Web-Page Recommendation In Information Retrieval Using Domain Knowledge And Web Usage Mining

NazneenTarannumS.H.Rizvi, Prof. R.R. Keole

Abstract—**It has become much more difficult to access relevantinformation from the Web as it has demonstrated as a wealthy, remarkable and marvelous data source of information.The tremendous growth in volume of web usage data results in the boost of web mining research with focus on discovering potentially useful knowledge from web usage data. Extracting the exact information from a large volume repository of unstructured or semi structured web is a big challenge. Objectives of web data mining is taken in searching relevant and reliable and meaningful knowledge from web learn about the particular user and web synthesis also.This paper proposed a desktop search utility which uses web usage mining process for finding term patterns in web query data which can be used for predicting the possible next pages in browsing sessions. This process consists of four main stages: Extracting web pages from google, creating term patterns, probability calculation of term patterns in name of web pages and recommending new list of web-pages based on term frequencies. In this paper, a framework is presented for recommending better web-pages based on the queries fried by the users over google search engine. The proposed system is described, and its performance is also evaluated**.

Index Terms—**Term pattern, Term frequency, search utility, Web-page recommendation, Domain knowledge, Web usage mining.**

I. INTRODUCTION

The continuous growth in the size of the World Wide Web has resulted in intricate Web sites, demanding enhanced user skills and more sophisticated tools to help the Web user to find the desired information. Due to the enormous growth of usage of WWW by the users, transactions are growing very quickly. Many experts forecast that the subsequent huge growth is forwarded in web information technology by adding semantics to web data, and will almost certainly consist of the semantic web. The recommendation accuracy of usage based techniques can be improved by integrating Web site content and site structure in the mining process. The goal of the intelligent recommendation system is to determine which web pages are more likely to be accessed by the user in the future

Manuscript received May, 2015.

NazneenTarannum S.H. Rizvi*,Computer Science and Information Tecnology,S.G.B.AUniversity, Amravati, India, 9403307299*

Prof.Ranjit R. Keole*,Computer Science and Information Tecnology,S.G.B.AUniversity, Amravati, India, 9823852893*

Web usage mining is one of the frequent usage areas of web mining. The awareness of Web mining lies in analyzing user's behaviour on the web after exploring access logs and its popularity is increasing at a faster face especially in E-services areas. The applications in these web semantic search areas added its approval and made it as an inevitable part in computer and information sciences. Details like user log files demand for resources and maintain web servers, which is the core mining area of web usage. The semantic analysis gives the user browsing patterns utilized for target advertisement, development of web design, fulfilment of users and making market analysis.

Also, Web semantic search is a key technology of the web database, since it is the major process through which the access content in the web data can be performed. Current web search technologies are fundamentally based on grouping of textual keyword search using ranking via the link structure of the web. Present web semantic search does not permit semantic processing of web search queries, which analyzed based on both web search queries and web pages. Semantic Knowledge-Based as presented shows how to abstract away from the raw real-world information step by step by means of semantic technologies

Web-page recommender systems are one kind of recommender systems, which can automatically recommend Web-pages that are most interesting to a particular user based on the user's current Web navigation behaviour. Since a website is usually designed to show the index pages on the home page, the index pages take the role of guiding users to the content pages on the website through Web-page links, whereas with the index pages, a user usually has to navigate a number of Web-pages to reach the content page they are interested in. If the index pages of a website are not well designed, which is often the case, Web users will struggle to find useful pages and are very likely to leave the site. For a commercial website, this means losing potential customers. Therefore, Web-page recommender systems have become increasingly valuable for helping Web users to find the most interesting and useful Web-pages on specific websites. Good Web-page recommendations can improve website usage and Web user satisfaction.

Different from the majority of the existing web recommendation techniques, we propose an intelligent web recommendation system that uses a term pattern frequency mining technique. which is very suitable for predicting the next web pages. Different evaluation measures including time constraint, precision, satisfaction and applicability are proposed to measure the performance of the recommendation system.

The rest of this paper is organized as follows. Section II presents the related work and literature review.

ISSN: 2278 – 7798

Section III presents the proposed system architecture of WRS-SU. Section IV discusses the performance evaluation and experimental results. Finally, Section V concludes the paper with its future work.

II. RELATED WORK AND LITERATURE REVIEW

The related work and literature review covers the background, latest development of and related techniques for recommender systems using usage mining.

In 2005, GediminasAdomavicius describes various limitations of current recommendation methods and discusses possible extensions that can improve recommendation capabilities and make recommender systems applicable to an even broader range of applications.

