© 2019 IJSRSET | Volume 6 | Issue 3 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology
DOI : 10.32628/ IJSRSET196344

Web Page Recommendation Using Key Information Extraction and PREWAP

Rupali P. Patil

Department of Computer Engineering, JSPMs Imperial College of Engineering and Research, Wagholi, Pune, India

ABSTRACT

Web page recommendation is the technique of web site customization required by individual user or group of users. The web page recommendation system exploits the patterns of the web pages visited by users. Our proposed system is oriented towards improves performance of a web page recommendation system with minimum time computation and memory usage achieved by using various models; the first model is Web Usage Mining which utilizes the web logs. The second model also utilizes web logs to represent the domain knowledge, here the domain ontology is used to solve the new page problem. Extracting key terms from the generated ontology, use of redundant ontological information reduced using Key Information Extraction Algorithm. Also implements PREWAP algorithm which gives better results than existing PLWAP for updated dataset. Likewise, the prediction model, which is a network of domain terms, which is based on the frequently viewed web-pages and represents the integrated web usage. The recommendation results have been successfully verified based on the results which are acquired from a proposed and existing web usage mining (WUM) technique.

Keywords : Web-page recommendation, domain ontology, PLWAP, PREWAP, knowledge representation, key information extraction, web mining.

I. INTRODUCTION

The explosive development of data on the World Wide Web (WWW) with the development of advanced electronic devices has made Web data increasingly critical in just about everyone's life. The fast presentation of current websites has overwhelmed Web users by offering various choices. Consequently, Web users tend to make poor decisions when surfing the Web due to a failure to cope with enormous measures of data. Recommender systems have proved in current years to be a valuable means of helping Web users by giving useful and effective recommendations. The core methods in recommender systems are the learning and prediction models which learn user's behavior and evaluate what users might want to view in the future. Specifically, a recommender system can suggest various popular items from a large set of items based on the knowledge acquire about an active user [1]. Web-page recommender systems are some sort of recommender systems, which can naturally recommend Web-pages that are many times stimulating to a specific user based on the user's current Web navigation behavior. Since a website is normally designed to present the index pages on the home page, the index pages take the role of directing users to the recent pages on the website through Web-page joins whereas with the index pages, a user normally needs to navigate a number of Webpages to reach the content page they are interested in. On the off chance that the index pages of a website are not well designed, which is often the case, users will struggle to discover useful pages and are likely to leave the site. For a commercial website, this will like the citizen's will unsatisfied. So, web-page recommender systems have

become important for helping the web users to discover the most interesting Web-pages on particular websites. Better Webpage recommendations can improve website usage and user fulfillment [2]. To make attractive Web-page recommendations to users without extra data from those users is a hot research theme. Noteworthy effort has been devoted to developing effective Web-page recommender systems; however, a number of challenges have been encountered in the development of web-page recommender systems.

The use of semantic information in recommender systems is exploring because it can offer the chance to enrich knowledge about Web-pages, and it has become a perfect answer for offering efficient web-page recommendations. The backbone method for knowledge representation is ontologies. Ontological representation of the knowledge can be machine understandable and can help in interpreting and reasoning about the Web access patterns discovery in the mining step. Machine enables knowledge integration and automated processes [3]. This study enabled a system for a new semantic-enhanced webpage recommender system (SWRS) and techniques to resolve the limitation. Based on the system, an attractive Web-page recommender system can be developed to offer web users the N most visited web-pages from the recently visited Webpage. The knowledge bases utilize in this framework, containing the site domain and Web use knowledge bases [14]. Ontologies in recommender systems have to predominantly been constructed physically by system developers in meeting with domain experts. Ontology development is a critical process which is immoderate and work intensive, and demands a high state of proficiency in the domain [2]. It is a huge challenge to design and develop ontology for a website because there are typically a huge number of pages on one website and some vital concepts in the ontology may be focus by developers.

This paper is composed further as: Section II talks about related work studied till now. Section III presents implementation details, algorithm used and mathematical model and experimental setup. The section IV contains results and discussion of the project work done so far. Section V ends with the conclusions and presents future work. At the end we have mentioned various references used in this paper.

