with the additional concept of weight. Here weights on edge represent the number of visits by users.

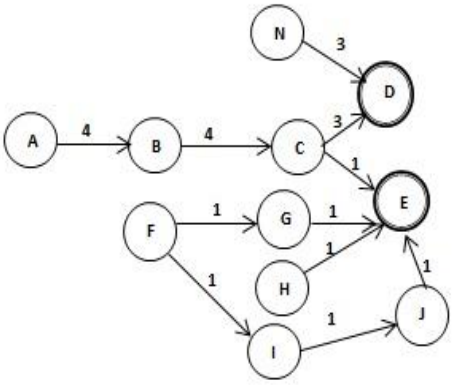

Figure3. Web Graph with navigational weight

Above figure represents that page D has visited by more people than page E, therefore it may claim that in this navigational web graph C → D seems to be more important than C→E in terms of user's interest. Therefore, PageRank algorithm may extend with the claim that, in the web navigation context, a page is considered important if many users have visited it before. On the basis of this assumption, this paper proposed a personalized page rank algorithm based on user's navigational behavior. Proposed algorithm represents the importance of the web site's pages both in terms of link connectivity, as well as their visit frequency.

2.3.1 Algorithm's Preliminaries

The input to the proposed algorithm is the Navigational Graph (NG). NG is a weighted directed graph that represents the user sessions. NG can be used to determine page and path probabilities. Since it contains all the distinct user sessions, therefore it is a representation of the actual user paths followed in the past.

The proposed algorithm obtained Navigational Tree (NT) from NG. One can consider NG as a tree with nodes and labeled-edges. Nodes of NG represent the web page and the edges between nodes represent the links between web pages, i.e. the paths followed by the users. Edge labels (weights) in NG represent the number of link traversals by users in past. Let the NG has a special node R termed as root node, represents the initial point of user's visit. The weighted paths from the root towards the leaves represent all the user's session paths that are included in the web logs. All tree paths terminate in a special leaf-node Φ denoting the end of a path. The NT creation algorithm is as follows: Every user session US in the web logs creates a path starting from the root of the tree. If a subsequence of the session already exists it updates the weights of the respective edges, otherwise creates a new branch, starting from the last visited common page in the path.

2.3.2 Algorithm for generating Navigational Tree

[This algorithm scans all nodes of Navigation Graph (NG) and returns a Navigational Tree (NT) for each User Session (US)]

1. root := NG
2. temp := root
3. W := 0
4. Repeat For each US ϵ U

   While temp ≠ Φ do

   a. Node= first_state(US)

   b. If Parent(temp)=Node then

      a. W:= W+1

      b. Temp:=Node

   Else

      a. Temp->child:=Count

      b. W:=1

      c. Temp:= Node

   End if
5. Exit

Above algorithm can be described by a simple example. Assume that the user sessions of a web site are included in the following Table.

Table -User Sessions

US_No	Path
1	A→B→C→D
2	B→C→D→E
3	A→B→C→E

4	A→B→F→G
5	B→F→G→H
6	A→F→I→J
7	A→N→D→C

The Navigational Tree created after applying the algorithm is shown in Figure.

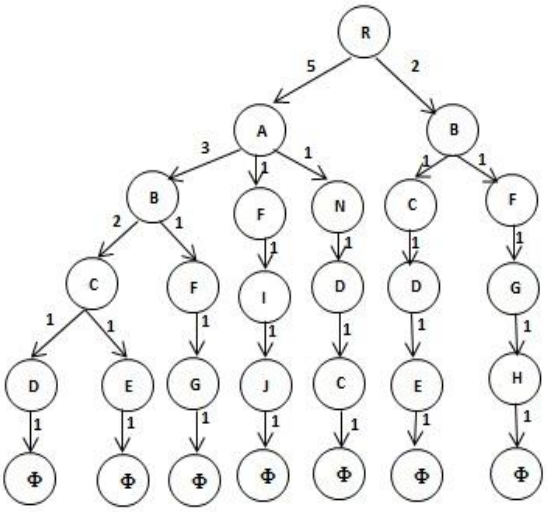

Figure4. Navigational Graph

2.3.3 Navigation based PageRank (NPR) Algorithm

The proposed Navigation based PageRank algorithm is based on the most popular PageRank algorithm of link analysis, including the user's navigational behavior.

2.2.3.1 PageRank Algorithm

PageRank algorithm was proposed by L. Page and S. Brin [11] and is the most prominent link analysis algorithm used for web search engine. PageRank is a "vote", by all the other pages on the Web, about how important a page is. A link to a page counts as a vote of support. If there is no link towards a page, there is no support (but it is an abstention from voting rather than a vote against the page).

Quoting from the original Google paper, PageRank is defined like this:

> It is assumed that page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a dumping factor which can be set between 0 and 1. Usually d is set to 0.85. Also C (A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:
>
> PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))
>
> Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one.