In 2006, A. Loizou presents a semantics-based approach to Recommender Systems (RS), to exploit available contextual information about both the items to be recommended and the recommendation process, in an attempt to overcome some of the shortcomings of traditional RS implementations.

In 2006, M. Eirinaki proposed a semantic web personalization system, focusing on word sense disambiguation techniques which can be applied in order to semantically annotate the web site's content.

In 2007, B. Mobasherpresents an overview of Web personalization process viewed as an application of data mining requiring support for all the phases of a typical data mining cycle. These phases include data collection and preprocessing, pattern discovery and evaluation, and finally applying the discovered knowledge in real-time to mediate between the user and the Web.

In 2007,Sinéad Boyce and Claus Pahl present a method for domain experts rather than ontology engineers to develop ontologies for use in the delivery of courseware content.

In 2008, S. A. Rios proposed a concept-based approach to add semantics into the mining process. The solution proposed, was applied to a real web site to produce offline enhancements of contents and structure.

In 2008, Dale Dzemydiene& Lina Tankeleviciene presents the scope and purpose of ontology for "E-learning technologies" course, argue about manual development of domain ontology, and provide a brief introduction on formalisms (classes, relations, formal axioms, and instances) for knowledge representation on the ontological level.

In 2009, S. Salin presents a framework for integrating semantic information with Web usage mining is presented. The frequent navigational patterns are extracted in the form of ontology instances instead of Web page addresses and the result is used for generating Web page recommendations to the visitor.

In 2009, C. I. Ezeife& Y. Liu presents incremental mining of web sequential patterns to generate current frequent patterns for the updated database (consisting of both old and incremental data) using mostly only the incremental (or newly added) data and previously mined frequent patterns.

In 2010, Xiaogang Wang proposed an efficient sequential pattern mining algorithm used to identify frequent sequential web access patterns. The access patterns are then stored in a compact tree structure, called Pattern-tree, which is then used for matching and generating web links for recommendations

In 2010,Sang T.T. Nguyen presents a new web usage mining process for finding sequential patterns in web usage data which can be used for predicting the possible next move in browsing sessions for web personalization.

In 2011, S. Grimm discusses the development of a new information representation system embodied in ontology and the Semantic Web. The new system differs from other representation systems in that it is based on a more sophisticated semantic representation of information, aims to go well beyond the document level, and designed to be understood and processed by machine.

In 2011, C.Rameshproposed a novel framework integrating semantic information in the Web usage mining process. Sequential Pattern Mining technique is applied over the semantic space to discover the frequent sequential patterns.

In 2012, V. SithaRamulupresents an overview of the semantic web mining- Integration of domain knowledge in to web mining to form semantic web mining, the concepts of semantic web mining.

In 2012, ThiThanh Sang Nguyen presents a new framework for a semantic-enhanced Web-page recommender (WPR) system, and a suite of enabling techniques which include semantic network models of domain knowledge and Web usage knowledge, querying techniques, and Web-page recommendation strategies. The paper enables the system to automatically discover and construct the domain and Web usage knowledge bases, and to generate effective Webpage recommendations

In 2014, ThiThanh Sang Nguyen proposed a the conceptual prediction model to automatically generate a semantic network of the semantic Web usage knowledge, which is the integration of domain knowledge and Web usage knowledge.

In 2014, Suresh Shirgave propose semantically enriched Web Usage Mining method for Personalization (SWUMP), which is a combination of the fields of Web Usage Mining and Semantic Web. In this method, the undirected graph is derived from usage data with rich semantic information extracted from the Web pages and the Web site structure

III. PROPOSED SYSTEM ARCHITECTURE

The proposed system gives a novel method to efficiently provide better Web-page recommendation through semantic enhancement by integrating the domain and Web usage knowledge of a website. In this system, we are implementing the system on the basis of already available data of Microsoft, to provide there analytics. Using the current visited Web-page (referred to as a state) and k previously visited pages (the previous k states), the Web-page(s) that will be visited in the next navigation step can be predicted. We will collect the terms available in metadata of the web-pages from the google and then depending upon the metadata, they will be updated in order i.e. define the term patterns and after this we will calculate the term pattern frequency of the query. By considering above system, semantic knowledge representation model of web usage of website for webpage recommendation will be considered. With the help of this system, the pages can be predicted.