II. LITERATURE REVIEW

Web usage mining means to discover several meaningful patterns from the Web usage information, for example, click streams, user exchanges and users Web access activity, which are often stored in web server logs [4]. A Web server log records user sessions of going by Web-pages of a website step by step. It can be used to discover potentially useful Web usage knowledge, e.g. the navigational behavior of users, [5]. Normally speaking, a Web usage mining technique includes three steps: pre-processing, mining, and interrogate mining results [6]. After pre-processing web log files, Web access sequences (WAS), for example, are made and filed in a dataset [7]. An element of dataset is a sequence of generating a user session for browsing. In the mining step, some mining methods, for example, clustering, association rules, and sequential pattern discovery [8], can apply to the WAS to extract the frequent Web access patterns (FWAP), which is useful in web usage. In the third phase, the knowledge will be used in a particular application, e.g., a recommender system for web page, in which FWAP is required for generating the rules to backing on-line Web-page recommendation. The mining step utilizing sequential pattern mining method is the core phase in a WUM method and assumes a urgent role in a Web-page recommender system to help users to make better decision based on their recent history of web navigation.

A. Markov model for likelihood web pages:

As per [9], the Markov model is an efficient and probabilistic model to calculate the likelihood of going to Webpages. Each Web-page is checked to as a state in the Markov model. Specifically, the N-order Markov model can be known to the next recent visited page based on the previous N-1 visited pages. The probability of the N-order Markov model is greater than the lower-order system; however, the number of steps used in a large-order Markov model will gradually increases. Because the complexity is calculated by the several of stages, the complexity of a greater-order Markov model increases when utilizing it to model a large number of Webpages. Crossover probabilistic predictive models based on the Markov model, for example, the element of clustering-based Markov model of [9], have indicated improved prediction exactness over the Markov model. So, the complexity of the Markov-based models has caused about when they utilize in Web-page recommender systems reason is there are a large number of pages in a website. One efficient approach to minimize complexity of a Markov-based model is to filter out the Web-pages in the Web usage information which is not relevant. Some clustering mechanism has been applied to filter Web-pages, for example, Expectation maximization (EM) and k-means in [10]. Moreover, by brushing with connection, probabilities to the fetch Web-pages based on their need in the website's navigational chart and fabricate a crossover probabilistic predictive system based on Markov models for Web-page recommendation. This methodology achieves more objective and predictions, and gives ranked recommendations to Web users. Then again, the tree-based methods are additionally contributed fundamentally to modeling the navigational technique in websites [4]. Using such approach, a complete tree structure is manufactured to model the navigational trails in session information. Every node in the tree presented a navigational subsequence from the root to page, and cite by the page and the number of occurrences of that subsequence in the session information. The best thing of this methodology is that it is very efficient to see navigational patterns, and not really tough to get the backing and confidence of navigational patterns. The support factor and confidence of navigational pattern is evaluated which is based on its occurrence in the sequences [4]. The limitation of this technique is the crucial state of space complexity, especially for websites that have a large number of Web-pages. Recent study has evaluated that the WAP-tree based outperforms the other pattern mining method, e.g. Apriori-based, and pattern-development based technique, in terms of memory [11]. This states that the large WAP-tree based methodology, for example, pre-order linked WAP-tree mining (PLWAP-Mine) and conditional sequence mining (CS-Mine), in recommender models [12].

B. Web usage mining

Web usage mining (WUM) is a valuable technique for investigating Web usage information to get it Web client navigation practices and discover valuable Web usage learning. For an e-commerce organization, WUM can be utilized for discovering viewpoint clients who likely make a huge number of buys, or foreseeing e-commerce exchanges focused around the perception of past visitors. In the setting of web-page recommender frameworks, WUM can be utilized to find Web usage learning to help clients to settle on better choices by recommending prevalent Web-pages to the clients or a more effective approach to arrange sites for Web-based applications. Picking a successful mining algorithm assumes an imperative part in prescribing the right level of data to online users. The objective of WUM is to capture, display, and investigate the behavioral patterns and profiles of clients associating with a site [4]. The found patterns are generally a situated of successions of pages that are often gotten to by gatherings of clients with basic investments. Mining algorithms are fitting for this reason since they can take the Web Access Successions (WAS) as the info and yield the Frequent Web Access Patterns (FWAP). The vision of the Semantic Web is to empower machines to translate,

comprehend, and process data in the World Wide Web so as to react clients appeals [13]. Ontology has been considered as the foundation of the Semantic Web innovation for speaking to and imparting information between Web applications since 1980s. This study creates learning representation models for space and Web usage learning utilizing ontology technique. This segment will have different meanings of ontology; quickly express the parts of ontology in applications, and language on present ontology, variations of ontologies, and illustrations of ontologies. It then depicts the issues of ontology development, learning ontology, and ontology thinking. At long last, it records a few systems for ontology assessment.

