ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 2, March 2016

A Survey of Web Page Recommendation Based on Web usage

[1]P. vinothini, Asst.Prof /CSE,[2]M.Lavanya, Asst.Prof /CSE,
[3]M.SasikalaAsst.Prof /CSE, [4]P.Dhivya, Asst.Prof /CSE
Bharathiyar Institution of engineering for women,Salem,Tamilnadu
Vinvinothini2@gmail.com

***Abstract*—. The Recommendation of the web page is an unique role in intelligent Web systems. The challenging issues in the web page recommendations are useful for gaining knowledge discovery from web usage data to satisfy knowledge representation. The existing work presents semantic enhancement scheme by integrating domain and web usage for the efficient webpage recommendation. The ontology model was used to represent the domain knowledge. The conceptual prediction model integrates the domain knowledge and web usage knowledge. The drawbacks of existing work include the issues in extracting the structured records from semi-structured web pages and the web page recommendations made were bit ambiguous. The proposed work presents the key information extraction technique for the website to suggest the web page recommendation. The annotations of web pages are derived from the corresponding sites in the target domain. It allows the automatic extraction from potentially other sites within the same domain. The clarity of web page recommendations is also increased. The recommendation results is compared with existing Web Usage Mining (WUM) method. The experimental results showed the proposed method to produces significantly higher performance than the WUM method.**

***Keywords*-. Web usage mining, domain ontology, semantic network, knowledge representation**

I. INTRODUCTION

Web Page recommendation is developing popular websites, and it is links to related or similar stories, books, or most viewed pages at websites. When a user browses a website, a visited Web-pages sequence in a session automatically (the period from starting, to existing the browser by the user) can be generated. The objective of a Web-page recommender system is to effectively predict the Web-page and that pages will be visited from a given Web-page of a website.

A web access sequence (WAS) is a Web usage data. This approach known from the training datasets to build the links between Web-pages. By using these approaches, the newly visited Web-page (state) and *k* already visited pages (the old *k* states), the Web-page(s) that will be visited in the next step of navigation and it can be predicted. The performance of these approaches depends on the size of training datasets.

Domain ontology is the representation of semantics Web-pages of a website. It shows the integrating domain knowledge with Web usage knowledge and it enhances the performance of recommender systems using ontology-based Web mining techniques. Integrating semantic information with Web usage mining achieves higher performance than classic Web usage mining algorithms.

It extracts data from detail pages of web sites. These are the pages which correspond to a single data entity, and which different attributes of new entity in a human-understable one. The methods could be easily adapted to the case where various data records exist on one page by use of existing data record detection algorithms.

Extracting structured records from semi-structured web pages is an important problem because all information on the internet is presented in basic form. Moreover, the structured records are particularly valuable to downstream learning or querying systems its attribute/value structure and conformance to a fixed domain schema. The problem to be solved by without requiring human supervision and the specific target web site. Most usual work on this is to required human supervision by a person to either annotate example web pages from each target website or map the new data fields for each target site to columns in a domain based schema. It does not require supervision for each target site and it was not robust to typical variations on semi-structured web sites, for example data fields do not have a label and it do not match some unique pattern. This method based on human supervision , the web pages from sites in the target domain, allowing minimally supervised extraction in hundreds or thousands of other sites in the same domain.

II. LITERATURE REVIEW

Semantic Web Mining aims at combining the two fast-developing research areas Semantic Web and Web Mining. Web Mining aims at discovering insights about the meaning of

Web resources and their usage[1]. A popular approach for modeling sites and their usage is related to OLAP techniques: a modeling of the pages in terms of possibly multiple) concept hierarchies, and an investigation of patterns at different levels of abstraction, i.e. a knowledge Discovery cycle which iterates over various “roll-ups” and “drill-downs”. Concept hierarchies conceptualize a domain in terms of taxonomies such as product catalogs, topical thesauri, etc. Semantic Web Usage Mining can improve the results of ‘classical’ usage mining by exploiting the new semantic structures in the Web. Web mining methods should increasingly treat content, structure, and usage in an integrated fashion in iterated cycles of extracting and utilizing semantics, to be able to understand and (re)shape the Web.

