Importance of Domain Knowledge in Web Recommender Systems

Saloni Aggarwal
Student
UIET, Panjab University
Chandigarh, India

Veenu Mangat
Assistant Professor
UIET, Panjab University
Chandigarh, India

ABSTRACT
Web usage data is extensively used in every domain to analyze the browsing behavior of the users who visited the website or search engines. Web usage data is contained in server logs called as Web Logs. This data enables the website owners to infer the needs and interests of the users for using this information to increase the revenue from their web business. The website owners employ recommender systems for this purpose. The recommender systems exploit web usage data to predict what web pages the user will visit next and therefore offer the recommendations for those very pages to the user and offers them support while browsing. This in turn helps users to have a better browsing experience, personalized support and hence, probability of user buying out the products from that website increases. Web usage mining alone is used in traditional recommender systems. Modern recommender systems employ semantic knowledge base i.e. domain knowledge in addition to web usage mining for efficient prediction of pages as this helps in avoiding the new page problem. This paper presents a comparative and comprehensive study of modern and traditional recommender systems.

General Terms
Web Recommender System, Domain Knowledge, Web Usage Mining.

Keywords
Semantic Knowledge; Domain Knowledge; Web Usage Data; Personalized Services.

1. INTRODUCTION
Web usage mining is a sub field of web mining [1]. Web mining is divided into three categories- Web usage mining, Web content mining, Web structure mining. Web content mining means mining and analyzing the text, images, audio , video and other type of content that is available on the web pages to extract useful content for the desired purpose. Web structure mining is used to extract structuring information from the web at the document level or the hyperlink level in order to discover structured summaries about the information on web. Web Usage Mining is defined as the extraction of useful knowledge from the web logs to capture information about how the web users browse the web in order to employ it for suitable purposes. Web Usage Mining forms the core of any type of Recommender Systems on web. Almost all web portals employ such systems. Web Usage Mining is based on web log data of websites. All the data like the pages visited by users, the order of visiting the pages etc. is present in logs.

2. TRADITIONAL RECOMMENDER SYSTEMS
The traditional recommender systems rely on the web usage data alone for generating recommendations. Many new approaches have come forward with web usage data as the base and taking other aspects like domain knowledge of website [2], relational databases [3] and other techniques to improve the efficiency of recommendations. Web Usage Data is however the core of recommender systems.

Web Usage Data: Web logs contain the web usage data. These logs contain clickstream data that is whatever the user clicks in his browsing session is all stored in the web logs. The analysis of clickstream data is also called as clickstream analysis.

The type of data that can be mined from the web logs is[4]: analyzing the time of maximum visits to the website, the type of data is most searched for on the search engines, what type of users visit the website during a particular period of the month or year, the frequency of visits to a web page, the time duration for which the user stays on a web page, user rating about the web page, the peak hours of the website visits in a day, the sequence of visiting the web pages by a category of user, which webpage is mostly visited given a particular current web page and much more. Depending upon the needs, different sectors extract different type of information from the logs to improve their profits from the web. All credit goes to web usage mining which is responsible for providing such benefits to the web community. The various type of sectors and how they use web usage mining are listed as follows:

2.1 Finance Sector
The most crucial information for banks is the needs and purchasing patterns of their customers. This information is used for offering personalized offers to customers, retaining the customers and managing the customer drop outs and doing resource allocation efficiently [5]. The tasks that can be done accomplished are as follows: Identifying the fraudulent transactions, Giving personalized offers to customers depending on their past purchasing pattern, Improving the performance of the bank's website, Identifying the potential customers and developing a target market, Acquiring new customers.

2.2 Web Commerce Sector
Web commerce includes all business that is carried on web: online tourism, shopping portals, web food websites, entertainment websites, online book stores. The advantages include: Offering the products that the user has interests in readily visible on the webpages so as to provide easy and quick access to those products [6], Identifying the target

customers and avoiding the marketing cost of marketing to non-interested customer segments, Web site layout is optimized for better browsing quality for users, Intelligence business system for predicting the demands of customers by mining customer's probability of purchase.

