WEB-PAGE RECOMMENDATION BASED ON CONCEPTUAL PREDICTION MODEL SUPPORTED BY WEB-SEMANTICS AND SESSION TIME

[1]V.Thanammal Indu, [2]M.Gokuldhev

[1]Assistant Professor, [2]Assistant Professor
[1]Department of Computer Science and Engineering,
[1]Amrita College of Engineering and Technology, Kanyakumari, India

Abstract: Web-page recommendation plays an important role in intelligent Web systems. Domain ontology represents the semantics of web-pages of a website. Web usage data gives knowledge about user's browsing patterns. Web semantics mining defines the context of a web page. The time spent on a web page indicates the interest level of a user. This paper proposes a system which evaluates user's interest based on combination of semantic enhancement, web usage data, content of web pages and the time spent by users on web pages. Domain ontology and semantic network are proposed to represent the domain knowledge; a conceptual prediction model is proposed to represent a semantic network of the semantic web usage knowledge. Term frequency keyword extraction method is proposed to perform semantic mining. A number of effective queries have been developed. Based on these queries, recommendation strategies have been proposed.

***Index Terms* - Web-page recommendation, Web usage mining, Semantic mining, Domain ontology**

I. INTRODUCTION

The objective of a Web-page recommender system is to effectively predict the Web-page or pages that will be visited from a given Web-page of a website. With the immense increase in the number of websites and Web-pages on the internet, the concern of suggesting users with the web pages in the area of their interest needs to be addressed as finest as possible. Various approaches have been proposed over the years by many researchers and each of them has taken the solution of creating personalized web recommendations one step ahead.

Earlier approaches were based on tree structures and probabilistic models, which can efficiently represent Web access sequences (WAS) in the Web usage data [2]. The predicted pages are limited within the discovered Web access sequences, i.e., if a user is visiting a Web-page that is not in the discovered Web access sequence, then these approaches cannot offer any recommendations to this user. This problem is referred to as "new-page problem". Semantic-enhanced approaches are effective to overcome new-page problem [1], [3], [4].

Domain ontology is commonly used to represent the semantics of Web-pages of a website. Web Usage Mining (WUM) aims to discover potential knowledge hidden in the web browsing behavior of users. Integrating semantic information with Web usage mining achieved higher performance in web recommendations than classic Web usage mining algorithms [3]-[7].

Web semantics, which defines the context of a web page, is an equally important concept to be considered [10]-[12]. Web semantics mining finds similar web pages based on the content of the web pages and clusters them together.

To improve the web recommendations, the time spent on web pages are taken into consideration in addition to using web usage and web semantics mining [8], [9]. Generally users will spend more time on a web page when they find the content of the web page interesting. Thus considering time spent narrows the suggestions list while making it even more focused to the current user's interest.

This paper presents a novel method to provide better Web-page recommendation based on domain knowledge, Web usage, Web semantics, and the session time. This is supported by three knowledge representation models [1], keyword extraction algorithm and a set of Web-page recommendation strategies. The first model is an ontology based model that represents the domain knowledge of a website. The construction of this model is semi-automated so that the development efforts from developers can be reduced. The second model is a semantic network that represents domain knowledge, whose construction can be fully automated. This model can be easily incorporated into a Web-page recommendation process because of this fully automated feature. The third model is a conceptual prediction model, which is a navigation network of domain terms based on the frequently viewed Web-pages and represents the integrated Web usage and domain knowledge for supporting Web-page prediction. The construction of this model can be fully automated.

A web page holds text, image, video and/or audio contents. For the purpose of mining, normally only the text of the web pages is considered. One of the most popular ways to determine what the text of a document is about is by using term frequency where term is the keyword (word of relevance) in the document [12]. The keywords from different Web-pages are categorized using a similarity measure. Instead of considering the time spent as it is, associating time slots with the web pages would be more sensible [9].

The recommendation strategies make use of the domain knowledge and the prediction model through two of the three models to predict the next pages with probabilities for a given Web user based on his or her current Web-page navigation state. To a great extent, this new method also includes the web semantics knowledge and time spent by the users from the web logs. This method yields better performance compared with the existing Web usage and Web semantics based Web-page recommendation systems.

II. Related work

We can classify the research work into three categories.