Incorporating hierarchical of new model that effectively combines usage information with Information from the conceptual structure of the website to generate our recommendations. Such a structure is termed the concept hierarchy of the website. It is important to emphasize here that the author limit ourselves only to the hierarchy of pages of an individual website and do not deal with a hierarchy of the entire web. Conceptual and structural characteristics of a website can play an important role in the quality of recommendations provided by a recommendation system. Resources like Google Directory, Yahoo! Directory and web-content management systems attempt to organize content conceptually. Most recommendation models are limited in their ability to use this domain knowledge. Recommendation models based only on usage information are inherently incomplete because they neglect domain knowledge. Better predictions can be made by modeling and incorporating context dependent information[2]: concept hierarchy, link structure and semantic classification allow us to do so.

semantic information [3] is to prune states in Selective Markov models SMM, semantic information can lead to context-aware higher order Markov models with about 16% less space complexity. The integration of semantic information directly in the transition probability matrix of lower order Markov models, was presented as a solution to this tradeoff problem resulting in semantic-rich lower order Markov models. This integration also solves the problem of contradicting prediction. The integration of semantic information, drawn from underlying domain ontology, into probabilistic low-order Markov models is discussed. Semantic information is infused into the Markov transition probability matrix to convert it to a matrix of weights for better-informed prediction, and to overcome the problem of contradicting prediction. This paper takes this idea a step further by proposing to use maximum semantic distance as a measure for pruning higher-order Markov models.

Collaborative recommendation [4]has emerged as an effective technique for personalized information access. However, there has been relatively little theoretical analysis of the conditions under which the technique is effective. To explore this issue, the author analyzes the robustness of collaborative recommendation: the ability to make recommendations despite (possibly intentional) noisy product Ratings. There are two aspects to robustness: recommendation accuracy and stability. The accuracy of various collaborative recommendation algorithms has been empirically validated for many domains, and the technology has been successfully deployed in many commercial settings. Collaborative recommendation has been demonstrated empirically, and has been widely adopted commercially. Unfortunately, the author does not yet have a general predictive theory for when and why collaborative recommendation is effective. The author has contributed to such a theory by analyzing robustness, a recommender system’s resilience to potentially malicious perturbations in the customer/product rating matrix.

The Semantic Web[5]is the second-generation WWW, enriched by machine process able information which supports the user in his tasks. Given the enormous size even of today’s Web, it is impossible to manually enrich all of these resources. Therefore, automated schemes for learning the relevant information are increasingly being used. Web Mining aims at discovering insights about the meaning of Web resources and their usage. Markup and mining approaches that refers to an explicit conceptualization of entities in the respective domain. These relate the syntactic tokens to background knowledge represented in a model with formal semantics. The author discussed how Semantic Web Mining can improve the results of Web Mining by exploiting the new semantic structures in the Web; and how the construction of the Semantic Web can make use of Web Mining techniques. Furthermore, mining the Semantic Web itself is another upcoming application. The author argue that the two areas Web Mining and Semantic Web need each other to fulfill their goals, but that the full potential of this convergence is not yet realized.

Markov models have been widely used for modeling users' web navigation behavior. In previous work the author have presented a dynamic clustering-based Markov model [6] that accurately represents second order transition probabilities given by a collection of navigation sessions. Herein, the author proposes a generalization of the method that takes into account higher-order conditional probabilities. An alternative method of modeling navigation sessions are tree-based models. Schechter use a tree-based data structure that represents the collection of paths inferred from log data to predict the next page accessed, while Dongshan and Junyi proposed a hybrid- order tree-like Markov model to predict web page access. In addition, Chen and Zhang use a Prediction by Partial Match tree that restricts the roots to popular nodes. The resulting dynamic high-order Markov model is such that the probabilities of the out links from a given state reflect the *n*- order conditional probabilities of the paths to the state. Thus, the model is able to capture a variable length history of pages, where different history lengths are needed to accurately model user navigation. In addition, the method makes use of a probability threshold together with a clustering technique that enables us to control the number of additional states induced by the method at the cost of some accuracy.