III. SYSTEM ARCHITECTURE

A. Architecture Overview

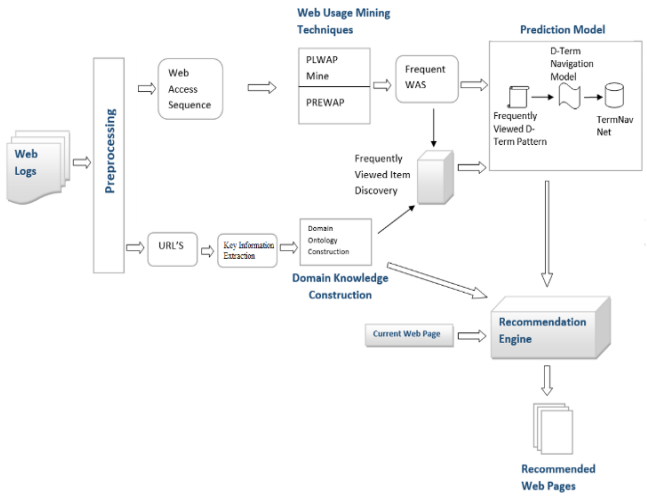

Fig 1. System Architecture

Our work represents a method to provide better Web page recommendation based on Web usage and domain knowledge, which is supported by knowledge representation models and a set of Web-page recommendation strategies.

1. PREWAP that will generate frequent view term discovery gives better result than existing PLWAP for updated dataset.

2. Domain Read Web Log / Dataset: It consist of raw data that is surfing history of web users. Web log contain the web access sequences and the URL's set which are divided in the preprocessing phase to pass the content.

3. Preprocessing: Used to extract useful information from non-useful data that is raw data. Web log having data in the form of URLs and web access sequence and then passed to further modules.

4. Web Usage Mining: This technique used to analyze web access pattern that are frequently used by web users. Because of analysis of web access pattern recommendation system will perform better then usage mining techniques are applied Ontology Construction: Generate Ontology using a key information extraction is designed to extract terms from the Web-page titles. For key information extraction, TF-IDF algorithm is used and then extract the important data / terms and relations between terms which will overcome the new page problem improves the efficiency of system.

5. Prediction Model: From frequent view term discovery prediction model will generate frequently viewed D-Term Patterns.

6. Recommendation Engine: On User Input the recommendation engine will recommend the related web pages which user likely to visit.

In the prototype, the domain and Web usage knowledge bases are implemented in OWL, which is a commonly used ontology language. The various algorithms, PREWAP mine, TF IDF, DomainOntoWP, [14] are seamlessly integrated in an automated fashion. These support the set of recommendation strategies which can predict the Web-pages those are next most likely to be visited for a given user.

B. Algorithms

1) Key Information Extraction Algorithm (TFIDF)

Input: Set of URL titles and Set of Web Pages(WP), threshold(th)
Output: Set ofKeyTerms(KE)
For All URL Titles
1: Stopword Removal
2: Stemming Will Give set of terms
3: END LOOP
4: for each term ti //Calculate Term Frequency
5: tf=frequency of term ti in all WebPages
6: IDF = log | (N / (ti WP)) |
7: tf-idf = tf ∗ idf
8: Consider terms as key term
9: If th<=tf-idf of ti

2) Construction of PREWAP-Tree

Input : A database WASD, Minimum support threshold $\lambda(0 \prec \lambda \leq |WASD|)$
Output: PREWAP-tree
Process:

1. /* produce the root */ Add a 'root' node as the root node of PREWAP-tree T;
2. /* produce 'vent name' and 'occur' values of node*/ For each Web access sequence S in WASD, do /* doing (a) and (b)*/
   (a) Delete all the events in S which don't meet the support λ, and gain frequent subsequence S' (e1e2...en).
   Set current Node to the leftmost child of root in T;
   (b) For i=1 to n (the length of S') do /* doing (A) and (B)*/