2.1. Traditional Approaches that use Sequence Learning Models

Association rules and probabilistic models have been commonly used in applying sequence learning models to Web-page recommendation. Pre-Order Linked WAP-Tree Mining (PLWAP-Mine) [13], are outstanding in supporting Web-page recommendation, compared with other sequence mining algorithms.

2.2. Semantic-Enhanced Approaches

Semantic information can be integrated into Web-page recommendation models. By making use of the ontology of websites, Web-page recommendation can be enriched and improved significantly in the systems. Depending on the domain of interest in the system, we can reuse some existing ontologies or build a new ontology, and then integrate it with Web mining. [14-16]

2.3. Web Mining

Web Mining aims to discover useful information from web hyperlinks, page contents and usage logs. Based on the data, web mining is categorized as web structure mining, web content mining and web usage mining. Web structures discovers knowledge from hyperlinks which represents structure of web site, web content mining discovers knowledge from web page contents and Web usage mining mines user access patterns from usage logs which record clicks made by every user [17],[18]. This helps to understand user's interest in a web site. The discovered patterns are usually represented as collections of pages, objects or resources that are frequently accessed by groups of users with common needs or interests.

III. Proposed Work

An efficient Web-page recommendation system is proposed which integrates a conceptual prediction model, semantic enhanced recommendation with the web semantics and the time spent on the web-pages. While web usage data gives insights to the users' browsing patterns and web semantics gives insights to the web page's content, the time spent on a web page indicates the interest level of a user for the area that the web page covers.

3.1. Preliminary Concepts

3.1.1. Domain ontology

It is defined as a conceptual model that specifies the terms and relationships between them explicitly and formally, which in turn represent the domain knowledge for a specific domain. The three main components are listed as follows:

i. *Domain terms (concepts),*
ii. *Relationships between the terms (concepts),*
iii. *Features of the terms and relationships.*

3.1.2. Semantic network

Semantic Network of a website is a kind of knowledge map which represents domain terms, Web-pages, and relations including the collocations of domain terms, and the associations between domain terms and Web-pages.

3.1.3. Web Semantics

It defines the context of a web page. Web semantics mining finds similar web pages based on the content of the web pages and clusters them together. This helps in the recommendation of those pages that were not browsed by previous users because they are not structurally linked to the browsed pages.

In order to make better Web-page recommendations, we need semantic Web usage knowledge which can be obtained by integrating the domain knowledge model or the semantic network with Web usage knowledge that can be discovered from Web log files using a Web usage mining technique. Additionally the web semantics information and the time spent by the users on a particular web-page are also considered to narrow down the recommendations.

3.2. System Architecture

Domain Ontology specifies the terms and their relationships. Ontology is constructed based on the titles of visited Web-pages. Semantic Network represents domain-terms, web-pages and relations. It infers how closely the Web-pages are semantically related using terms. The domain knowledge or semantic network is integrated with Web usage knowledge. Conceptual Prediction Model is designed to automatically generate a weighted semantic network. Figure.1. illustrates the architecture of the proposed system.

Web semantics, which defines the context of a web page, is an equally important concept to be considered. Web semantic mining aims at mining the content of web pages and finding similarity between web pages based on the content. A web page can have text, images, video and/or audio as its content. For the purpose of mining, considering only the text of the web pages helps the most. One of the most popular ways to determine what the text of a document is about is by using term frequency where term is the keyword (word of relevance) in the document.

The inclusion of the additional dimension, time spent on page, appears to improve the recommendations made. Users often spend considerable time browsing the web pages for getting the right information. If the users' intention and interest for browsing a web page can be identified, it will be easier to make available that area of information with higher priority.