The new web usage mining process for finding sequential patterns[7] in web usage data which can be used for predicting the possible next move in browsing sessions for web personalization. This process consists of three main stages: preprocessing web access sequences from the web server log, mining preprocessed web log access sequences by a tree-based algorithm, and predicting web access sequences by using a dynamic clustering-based model. Web Usage Mining algorithms can be classified into many categories, such as clustering, classification, association rules, and sequential pattern discovery. There are two major methods of sequential pattern discovery: deterministic techniques (recording the navigational behavior of the user) and stochastic methods. The frequent web navigation patterns will then be modeled by a higher-order Markov model to predict the next page accessed by a user. The new web usage mining process proposed in this paper is the combination of the PLWAP algorithm and the dynamic clustering-based Markov model. It inherits the advantages of the PLWAP-tree and the dynamic clustering-based Markov model and overcomes their drawbacks by omitting uninteresting web pages. The resultant web page link graph not only presents all possible links of the websites, but also predicts the frequently visited links or the frequent web navigations. This makes the new mining process be very useful for web personalization systems using recommender systems. In contrast, a web usage mining process only using the Markov model does not highlight frequent web navigation or user interests.

The semantics-based approach for Recommender Systems (RS), to exploit available contextual information about both the items to be recommended and the recommendation process[8], in an attempt to overcome some of the shortcomings of traditional RS implementations. These shortcomings reflect the lack of computational support for humans who are interested in items they or the people who usually share their taste haven't previously come across. In addition, such systems do not allow for shifts of the user's interest over time, since all ratings provided by a user have an equal bearing on the recommendation selection. To overcome such issues, a system should be able to consider the semantics of both the recommendation context and those of the items at hand to constrain the recommendation process. Information specific to the recommendation context for both user clustering and content-based comparisons have been shown to improve overall recommendation performance. The analysis carried out to this point has shown that contextual relationships between artists and arbitrary resources can be successfully used to build feature vectors and to produce clustering reflective of real users' listening preferences. The analysis carried out to this point has shown that contextual relationships between artists and arbitrary resources can be successfully used to build feature vectors and to produce clustering reflective of real users' listening preferences. It is intended that the collaborative filtering information, available from Last.fm will be imported into the system in order to assess the circumstances under which selection and combination of appropriate sub-spaces of the full high dimensional recommendation space is beneficial, with respect to the predictive ability of the system.

Ontology [9] has the potential to play an important role in instructional design and the development of course content. They can be used to represent knowledge about content, supporting instructors in creating content or learners in accessing content in a knowledge-guided way. While ontology's exist for many subject domains, their quality and suitability for the educational context might be unclear. Ontology defines the kinds of things that exist in an application domain. In the computing context, ontology is a framework for representing concepts (things, or ideas about things) and the relationships that exist between those concepts. Ontology's have been used in various educational-technology systems (Sampson et al., 2004). In particular, they can capture the knowledge aspects of educational content (Arroyo et al., 2002). However, ontology's for a particular subject may not exist or it might be unclear if existing ones are suitable. The author has therefore addressed how content ontology's should appear with regard to their structure and quality, and how to develop content ontology's for educational technology. The ontology's can provide an interface to the content. As the author have discussed, these ontology's can guide the instruction design of a course. Learners (or instructors) can browse through the content guided by the dependencies expressed in the concept ontology, thus allowing for the delivery of a course in a way that matches the preferred learning style of the user by varying the sequentialization of content elements. Christo Ananth et al. [16] discussed about a Secure system to Anonymous Blacklisting. The secure system adds a layer of accountability to any publicly known anonymizing network is proposed. Servers can blacklist misbehaving users while maintaining their privacy and this system shows that how these properties can be attained in a way that is practical, efficient, and sensitive to the needs of both users and services. This work will increase the mainstream acceptance of anonymizing networks such as Tor, which has, thus far, been completely blocked by several services because of users who abuse their anonymity. In future the Nymble system can be extended to support Subnet-based blocking. If a user can obtain multiple addresses, then nymble-based and regular IP-address blocking not supported. In such a situation subnet-based blocking is used. Other resources include email addresses, client puzzles and e-cash, can be used, which could provide more privacy. The system can also enhanced by supporting for varying time periods.

III. RESULT AND DISCUSSION

A. Experimental Evaluation.

In this section we describe the experimental methodology and metrics used to compare different prediction algorithms; and present the results of our experiments. Experiments are conducted with the available implementation of Naive Bayes classifier in WEKA using 5-fold cross validation.