> PageRank or PR(A) can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the Web.

2.2.3.2 Proposed NPR Algorithm

Based on the concept that a web page is important if many users have visited it in past, a new link analysis algorithm, NPR (Navigation based PageRank) is proposed, which is based on user's navigational behavior. NPR extends the PageRank algorithm, by combining the page ranking with knowledge acquired from previous user visits, which are recorded in the user sessions. The proposed algorithm results the ranking of web pages that is related to the frequency of visits to them.

Section 2.2.2, defines the directed navigational graph NG, where the nodes represent the web pages of the website and the edges represent the paths followed by previous users. Both nodes and edges contain weights. The weight W_i on each node represents the number of times page X_i was visited and the weight $W_{i\to j}$ on each edge represents the number of times X_j was visited immediately after X_i. NPR algorithm denotes the set of pages pointing to X_i (incoming links) as $In(X_i)$ and the set of pages pointed to by X_i (outgoing links) as $Out(X_i)$.

Definition: Navigation based PageRank Algorithm is the recursive algorithm and can be defined by following formula:

$$NPR(X) = d\left(\sum_{i=1}^{n} \frac{NPR(In(X_i))}{Out(X_i)} \times \frac{W_{in\to X}}{W_{out\to X}}\right) + (1-d)W_X$$

The proposed algorothm uses the basic idea of PageRank that page having more inlinks has more chances of visiting, while outlink increases the chance of dumping the page and to navigate another. NPR bias the PageRank calculation to assign a higher rank to the pages that were visited more often by users in the past. Further, this hybrid ranking combined with the structure and the usage data of the site, to provide a ranked recommendation set to current users.

3. CONCLUSION

This paper has shown the integration of content semantics and link analysis techniques can improve the recommendation process. The proposed framework can be used to discover interesting user navigation patterns which then can be applied to real world problems such as website improvement, additional topic recommendations, customer's behavior study etc. Web personalization system not only provides user a set of personalized pages but also gives user a list of domains the user may be interested in. Thus user can switch to different interests when surfing on the web for information. Besides this web personalization has increased the accuracy of recommendations very significantly.

Our future plans involve the location based website personalization. Further work will bias the location factor with the proposed Navigation based PageRank (NPR) algorithm to generate more efficient website personalization.

4. REFERENCES

[1] Eirinaki M., Lampos H., Vazirgiannis M., Varlamis I., "SEWeP: Using Site Semantics and a Taxonomy to Enhancethe Web Personalization Process", in Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2003), Washington DC (2003).

[2] Cooley R., Mobasher B. and Srivastava J, "Data preparation for mining World Wide Web browsing patterns", Journal of Knowledge and Information Systems 1, (1), 1999, pp 05-32.

[3] Chen, M., Park, J. and Yu, P. "Efficient data mining for path traversal patterns," in IEEE Transactions on Knowledge and Data Engineering Vol 10, No2, March/April 1998 pp 209-221.

[4] Jalali, M., et al. "A new clustering approach based on graph partitioning for navigation patterns mining," in International Conference on Pattern Recognition, 2008, pp.1-4.

[5] Sujatha, V. and Punithavalli. "Improved User Navigation Pattern Prediction Technique From Web Log Data," in International Conference on Communication Technology and System Design, 2001, pp.92-99.

[6] Tug, E., Sakiroglu, M. and Arslan, A. "Automatic discovery of the sequential accesses from web log data files via a genetic algorithm," in Knowledge Based Systems, 2006, pp.180-186.

[7] Kim, S. and Zhang, B. "Genetic mining of HTML structures for effective web document retrieval," in Applied Intelligence 18, 2003, pp.243-256.

[8] Mobasher, B., et al., "Integrating web usage and content mining for more effective personalization," in First International Conference on Electronic Commerce and Web Technologies, 2000, pp.165-176.

[9] Sarukkai, R.R. "Link prediction and path analysis using Markov chains," in 9th World Wide Web conference, 1999.

[10] Kaur C.,Aggarwal R.R., "Reference Scan Algorithm for Path Traversal Patterns", International Journal of Computer Applications 48(7), June 2012 pp. 20-25.

[11] Page L., Brin S., Motwani R., and Winograd T. "The Pagerank Citation Ranking: bringing order to the Web", Technical report, Stanford Dig. Lib. Tech. Project, 1998 pp.1-17.

[12] M.S. Aktas, M.A. Nacar, F. Menczer, "Personalizing PageRank Based on Domain Profile"s, in Proc. of the 6th WEBKDD Workshop, Seattle, Washington, USA , August 22 2004 pp 83-90.

[13] Sharma A., Kumar S., Singh M., " Semantic Web Mining for Intelligent Web Personalization", Journal of Global Research in Computer Science, Volume 2, No. 6, June 2011, pp 77-81.