   (A) If current Node is NULL

   Create a new child node (ei:1);
   Else if current Node is labeled ei
   Set Node Exist to true;
   Else
   Set current Node to current Node's sibling until ei will be found or the sibling is NULL;

   If Node Exist

   Increase count ei by 1 and set current node;

   occur=occur+1;

3. Then pre-visit T from root: 'root left sub tree right sub tree', and at the same time add all nodes with the same event to a linkage queue and record the serial number when one node is visited;

4. End.

3) DomainOntoWP and the first-order CPM

1. Builds DomainOntoWP ;
2. Generates FWAP using PLWAP-Mine;
3. Guilds FVTP;
4. Guilds a 1st-TermNavNet given FVTP;
5. Identifies a set of currently viewed terms tk using query
6. Topicman(dk) on DomainOntoWP;
7. Infers next viewed terms tk+1 given each term in tk using query RecDTerm(tk) on the 1st-order TermNavNet;
8. Recommends pages mapped to each term in tk+1 using query Pageman (tk+1) on DomainOntoWP.

C. Mathematical Model

1) Conceptual Prediction Model: N = (tx, δx) j txT: a set of term along with the corresponding occurrences counts,

φ= (tx, ty, δx,y, px,y) | tx, tyT : a set of transitions from tx to ty, along with their transition weight

(δx,y), and first order transition probabilities (px,y)

M = (tx, ty, tz, δx,y,z ,...,px;y;z) | tx, ty, tzT : a set of transition from tx,ty, tz , along with their transition weights (δx,y,z) and second order transition probabilities (px,y,z) If M is non empty, the

CPM is considered as the second order conceptual predication model .

2) First Order Transaction Probability: CPM states= {S, t1..., tp,E}

N =|F| j is the number of term pattern in FPs = First Order Transaction Probability S and tx are the two states First Order Transaction Probability from state S and tx is given by:

$$\rho S,x = \frac{\delta_{S,x}}{\sum_{y=1}^{N} \delta_{S,y}}$$

First order Transaction Probability from state tx to ty is given by,

$$\rho S,y = \frac{\delta_{S,y}}{\delta_x}$$

The First Order Transaction Probability from state tx to Final State E is given by,

$$\rho x,E = \frac{\delta_{x,E}}{\delta_x}$$

IV. RESULT AND DISCUSSION

A. Experimental Setup

All the experimental cases are implemented in Java in congestion with Netbeans tools, algorithms and strategies, and the competing data mining approach along with various feature association rule generation technique, and run in environment with System having configuration of Intel Core i5-6200U, 2.30 GHz Windows 10 (64 bit) machine with 8GB of RAM .

B. Dataset Description

The data was created by sampling and processing the www.microsoft.com logs. The data records the use of www.microsoft.com by 38000 anonymous, randomly-selected users. For each user, the data lists all the areas of the web site (Vroots) that the user visited in a one-week timeframe. Users are identified only by a sequential number, for example, User 14988, User 14989, etc. The file contains no personally identifiable information. The 294 Vroots are identified by their title (e.g." NetShow for PowerPoint") and URL (e.g."/stream"). The data comes from one week in February, 1998. Each instance represents an anonymous, randomly selected user of the web site. Each attribute is an area ("vroot") of the www.microsoft.com web site. Missing Attribute Values: The data is very sparse.

C. Comparison Results

This section presents the performance of the PLWAP Mine and PREWAP algorithms. Table 1. shows the different threshold value and required time in millisecond for PLWAP and PREWAP respectively.

Table 1: Execution Time Comparison Between PLWAP and PREWAP Algorithm with different threshold values.

Threshold	PLWAP	PREWAP
10	3566	3547
20	2701	1989
30	2434	1266
40	2241	1226

Figure. 2 Shows the Time comparison of PLWAP and PREWAP with Key Terms Extraction Terms algorithms for various threshold. The X-axis shows various threshold and Y- axis shows Time in milliseconds (ms) to run the algorithms. The PREWAP takes less time than PLWAP for mining. With Increasing in threshold values the time required to run the algorithm is reduced as shown in graph.