It compared the performance of different techniques from traditional webpage recommendation approach. The new models of knowledge representation,queries,and recommendation strategies were done by using the five experimental cases as follows

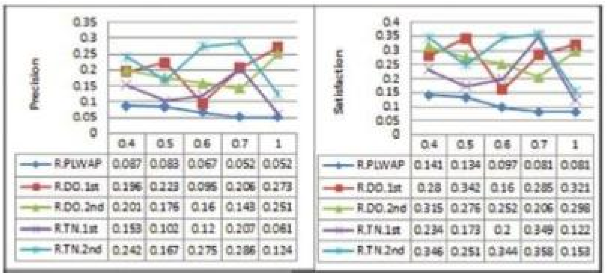
Case1(Recommendation.Preorder Linked-WAP):Set the threshold of the recommendation approach by using of web usage mining algorithm.

Case 2(Recommendation.Domain Ontology 1st):Test the effectiveness by integrating the domain ontology (DomainOntoWP) with TermNavNet using the first-order CPM.

Case 3(Recommendation Domain Ontology 2nd): Test the effectiveness recommendation by integrating the domain ontology (DomainOntoWP) with TermNavNet using the second-order CPM.

Case 4(Recommendation TermNetWP.1st): Test the effectiveness of the semantic- enhanced Web-page recommendation by integrating the semantic network of Web-pages (TermNetWP) with TermNavNet using the first-order CPM.

Case 5(Recommendation TermNetWP.2nd): Test the effectiveness of the semantic- enhanced Web-page recommendation by integrating the semantic network of Web-pages (TermNetWP) with TermNavNet using the second-order CPM.



	0.4	0.5	0.6	0.7	1
R.PLWAP	0.087	0.083	0.067	0.052	0.052
R.DO.1st	0.196	0.223	0.095	0.206	0.273
R.DO.2nd	0.201	0.176	0.16	0.143	0.251
R.TN.1st	0.153	0.102	0.12	0.207	0.061
R.TN.2nd	0.242	0.167	0.275	0.286	0.124

	0.4	0.5	0.6	0.7	1
R.PLWAP	0.141	0.134	0.097	0.081	0.081
R.DO.1st	0.28	0.342	0.16	0.285	0.321
R.DO.2nd	0.315	0.276	0.252	0.206	0.298
R.TN.1st	0.234	0.173	0.2	0.349	0.122
R.TN.2nd	0.346	0.251	0.344	0.358	0.153

Fig 1 Results for Experimental cases 1-5 (Rec.Len= 5)

IV. CONCLUSION

Privacy is an important initial problem in Online Social Networks (OSNs). While web sites are increasing rapidly in popularity, the existing policy-configuration tools are difficult for regular users to recognize and use. Incorporating content information into collaborative filtering can significantly improve predictions of a trusted system. This paper envisioned online social networks with trust values being incorporated to more critical systems to judge statements. Intelligence professionals can assign trust based on how much the online user's trust the information and analyses provided by other online users. That, in turn, is used with background about the experimental reports to increase the degree of information quality in the trusted system. This paper, represented a two methodology to integrating trust, provenance, and annotations in Semantic Web systems. First, implemented an algorithm for computing personalized trust recommendations using the provenance of existing trust annotations in online social networks. Then, introduced two applications that combine the computed trust values with the provenance of other annotations to personalize websites.

The metrics of the proposed model for content filtering using segmentation and personalization renders as follows:

- Alternatively of blocking the web page in cases where the content to be blocked is present only at a portion of the page, the proposed model provides a distinct benefit to online user.
- Immersion of personalization in the blocking process provides a tailor made content filtering system based on the online user's required.

The performance of our system can be boosted by using the methods described earlier. Experimental comparing the different approaches of combining content and collaboration, outlined in the previous section, are also needed.

As a future work, planning to support similarity search within our classes supplemented with semantic information gathered from URL information in the *Google+*. We trust that this will result in more precision and be particularly useful when *Google+* is accessed on portable devices where performance and accuracy are the major concerns.

V. ACKNOWLEDGMENT

For this work, I got sources from web crawler . This research was supported by Bharathiyar Institute of Engineering for Women.

REFERENCES

[1] Edmonton, "Content-Boosted Collaborative Filtering for Improved Recommendations," Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-2002), pp. 187-192, Canada, July 2002.

[2] Carullo. M.,"Clustering of Short Commercial Documents for the Web," Proc. 19th Int'l Conf. Pattern Recognition (ICPR '08), 2008.