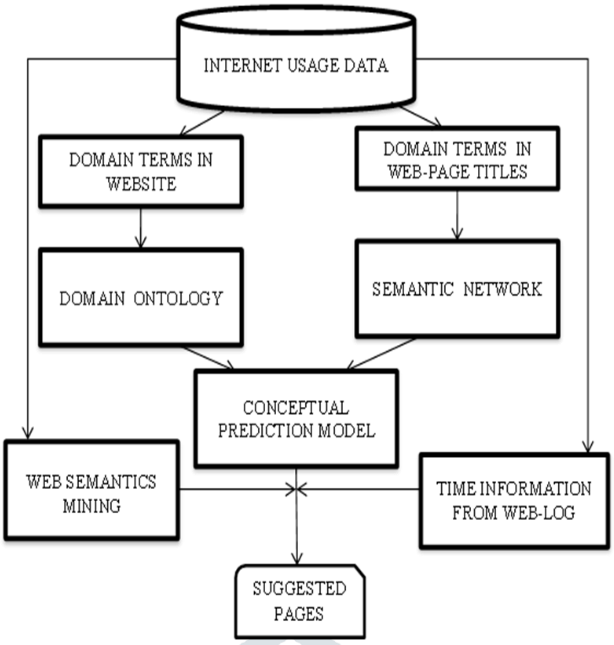


Figure.1. System Architecture of the proposed work

IV. IMPLEMENTATION

In the context of Web-page recommendation, the input data is Web logs that record user sessions on a daily basis. The user sessions include information about users' navigation details and time spent by users on each web-page. Each Web-page has a title, which contains the keywords that hold the semantics of the Web-page.

This section now presents procedures for constructing the semantic models using the Microsoft (MS) website (www.microsoft.com) as an example. Figure.2. explains the different models used in the recommendation system.

4.1. Domain Ontology Construction

The ontology will be constructed based on the titles of visited Web-pages so that it is the domain knowledge perceived by users. The procedure for constructing the domain ontology:

- Collect the terms from titles of Web-pages
- Define concepts of website based on extracted terms
- Define relationship between concepts
- Express as a model of set of domain terms, Web-pages and their relationships.

4.2. Semantic Network Construction

Semantic network of a website is a kind of knowledge map which represents domain terms, Web-pages, and relations including the collocations of domain terms, and the associations between domain terms and Web-pages.

The procedure for constructing the semantic network:

- Collect the titles of visited Web-pages
- Extract term sequences from the Web-page titles
- Build semantic network based on the term sequences and their occurrence weights

Queries are developed to retrieve the terms associated with a Web-page and Web-pages associated with a term from both the models.

4.3. CPM Construction

The Conceptual Prediction Model (CPM) integrates the domain or semantic knowledge with web usage mining data. CPM automatically generates a weighted semantic network. Several queries are developed on the knowledge databases. The recommendation strategies combine all the three models and suggest web-pages. The procedure for constructing CPM:

- The FWAP is generated using PLWAP-Mine algorithm
- Integrate FWAP with DomainOntoWP or TermNetWP to generate FVTP.
- Generate weighted semantic network, weight being the probability of transition between two adjacent terms based on FVTP
- First order transition probability based on transition from tx to ty
- Second order transition probability based on transitions (ty, tz) from (tx, ty)

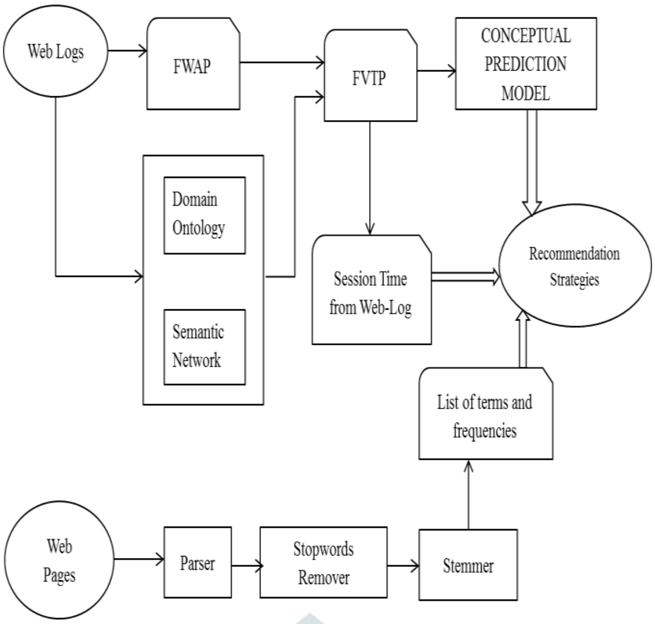

Web Logs
FWAP
FVTP
CONCEPTUAL PREDICTION MODEL
Domain Ontology
Semantic Network
Session Time from Web-Log
Recommendation Strategies
List of terms and frequencies
Web Pages
Parser
Stopwords Remover
Stemmer

Figure.2. Implementation of Web-Recommendation System

4.4. Semantic Mining - Keyword Extraction

Web semantic mining aims at mining the content of web pages and finding similarity between web pages based on the content. One of the most popular ways to determine what the text of a document is about is by using term frequency where term is the keyword (word of relevance) in the document.