ISSN: 2278 – 7798

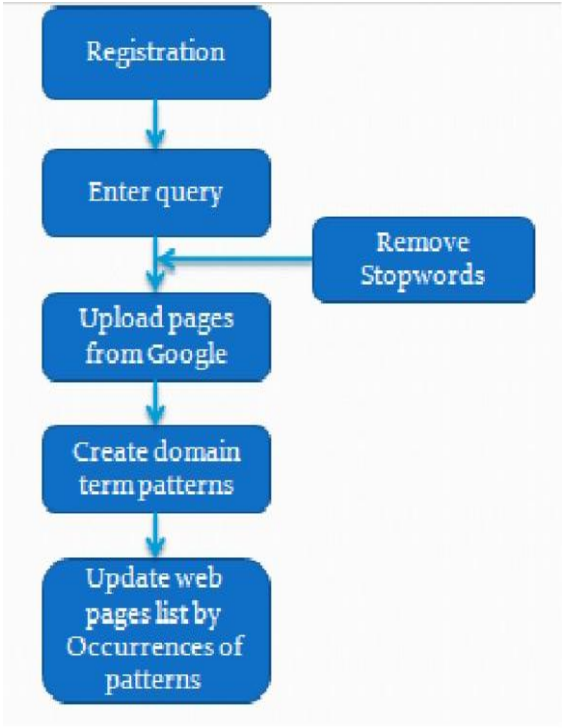

Fig: Flow chart of WRS-SU

THE MINING PROCESS FOR PROPOSED SYSTEM

The flow of work in WRS-SU can be illustrated as follows:

1. Firstly User login to the WRS-SU by providing its username and password. If in case user has no login account he has to signs up in WRS-SU to create a username and password.
2. After successful login, User enters a query in WRS-SU for accessing the desired web pages..
3. The query received by WRS-SU is send to Google API for extracting the relevant results or snippets containing desired information from the Google. At the same time every query is scanned to analyze whether the same query was fired by the user previously. If so, the already recommendated web-pages are provided to users.
4. The web pages extracted from the Google are stored in a folder. From these results, the URL's are extracted by discarding the unwanted links.
5. The stop-words are removed from the query entered by the users and we get the keywords or domain terms.
6. Term patterns are created from these keywords.
7. Term frequencies or probabilities of term occurrences are calculated on each web-pages of the results from the google.
8. Based on the term frequency calculated, the web-pages list are updated to bring the lower priority pages at a higher priority.
9. At the Recommendation system, the CPM Algorithm will be executed and the resultant web pages will be displayed by their priority based on frequency of their domain terms
10. Admin keeps an eye on every query provided by the user and the information accessed from the response generated by WRS-SU.

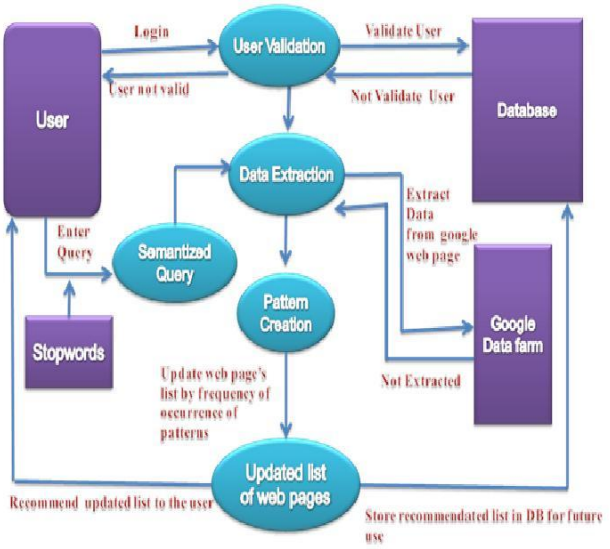

Fig: Data Flow Diagram

IV. PERFORMANCE EVALUTION

The performance of Web-page recommendation strategies is measured in terms of time required for recommending the web-pages list between the google search engine and the proposed search utility(WRS-SU)

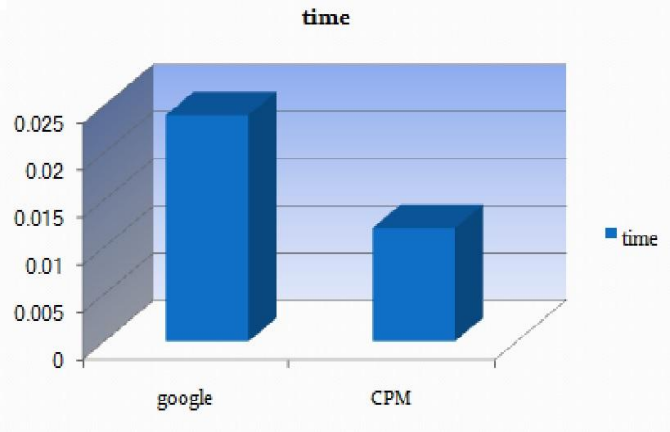

Fig: Time comparison between google and WRS-SU

*The two major performance metrics are :Precision*and *Satisfaction*. In order to calculate these two metrics, we introduce two definitions: Web-page recommendation rules and Support.