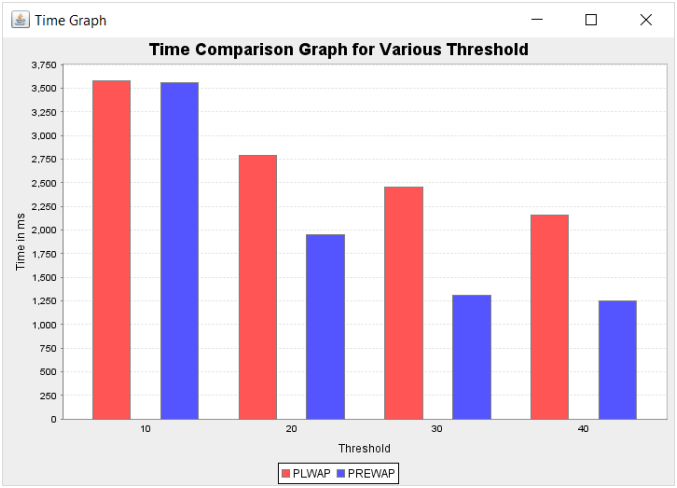

Fig 2. Comparison graph of PLWAP and PREWAP algorithm (X-axis: Number of web pages, Y-axis: Time in ms)

Figure. 3 Shows the Line graph of Time comparison between PLWAP and PREWAP with Key Terms Extraction Terms for various dataset size. The X-axis shows various datasets (size) and Y- axis shows Time in milliseconds (ms) to run the algorithms. The PREWAP takes less time than PLWAP for mining large dataset. Table 2. shows the different dataset sizes and required time in millisecond for PLWAP and PREWAP respectively.

Table 2: Execution time comparison between PLWAP and PREWAP Algorithm with various dataset sizes

Dataset	PLWAP	PREWAP
Dataset 1	1090	734
Dataset 2	1039	544
Dataset 3	1206	686
Dataset 4	1511	937
Dataset 5	1628	1047

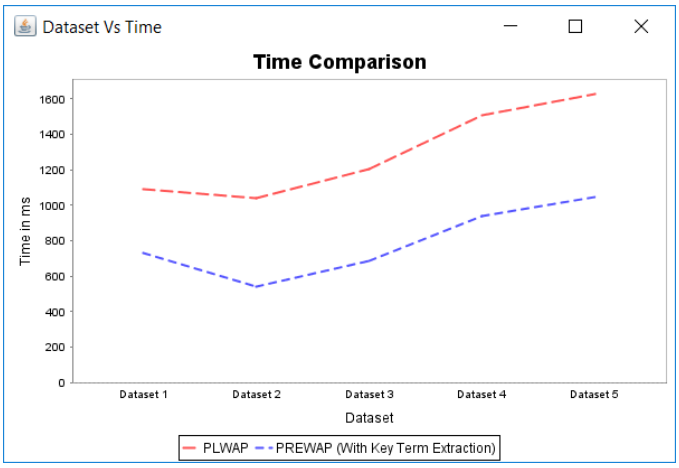

Fig. 3. Time Comparison Graph

Figure 4. Indicate the Memory comparison of PLWAP and PREWAP with Key Terms Extraction Terms algorithms. The X-axis shows algorithms PLWAP and PREWAP with Key Terms Extraction Terms algorithms and Y- axis shows memory in bytes. The PREWAP takes less memory than PLWAP as PREWAP uses header table at tree construction time which helps to reduce the time and memory at searching node in tree.