For web semantic mining, list of categories is formed based on the area of browsed web pages. The text from web pages is extracted and processed to get all the keywords and their frequencies. These keywords are mapped to the categories using a similarity measure. This mapping enables determining which categories a web page belongs to.

Once the similarity measure of each document to all the categories has been evaluated, the documents are then clustered to bring together pages with similar content. This process helps to cluster together those pages which are contextually similar but not structurally connected to each other. Thus, even though previous users may not have visited similar pages not connected via links, semantic clustering brings together such pages and recommends them to the current user.

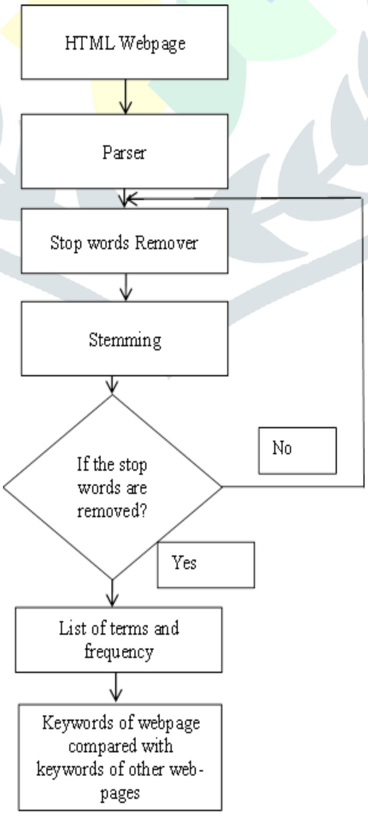

HTML Webpage
Parser
Stop words Remover
Stemming
If the stop words are removed?
No
Yes
List of terms and frequency
Keywords of webpage compared with keywords of other web-pages

Figure.3. Term Frequency Keyword Extraction Framework

Figure.3. shows the keyword extraction framework. The steps of the proposed method pass through three main modules which are loader/parser, Stopword remover and Stemmer. The input of the proposed approach is the webpage and the output is list of words and their frequencies.

The Parser module is the first step in which a parser is a program that breaks large units of data into smaller pieces which are called (tokenzs). After the webpage is fed to the system, the parser divides it into tokenzs. The output of the parser phase is the list of the words in the webpage. The second phase is the stopwords remover. Stopwords are common words that carry less important meaning than keywords. Usually search engineers remove Stopwords from a keyword phrase to return the most relevant result. Examples of the stop words are: the, an, a, are different Stopwords removers can be used as PorterStemAnalyzer [19].

4.5. Adding Time Dimension

The inclusion of the additional dimension, time spent on page, appears to improve the recommendations made. So a system is proposed which aims to improve the web recommendations by taking the time spent on web pages into consideration in addition to using web usage and web semantics mining. Users often spend considerable time browsing the web pages for getting the right information. If the users' intention and interest for browsing a web page can be identified, it will be easier to make available that area of information with higher priority.

A user session can be represented as a sequence of pairs, each pair containing the web-page accessed and the normalized time spent on that page. For instance a user session S can be represented as:

$$S = (< WP1, T1 >, < WP2, T2 >, < WP3, T3 >, \ldots\ldots, < WPn, Tn >) \qquad (4.1)$$

where, WP1, WP2, WP3, WP4,… are the pages in the session, and T1, T2,T3, T4,… are the respective normalized times.

A user will spend more time on a web page generally when he/she finds the content of the web page interesting. Thus, the time spent on a web page while browsing is an important metric to judge the user's interest in that page and thus the importance of the page.

To better understand this, consider web pages A, B, C, D, and E which are not semantically similar to each other. Consider a user who has navigated pages A,B,C,D and spent less than a minute, 3 minutes, 5 minutes and 10 minutes on these pages respectively and another user who has visited pages A, B, C, E and spent less than a minute, 10 minutes, 3 minutes and 6 minutes on these pages respectively. Now suppose we have a new user who visits pages A, B, C for duration less than a minute, 10 minutes and 3 minutes. Which page recommendation would be more relevant page D or page E? Recommending page E would make more sense and be closer to the current user's interest.