Definition (Web-page recommendation rules). *Let* $S = S_1S_2 \ldots S_kS_{k+1} \ldots sn$ $(n \geq 2)$ *be a WAS. For each prefixsequence* $S_{prefix} = S_1S_2 \ldots S_k$ $(k \leq n - 1)$*, a Web-page recommendationrule is defined as a set of recommended Web-pages generated by a Web-page recommendation strategy, denotedas RR* = {*r*1, *r*2, . . . ,*rM*}*, where ri (i*= [1 . . .*M*]*) is a recommended Web-page.*

A Web-page recommendation rule is deemed as *correct*, and/or *satisfied*, or *empty* based on the following conditions:

1) If *sk*+1 €*RR,RR*is *correct*.
2) If ∃ *si*€*RR*$(k + 1 \leq i \leq n)$, *RR* is *satisfied*.
3) If $M = 0$, *RR* is *empty*.

Definition (Support). *Given a set _ of WAS and a set P* = {*P*1, *P*2 . . . *Pn*} *of frequent (contiguous) Web access sequences over* Δ*, the support of each Pi* €*P is defined as:*

ISSN: 2278 – 7798

$\sigma(Pi) = |\{S \in \Delta : Pi \subseteq S\}|$
$|\Delta|$, *where S is a WAS.*

Support is used to remove infrequent Web-pages and discover FWAP from WAS. This is accomplished by setting a Minimum Support (*MinSup*) and using it as a threshold to checkWAS. TheWeb access sequences whose Support values are greater than, or equal to *MinSup*are considered as FWAP. The smaller MinSup is set, the more FWAP are discovered.

Performance evaluation of the proposed approach and the search engine is done based on Precision measure.

Precision is the basic measure used in evaluating search strategies. There is a set of records in the database which is relevant to the search. Records are assumed to be either relevant or irrelevant. The actual retrieval set may not perfectly match the set o relevant records.

It is the ratio of number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as percentage.

Precision measure is calculated based on the following formula.

$$\text{Precision} = \frac{tp}{tp + fp}$$

Where,
tp – True Positive (Correct result)
fp – False Positive (Unexpected Result)

Different Methods	tp	fp	Precision
Search Engine Recommendation	15	10	0.6
Domain & Usage Based Recommendation	4	0	1

Table 4.1: Precision Measure

From the table 4.1, it is understood that precision of the search-engine is 0.6, and precision of combine approach of Domain and usage mining is 1 out of 1. The results of the performance measure are plotted in Figure 4.1.

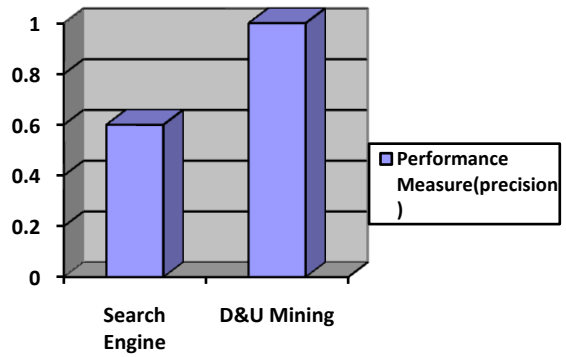

Figure 4.1 Performance Measure

V. CONCLUSION AND FUTURE RESEARCH

In this paper, we have proposed an intelligent web-page recommendation system known as WRS-SU based on Term Pattern frequency. In the proposed system, the sequential pattern mining algorithm CPM is used to mine frequent sequential web term patterns. The mined patterns are stored in the software folder, which is then used for matching and generating web links for online recommendations. The proposed system has achieved good performance with high satisfaction and applicability and the time required for predicting the next web-pages are better than the google search engine.

Future work will focus on further experiments with different combinations of the system's functionalities, further contextualization possibilities from the Semantic Web Mining area, and an evaluation of the proposed approach with respect to learning support and to open-corpus learning Future research can be considered as:

(i)Multi-site: The proposed framework and enabling techniques can be extended to make Web-page recommendations for multiple websites in the same domain.

(ii)Web usage data: To offer more effective Web-page recommendations, it will be highly desirable to develop advanced tools to identify and collect more appropriate Web usage data than Web logs, such as clickstream data.