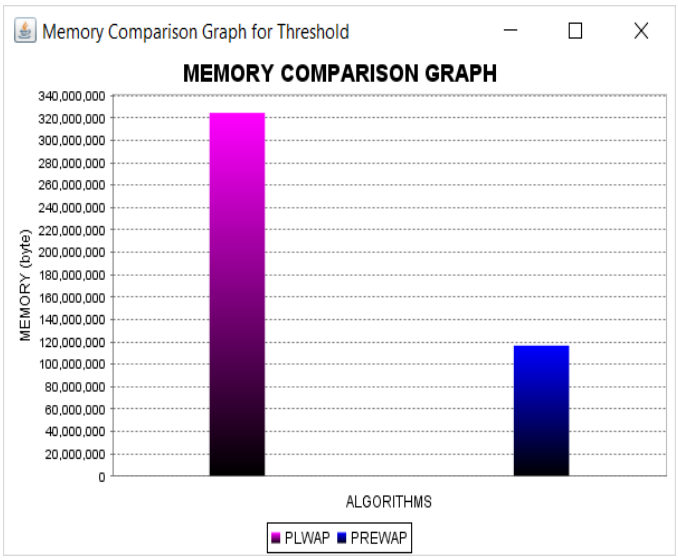

Fig. 4: Memory Comparison Graph

V. CONCLUSION

Thus we developed an efficient web page recommendation system based on domain ontology and web usages mining. It is observed that the adopted strategy gives better results than the existing algorithm. It reduces computation cost for web page recommendation. TF-IDF key term extraction to get the important terms from web logs from which domain ontology is build and also these terms used while PREWAP tree construction. When dataset gets updated and there are too many small frequent item sets generated in such case PLWAP Mine will not work properly in terms of execution time and it requires more memory. PREWAP take less time and less memory compare to PLWAP Mine. So PREWAP is better than PLWAP Mine algorithm. The study has made significant contributions from both theoretical and practical aspects in the area of Web-page recommender systems.

VI. REFERENCES

[1] Konstan, J. and Riedl, J. 2012, "Recommender Systems: from Algorithms to User Experience", User Modeling and User-Adapted Interaction, vol. 22, no. 1, pp. 101–123.

[2] Gunduz-Oguducu, 2010, "Web Page Recommendation Models: Theory and Algorithms", Morgan and Claypool.

[3] Zhengyu Zhu,Meiyu Zheng ,2015 ,"Web Log Frequent Sequential Pattern Mining Algorithm Linked WAP-Tree"

[4] Domingue, J., Fensel, D. and Hendler, J.A. 2011, "Introduction to the Semantic Web Technologies", in J. Domingue, D. Fensel and J.A. Hendler (eds), Handbook of Semantic Web Technologies, SpringerVerlag Berlin Heidelberg, pp. 3–41.

[5] Liu, B., Mobasher, B. and Nasraoui, O. 2011, "Web Usage Mining", in B. Liu (ed.), Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Springer–Verlag Berlin Heidelberg, pp. 527–603.

[6] Mobasher, B., Burke, R., Bhaumik, R. and Williams, C. 2007, "Toward Trustworthy Recommender Systems: An Analysis of Attack Models and Algorithm Robustness", ACM Transactions on Internet Technology, vol. 7, no. 4, p. 23.

[7] Woon, Y. K., Ng, W. K. and Lim, E.-P. 2005, "Web Usage Mining: Algorithms and Results", in A. Scime (ed.), Web Mining: Applications and Techniques, IGI, pp. 373–392.

[8] Ezeife, C. and Liu, Y. 2009, "Fast Incremental Mining of Web Sequential Patterns with PLWAP Tree", Data Mining and Knowledge Discovery, vol. 19, no. 3, pp. 376–416.

[9] Pierrakakos, D., Paliouras, G., Papatheodorou, C. and Spyropoulos, C.D. 2003, "Web Usage Mining as a Tool for Personalization: A Survey", User Modelling and User Adapted Interaction, vol. 13, no. 4, pp. 311-372.

[10] Borges, J. and Levene, M. 2004, "A Dynamic Clustering-Based Markov Model for Web Usage Mining", Available online at http://xxx.arxiv.org/abs/cs.IR/0406032.

[11] Khalil, F. 2008, "Combining Web Data Mining Techniques for Web Page Access Prediction", Doctoral thesis thesis, University of Southern Queensland.

[12] Mabroukeh, N.R. and Ezeife, C.I. 2010, "A Taxonomy of Sequential Pattern Mining Algorithm", ACM Comput. Surv., vol. 43, no. 1, pp. 1–41.

[13] Ezeife, C. and Liu, Y. 2009, "Fast Incremental Mining of Web Sequential Patterns with PLWAP Tree", Data Mining and Knowledge Discovery, vol. 19, no. 3, pp. 376–416.

[14] Henze, N., Dolog, P. and Nejdl, W. 2004, "Reasoning and Ontologies for Personalized E-Learning in the Semantic Web", Educational Technology and Society, vol. 7, no. 4, pp. 82–97.

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