2.3 Search Engine Sector

When a user enters a search query in the search engine form, the search engine displays a number of web pages on the result page. The web page of highest importance is displayed on the top and the lowest priority web page is displayed at the bottom. Hence, the web pages are displayed in the decreasing order of their relative importance. Grading system [7] is used for assigning the priorities in which a web crawler is deployed, web crawler analyses the possible list of web pages that are to be displayed as a result of the search query, generates a queue of their URLs and crawls them thereby bringing the highest grade web page to the user from the server. This web crawler is given higher access to the web pages having higher grade.

2.4 Mobile Search Engine Sector

The web search engines of handheld devices are little different from the normal search engines because of the short form space leading to shorter queries. Shorter the query, less relevant the results will be. Therefore to get good results in response to short query, mobile devices nee to employ an architecture that analyzes the web usage logs to bring relevant results in top and least relevant on bottom. PMSE client server architecture has been which works as follows [8]: firstly user profile information is used and interests of the user are captured by mining on past web sequences of the user in the web logs, A feature set is designed that contains user preferences personalized according to his profile information, then the results that were to be displayed otherwise are re-ordered as per user preferences to give personalized results.

2.5 Social Network Sector

Social networks use web usage mining to increase the number of connections among the social networks users and thereby increasing the amount of usage of their websites [9]. The option of "people you might know" is an example of this type of connection increasing strategy. The recommender system employed in social networks analyze the web usage of the users from the web logs and find out the users with similar interests, large number of common friends, relevance between two people based on their geographical location or background history like same education place etc. Another benefit [10] is "opinion mining" which is done extensively on twitter. Today the most extensive usage of web mining is in the field of social networks. Another application of web usage mining through social websites is the advertisement feature where the e-commerce websites can offer various products and services to the users on the social media. It acts as an advertising platform for e-commerce industry.

2.6 System Improvement

Web usage mining is employed for understanding the traffic on web that provides the basis for making plans and policies data distribution, network settings and transmission, Web caching or load balancing [11]. Also web usage mining helps discovering patterns that help detecting intrusion, attempted break-ins etc. corresponding to which web sites can be modified and the probability of such frauds can be decreased.

2.7 Web Designing Sector

Web usage mining also helps in maintaining the website attractiveness. The web site layout is an important factor for many websites for example, an e-commerce website would display a product catalog for its customers ad that catalog should be attractive and in correct layout both structurally and in term of content. A detailed feedback on user behavior can be acquired using web usage mining thereby, helping the website designer to design the website accordingly [12].

2.8 Intra-Organizational use

An organization might need to evaluate the browsing behavior of its members in order to carry out some survey task or take some decisions about the organization as a whole. This can be done by mining the log of the website. The members could be students, employees, managers, faculty etc. depending upon the domain. Information extracted would be- what is the peak time of usage of the website, what type of modules are more frequently visited, what subjects interest the users more etc.[13].

3. TWO APPROACHES

Web page recommender systems are based on either of the two approaches [14]:

3.1 Content based Filtering

In content based filtering the recommendations are made based on the amount of interest of the user in different web pages/ products. The interest is estimated from past navigational behavior of the user- the rating of the web pages by the user, the sequence of web pages mostly visited by the user and the user responses to various questions asked to the user. It has a limitation that the users do not always give accurate ratings or correct answers to the survey questions. Also new user problem may be encountered.

3.2 Collaborative Filtering

In collaborative filtering, recommendations are based on the user's similarity with other users. Similarity may be calculated using 2 factors:

3.2.1 User profile information

The users with similar profile like age, sex, geographical location are grouped into one cluster. Limitation is that it may lead to irrelevant recommendations because of no knowledge about user interests.

3.2.2 User interests

The users with similar browsing patterns are grouped. The users' past navigational patterns are analyzed and users with similar interests are put together. But it has a limitation of Sparsity and scalability [14]. The computation time of similarity increases with increase in number of users.