The web personalization approaches that rely solely on web usage data would recommend both, pages D and E. Those approaches that rely on web usage data and web semantics would recommend pages D, E and other pages semantically similar to both these pages. But now that we have considered the time spent on the web pages, we know that the new user's interest will be more similar to those users who have spent approximately the same time on web pages browsed prior to web page E. Thus, now when we combine web usage data and web semantics along with time spent on the web pages we get the new recommendation as web page E and all other pages that are semantically similar only to web page E.

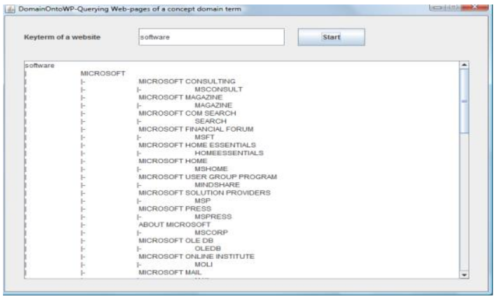
Also, instead of considering the time spent as it is, it makes more sense to have time slots and associate the web pages with the corresponding time slot. For example, one could have three time slots where time spent on a web page is: less than 2 minutes, between 2 minutes to 5 minutes and greater than 5 minutes. The range for time slots can be decided based on the average time spent on web pages of the web site being considered. The advantage of using time slots over individual time duration is that it is more flexible and does not differentiate between pages that have similar but not the same amount of time spent on them by different users.

V. Experiments -Queries and Recommendation Strategies

In order to evaluate the effectiveness of the proposed models of knowledge representation and the recommendation strategies along with the queries, we implement these models, algorithms and strategies to test their performance of Web-page recommendation using a public dataset. The Microsoft (MS) website (www.microsoft.com) was used to run the experimental cases. The dataset was downloaded from http://kdd.ics.uci.edu/databases/msweb/msweb.html.

5.1. Queries based on Domain Ontology

- To query domain terms (topic) of a given Web-page
- To query Web-pages of a given domain term



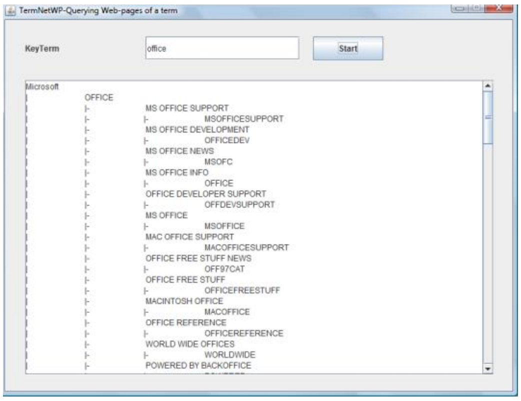
Keyterm of a website: software

```
software
            MICROSOFT
            |-          MICROSOFT CONSULTING
            |-          |-          MSCONSULT
            |-          MICROSOFT MAGAZINE
            |-          |-          MAGAZINE
            |-          MICROSOFT COM SEARCH
            |-          |-          SEARCH
            |-          MICROSOFT FINANCIAL FORUM
            |-          |-          MSFT
            |-          MICROSOFT HOME ESSENTIALS
            |-          |-          HOMEESSENTIALS
            |-          MICROSOFT HOME
            |-          |-          MSHOME
            |-          MICROSOFT USER GROUP PROGRAM
            |-          |-          MINDSHARE
            |-          MICROSOFT SOLUTION PROVIDERS
            |-          |-          MSP
            |-          MICROSOFT PRESS
            |-          |-          MSPRESS
            |-          ABOUT MICROSOFT
            |-          |-          MSCORP
            |-          MICROSOFT OLE DB
            |-          |-          OLEDB
            |-          MICROSOFT ONLINE INSTITUTE
            |-          |-          MOLI
            |-          MICROSOFT MAIL
```

Figure.4. Results of queries based on domain ontology

5.2. Queries based on Semantic Network

- To query domain terms (topic) of a given Web-page
- To query Web-pages mapped to a given domain term