[3] Vanetti. M., "Content-Based Filtering in On-Line Social Networks," Proc. ECML/PKDD Workshop Privacy and Security Issues in Data Mining and Machine Learning (PSDML '10), 2010.

[4] Sriram.B., "Short Text Classification in Twitter to Improve Information Filtering," Proc. 33rd Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '10), pp. 841-842, 2010.

[5] Fang. L and LeFevre. K, "Privacy Wizards for Social Networking Sites," Proc. 19th Int'l Conf. World Wide Web (WWW '10), pp. 351-360, 2010.

[6] Kuppusamy. K. S and Aghila. G, "A personalized web page content filtering model based on segmentation," International Journal of Information Sciences and Techniques, Vol.2, No.1, January 2012.

[7] Marco Vanetti et al, "A System to Filter Unwanted Messages from OSN User Walls",IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 2, February 2013.

[8] Hongyu Gao Northwestern University Evanston, IL, USA, "Towards Online Spam Filtering in Social Networks," February 2012.

[9] Adomavicius. A and Tuzhilin.G, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," IEEE Trans. Knowledge and Data Eng., Vol. 17, No. 6, pp. 734-749, June 2005.

[10] Chau. M and Chen. H, "A Machine Learning Approach to Web Page Filtering Using Content and Structure Analysis," Decision Support Systems, Vol. 44, No. 2, pp. 482-494, 2008.

[11] Mooney. R. J and Roy. L, "Content-Based Book Recommending Using Learning for Text Categorization," Proc. Fifth ACM Conf. Digital Libraries, pp. 195-204, 2000.

[12] Sebastiani. F, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, Vol. 34, No. 1, pp. 1-47, 2002.

[13] Belkin. N. J and Croft. W. B, "Information Filtering and Information Retrieval: Two Sides of the Same Coin?" Comm. ACM, Vol. 35, No. 12, pp. 29-38, 1992.

[14] Denning. P .J, "Electronic Junk," Comm. ACM, Vol. 25, No. 3, pp. 163-165, 1982.

[15] Foltz. P. W and Dumais. S.T.,"Personalized Information Delivery: An Analysis of Information Filtering Methods," Comm. ACM, Vol. 35, No. 12, pp. 51-60, 1992.

[16] Christo Ananth, A.Regina Mary, V.Poornima, M.Mariammal, N.Persis Saro Bell, "Secure system to Anonymous Blacklisting", International Journal of Advanced Research in Biology, Ecology, Science and Technology (IJARBEST), Volume 1,Issue 4,July 2015,pp:6-9

[17] Pollock. S, "A Rule-Based Message Filtering System," ACM Trans.Office Information Systems, Vol. 6, No. 3, pp. 232-254, 1988.

[18] Baclace. P. E, "Competitive Agents for Information Filtering", Comm. ACM, Vol. 35, No. 12, pp. 50, 1992.

[19] Hayes. P. J., "TCS: A Shell for Content-Based Text Categorization," Proc. Sixth IEEE Conf. Artificial Intelligence Applications (CAIA '90), pp. 320- 326, 1990.

[20] Amati. G and Crestani. F, "Probabilistic Learning for Selective Dissemination of Information", Information Processing and Management, Vol. 35, No. 5, pp. 633-654, 1999.

[21] Pazzani. M.J. and Billsus. D, "Learning and Revising User Profiles: The Identification of Interesting Web Sites," Machine Learning, Vol. 27, No. 3, pp. 313-331, 1997.

[22] C. Apte.,, "Automated Learning of Decision Rules for Text Categorization," Trans. Information Systems, Vol. 12, No. 3, pp. 233-251, 1994.

[23] Dumais. S, Platt. J, Heckerman.D, and Sahami. M., "Inductive Learning Algorithms and Representations for Text Categorization,"Proc. Seventh Int'l Conf. Information and Knowledge Management (CIKM '98), pp. 148-155, 1998.

[24] Lewis. D. D, "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task," Proc. 15th ACM Int'l Conf. Research and Development in Information Retrieval (SIGIR '92), N.J. Belkin, P. Ingwersen, and A.M. Pejtersen, eds., pp. 37-50, 1992.

[25] Schapire. R. E and Singer. Y, "Boostexter: A Boosting-Based System for Text Categorization," Machine Learning, Vol. 39, No. 2/3, pp. 135-168, 2000.