4. GAPS IN TRADITIONAL RECOMMENDER SYSTEMS

The various limitations of traditional approaches have been mentioned in the previous section. However these problems can be generalized and categorized under two broad categories. The two main gaps in the traditional recommender systems are that either it can show irrelevant recommendations or it might show no recommendations [15] at all. In the first case, the results are irrelevant providing zero support to the user and rather spoiling the browsing quality by offering wrong recommendations. In the latter case, in case of no recommendations, the very essence of the recommender

systems fails that is the system fails and can no longer offer any recommendations to the user. Both the cases have been discussed as follows:

4.1 Irrelevant Recommendations

New User: This happens when the User is new to the website and has no past browsing history with the website he is visiting. This means the recommender system tries to fetch predictions for the user by matching his profile information with the profile information(content based filtering) of the existing users of the website based on aspects like- same sex, same age, same geographical location etc. and then offering him the webpages based on the clickstreams of the similar users. That is the recommender system tries to offer him predictions as it would have offered to the existing users of similar nature. This might lead to inaccurate recommendations as the interests of the user might be different despite having same attributes like same sex and same age etc. Also for the new user if the system tries to compute the similarity with other users based on the interests(collaborative filtering) the problem of accommodating large number of users arises so a calculating similarity for every new user is a practical limitation, some generalization technique is required that can be applied to all the new and existing users alike.

4.2 No recommendations

New Page Problem [15]: This is the major problem that has not been addressed by either of the two traditional approaches. For large websites, this happens when a user visits some web page that has not been visited previously by anyone, neither by other users, nor by himself. Therefore, no web usage mining based recommender system (be it any approach) can offer any recommendations at all, not even in infinite amount of time. This happens because the system cannot find any clickstream sequences in the web logs that contain that very particular web page on which the user is currently which in turn means it cannot find any web page that has been visited after this web page, thereby leading to failure of recommender system in offering any kind of web page predictions to the user.

5. MODERN RECOMMENDER SYSTEMS

Modern recommender systems use the concept of Semantic Web Usage Mining. Semantic web usage mining is based upon the semantic (domain knowledge) of the website that is visited by the user. Domain knowledge or the Semantics knowledge of the website means the information about the content that is present on every webpage of the website, for example, the "contact us" web page of a website will contain content like address of the company, phone number of the company HR, the directions to reach the company, email ID etc. This type of knowledge is extracted from every web page and incorporated into a new model called as Domain Knowledge Base or the Semantic Base. This Domain Knowledge base is combined along with the Web Usage Data and then mining is done on the integrated output model that contains both web usage information and website's content knowledge to predict efficient web pages for users.

Modern recommender systems as a solution to the problems of traditional systems: The modern recommender-systems are completely free from the problem of "No recommendations" i.e. the New Page problem [15]. However the first problem i.e. New User problem has not been completely eliminated but addressed considerably because of domain-knowledge being the other pillar in addition to the similar-users being the basic pillar for offering recommendations by improving the response time.

The information about the web domain is extracted and developed into Domain Ontology (formal representation of domain knowledge as a graphical schema), then web usage data is extracted and integrated into the ontology to form a hierarchical model, then a web usage mining algorithm is implemented on top of it to offer recommendations.

HOW IT WORKS: When a user visits a New Page, Modern recommender system can offer recommendations to the user by mapping the existing web page sequences in the web logs onto the domain term sequences extracted using Domain Ontology [16]. Initially all the web pages are mapped to keywords (Domain Terms) that best represent the information contained in the web page. For example, a web page named as "*www.businessindia.com/shop-portalsupport*" is mapped to Domain terms: shop-portal and support. So when a user visits a page, say, "*www.businessindia.com/endusershop-portalsupport*" that has never been visited previously by any user then this new web page can be related to keywords shop-portal and support. And the domain term sequences (that have already been generated by mapping web page sequences to the domain terms)will be searched for, and the recommendations about which domain terms can be visited next will be generated- this in turn can again be mapped to the web page names corresponding to generated domain terms and hence, those pages will be offered as recommendations to the user.