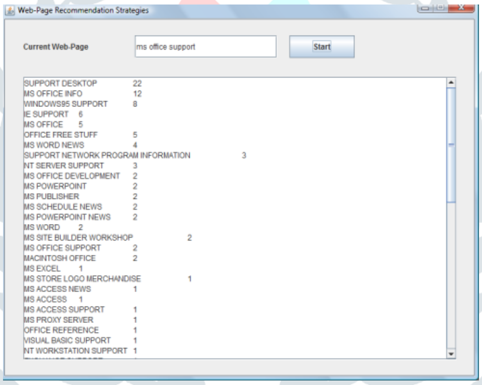
KeyTerm office

```
Microsoft
|            OFFICE
|            |-            MS OFFICE SUPPORT
|            |-            |-            MSOFFICESUPPORT
|            |-            MS OFFICE DEVELOPMENT
|            |-            |-            OFFICEDEV
|            |-            MS OFFICE NEWS
|            |-            |-            MSOFC
|            |-            MS OFFICE INFO
|            |-            |-            OFFICE
|            |-            OFFICE DEVELOPER SUPPORT
|            |-            |-            OFFDEVSUPPORT
|            |-            MS OFFICE
|            |-            |-            MSOFFICE
|            |-            MAC OFFICE SUPPORT
|            |-            |-            MACOFFICESUPPORT
|            |-            OFFICE FREE STUFF NEWS
|            |-            |-            OFF97CAT
|            |-            OFFICE FREE STUFF
|            |-            |-            OFFICEFREESTUFF
|            |-            MACINTOSH OFFICE
|            |-            |-            MACOFFICE
|            |-            OFFICE REFERENCE
|            |-            |-            OFFICEREFERENCE
|            |-            WORLD WIDE OFFICES
|            |-            |-            WORLDWIDE
|            |-            POWERED BY BACKOFFICE
```

Figure.3. Results of queries based on semantic network

5.3. Recommendations based on Conceptual Prediction Model

- Given a current web-page and time-slot, a set of web-pages are recommended based on the semantic enhanced model, which integrates domain knowledge with the web-usage knowledge. This is done by calculating the first order transition probability between the given term and the next set of terms from a frequently viewed term patterns.
- The keywords of the web-pages of the Microsoft (MS) website are extracted and stored in a database along with its frequency of occurrence. The set of web-pages from CPM are compared for matching keywords and the information is added to the recommendations made.
- The given time-slot of the current user is compared with the time-slot of previously viewed user patterns and a flag is set, finally those set of pages that match with the given time-slot are highlighted in the recommendations made.



Current Web-Page ms office support

```
SUPPORT DESKTOP          22
MS OFFICE INFO           12
WINDOWS95 SUPPORT        8
IE SUPPORT    6
MS OFFICE     5
OFFICE FREE STUFF        5
MS WORD NEWS             4
SUPPORT NETWORK PROGRAM INFORMATION          3
NT SERVER SUPPORT        3
MS OFFICE DEVELOPMENT    2
MS POWERPOINT            2
MS PUBLISHER             2
MS SCHEDULE NEWS         2
MS POWERPOINT NEWS       2
MS WORD       2
MS SITE BUILDER WORKSHOP          2
MS OFFICE SUPPORT        2
MACINTOSH OFFICE         2
MS EXCEL      1
MS STORE LOGO MERCHANDISE         1
MS ACCESS NEWS           1
MS ACCESS     1
MS ACCESS SUPPORT        1
MS PROXY SERVER          1
OFFICE REFERENCE         1
VISUAL BASIC SUPPORT     1
NT WORKSTATION SUPPORT   1
```

Figure.3. Results of queries based on conceptual prediction model

VI. Conclusion And Future Work

This paper has presented a new method to offer better Web-page recommendations through semantic enhancement by three new knowledge representation models supported by Web-Semantics and the time spent by the users on a Web-page. Two new models have been proposed for representation of domain knowledge of a website. One is an ontology-based model which can be semi-automatically constructed, namely and the other is a semantic network of Web-pages, which can be automatically constructed. A conceptual prediction model is also proposed to integrate the Web usage and domain knowledge to form a weighted semantic network of frequently viewed terms. Term frequency keyword extraction method is used to mine the contents of Web-pages, adding this to the recommendation strategies would make the recommendation more accurate. The time spent by previous users are compared with the current user's time and the recommendations are further refined.

A number of Web-page recommendation strategies have been proposed to predict next Web-page requests of users through querying the knowledge bases. The experimental results are promising and are indicative of the usefulness of the proposed models.

For the future work, more attributes can be taken into consideration that increases the knowledge about a user. Attributes like user's gender, location, age, etc. tell us more about the user and considering the number of times that the web page was browsed help in increasing the knowledge about a web page. Taking these attributes for generating web page recommendations will make it more personalized and useful for the users.