(iii)Web usage knowledge base update: To ensure that the discovered Web usage knowledge is up-to date, new methods need to be developed to dynamically update the knowledge bases .

(iv)Domain knowledge discovery and representation: Advanced topic models from the area of information retrieval can be used to extract thedomain terms from the Web-pages on the website

ACKNOWLEDGMENT

My thanks to the Guide, Prof. R.R.Keole and Principal Dr.A.B.Marathe, who provided me constructive and positive feedback during the preparation of this paper.

REFERENCES

[1] Wang, X., Bai, Y. & Li, Y. 2010, 'An Information Retrieval Method Based on Sequential Access Patterns', 2010 Asia-Pacific Conference on Wearable Computing Systems (APWCS), pp.247-250

[2] 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.

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

[4] A. Bose, K. Beemanapalli, J. Srivastava, and S. Sahar, "Incorporating concept hierarchies into usage mining based recommendations," in *Proc. 8th WebKDD*, Philadelphia, PA, USA, 2006, pp. 110–126.

[5] N. R. Mabroukeh and C. I. Ezeife, "Semantic-rich Markov models for Web prefetching," in *Proc. ICDMW*, Miami, FL, USA, 2009, pp. 465–470.

[6] M. O'Mahony, N. Hurley, N. Kushmerick, and G. Silvestre, "Collaborative recommendation: A robustness analysis," *ACM Trans. Internet Technol.*, vol. 4, no. 4, pp. 344–377, Nov. 2004.

[7] G. Stumme, A. Hotho, and B. Berendt, "Semantic Web mining: State of the art and future directions," *J. Web Semant.*, vol. 4, no. 2, pp. 124–143, Jun. 2006.

[8] B. Zhou, S. C. Hui, and A. C. M. Fong, "CS-Mine: An efficient WAP-tree mining for Web access patterns," in *Proc. Advanced Web Technologies and Applications*. vol. 3007. Berlin, Germany, 2004, pp. 523–532.

[9] J. Borges and M. Levene, "Generating dynamic higher-order Markov models in Web usage mining," in *Proc. PKDD*, Porto, Portugal, 2005, pp. 34–45. C. I. Ezeife and Y. Lu, "Mining Web log sequential patterns with position coded pre-order linked WAP-tree," *Data Min. Knowl. Disc.*, vol. 10, no. 1, pp. 5–38, 2005.

[10] B. Zhou, S. C. Hui, and A. C. M. Fong, "Efficient sequential access pattern mining for web recommendations," *Int. J. Knowl.-Based Intell. Eng. Syst.*, vol. 10, no. 2, pp. 155–168, Mar. 2006.

[11] C. Ezeife and Y. Liu, "Fast incremental mining of Web sequential patterns with PLWAP tree," *Data Min. Knowl. Disc.*, vol. 19, no. 3, pp.

ISSN: 2278 – 7798

376–416, 2009.
[12] T. T. S. Nguyen, H. Lu, T. P. Tran, and J. Lu, "Investigation of sequential pattern mining techniques for Web recommendation," *Int. J. Inform. Decis. Sci*., vol. 4, no. 4, pp. 293–312, 2012.
[13]] S. T. T. Nguyen, "Efficient Web usage mining process for sequential patterns," in *Proc. IIWAS*, Kuala Lumpur, Malaysia, 2009, pp. 465–469.
[14] L. Wei and S. Lei, "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, 2009, pp. 114–125. A. Loizou and S. Dasmahapatra, "Recommender systems for the semantic Web," in *Proc. ECAI*, Trento, Italy, 2006.
[15] B. Liu, "Information retrieval and Web search," in *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, B. Liu, Ed. Berlin, Germany: Springer, 2011, pp. 183–236.
[16] H. Dai and B. Mobasher, "Integrating semantic knowledge with web usage mining for personalization," in *Web Mining: Applications and Techniques*, A. Scime, Ed. Hershey, PA, USA: IGI Global, 2005, pp. 205–232

NazneenTarannum S.H. Rizvi, received the B.E.degree in Information Technology from H.V.P.M's College Of Engineering And Technology, Amravati in 2012. She is currently persuing Master's Degree in Computer Science and Information Technology from H.V.P.M's College of Engineering And Technology, Amravati.

Prof. RanjitR.Keole, received the B.E.and M.E degree in Computer Science from Prof. Ram Megha Institute of Technology, Badnera in 1992 and 2008, respectively. His field of specialisation is web Mining. He is currently working as Associate Professor at H.V.P.M's college of Engineering and Technology,Amravati.

ISSN: 2278 – 7798