References

[1] Thi Thanh Sang Nguyen, Hai Yan Lu, and Jie Lu.2014."Web-Page Recommendation Based on Web Usage and Domain Knowledge" ieee transactions on knowledge and data engineering, vol. 26, no. 10.

[2] B. Liu, B. Mobasher, and O. Nasraoui. 2011."Web usage mining," in Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, B. Liu, Ed. Berlin, Germany: Springer-Verlag, pp. 527–603.

[3] B. Mobasher, 2007 , "Data mining for web personalization," in The Adaptive Web, vol. 4321, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Berlin, Germany: Springer-Verlag, pp. 90–135.

[4] G. Stumme, A. Hotho, and B. Berendt, 2004, "Usage mining for and on the Semantic Web," in Data Mining: Next Generation Challenges and Future Directions. Menlo Park, CA, USA: AAAI/MIT Press, pp. 461–480.

[5] H. Dai and B. Mobasher, 2005, "Integrating semantic knowledge with web usage mining for personalization," in Web Mining: Applications and Techniques, A. Scime, Ed. Hershey, PA, USA: IGI Global, pp. 205–232.

[6] S. A. Rios and J. D. Velasquez, "Semantic Web usage mining by a concept-based approach for off-line web site enhancements," in Proc. WI-IAT'08, Sydney, NSW, Australia, pp. 234–241.

[7] S. Salin and P. Senkul, 2009, "Using semantic information for web usage mining based recommendation", in Proc. 24th ISCIS, Guzelyurt, Turkey, pp. 236–241.

[8] Ahmad, A.M., Hijazi, M.H.A., & Abdullah, A.H, 2004: Using Normalize Time Spent on a Web Page for Web Personalization [Electronic version]. IEEE Region 10 Conference Volume B, 21-24 Page(s):270 - 273 Vol. 2

[9] Gündüz, S., and Ozsu, M. T., 2003, A Web Page Prediction Model Based on Click-Stream Tree Representation of User Behavior. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , 535-540.

[10] M., Eirinaki, C., Lampos, S., Paulakis, M., Vazirgiannis, 2004: Web Personalization Integrating Content Semantics and Navigational Patterns [Electronic version].WIDM'04, November 12-13, Washington, DC, USA.

[11] M., Eirinaki, G., Tsatsaronis, D., Mavroeidis, M., Vazirgiannis, 2005: Introducing Semantics in Web Personalization: The Role of Ontologies [Electronic version]. Semantics and Web Mining, LNAI 4289, 2006

[12] M., Eirinaki, I., Varlamis, M., Vazirgiannis, 2003: SEWeP: Using Site Semantics and Taxonomy to Enhance the Web Personalization Process [Electronic version]. SIGKDD '03, August 24-27, Washington, DC, USA.

[13] C. I. Ezeife and Y. Lu, 2005, "Mining Web log sequential patterns with position coded pre-order linked WAP-tree," Data Min.Knowl.Disc., vol. 10, no. 1, pp. 5–38.

[14] L. Wei and S. Lei, 2009, "Integrated recommender systems based on ontology and usage mining," in Active Media Technology, vol. 5820,J. Liu, J. Wu, Y. Yao, and T. Nishida, Eds. Berlin, Germany: Springer-Verlag, pp. 114–125.

[15] A. Loizou and S. Dasmahapatra, 2006, "Recommender systems for the semantic Web," in Proc. ECAI, Trento, Italy.

[16] M. Eirinaki, D. Mavroeidis, G. Tsatsaronis, and M. Vazirgiannis, 2006, "Introducing semantics in Web personalization: The role of ontologies," in Proc. EWMF, Porto, Portugal, pp. 147–162.

[17] Jose M. Domenech1 and Javier Lorenzo,2007"A Tool for Web Usage Mining", 8th International Conference on Intelligent Data Engineering and Automated Learning.

[18] Robert.Cooley,BamshedMobasher, and Jaideep Srinivastava,1997, " Web mining:Information and Pattern Discovery on the World Wide Web", In International conference on Tools with Artificial Intelligence, pages 558-567, Newport Beach, IEEE.

[19] L. Mostafa, M. Farouk, and M. Fakhry,2009, "An Automated Approach for Webpage Classification," ICCTA09 Proceedings of 19th International conference on computer theory and applications ,Alexandria,Egypt.