6. STEPS IN MODERN/SEMANTICS-BASED RECOMMENDER SYSTEMS

6.1 Construction of Domain Ontology

In this phase, the domain terms are extracted from the web pages using a strategy- web page titles, metadata, html tags etc. that can accurately represent the information contained in the web page. Web page titles are the easiest and most widely used approach to extract domain terms because it is assumed that the web page titles are named as per the content of the web page and they are informative enough. This task is automated using software agents, most widely used of which is protégé [16].

After extraction of domain terms, the relationship hierarchy among the extracted terms is developed using formal ontology language that consists of relations [15] like- consists of, belongs to, has page, links to, to develop a graphical hierarchy of the domain terms among themselves. This is done on the basis of the webpages i.e. how the web pages are related to each other in the website (the links between the pages), similarly domain terms are related to each other in the domain ontology.

6.2 Web Usage Data

This is the second phase of the recommender system development. In this phase, the Web usage logs are mined to extract the web sessions for the particular user. The web session is defined as the time interval during which the user browsed the website continuously with start time t and end time l. The web session of the user is extracted that consists of a number of web page sequences that were followed by the user in that particular session.

Next, the web page sequences are analyzed one after the other by looking onto their titles(which are containers of domain terms extracted in phase 1) are chosen on the basis of their

frequency of occurrence in the web usage data. The domain term sequences are extracted from the web sessions. The domain term sequences once extracted are then mapped onto the Web pages by applying an algorithm that maps the sequences of domain terms to Web pages.This is done page by page [16] as follows:

For every Web page
For every term sequence
For each term in the sequence
If web page contains that term
Increment term count for that webpage and go to next Term
End
If the complete term sequence occurs in webpage
Map that term sequence to the web page and move to next term sequence
End
Else
Move to next term sequence
End
Else
Move to next Web page.
End

Then for every pair of web pages, the common terms are identified and weights are allotted to every pair of web pages in the graph. More the number of common terms, more will be the weight of that pair of web pages. This process generates the semantically enhanced network of web pages and the relationship between the domain terms and the web pages is also established. The web page sequences that have more weights will be stored in the knowledge base.

6.3 Web page recommendation

The current web page on which the user is present, the previously visited webpages before the current web page and the web page sequences extracted from phase2 are taken together as input. The sequences are matched on the domain ontology constructed in phase 1. The most weighted links that come on the matched sequences are considered. The position of the current web page on the weighted path of links selected from the domain ontology is identified. Now from this position, various links of this current position to the different terms to which it is related in the domain ontology are analyzed in the graph (domain ontology), the link with the highest weight is chosen. The term connecting this highest weight link is chosen as the next to be visited web page and is recommended to the user.

The resulting recommendations are more accurate and more relevant as far as user interests are concerned. Also the new page problem has been addressed completely. The performance of the recommender systems is measured using precision and satisfaction as the two parameters of verdict where precision is defined as the number of correct recommendations/ Total number of recommendations. Satisfaction is defined as the number of accessed recommendations/ Total number of recommendations. Modern recommendation systems perform considerably well as compared to the traditional systems because of being semantically enriched.

7. CONCLUSION

Web Usage Mining is a very important field to carry out research and employ web intelligence. The web logs are storehouse of all the browsing information of the website and its users. Any type of information can be extracted from these logs depending upon the purpose. Based on these logs, web recommender systems have emerged out to be not only a useful but also an attractive filed of research with time. Recommender systems rely on these web logs and web usage mining for offering recommendations. However to exploit the benefit of web usage data in a complete manner, semantic data about the website is a very potential source for exploitation. It is the most emerging field of research today. The results shown by Domain Knowledge based web usage mining are way better than traditional mining algorithms. For future work, an efficient recommender system can be developed by applying a mining algorithm as well as automated domain ontology construction other than those applied already.

8. REFERENCES

[1] Kosala, Raymond, and Hendrik Blockeel. "Web mining research: A survey." *ACM Sigkdd Explorations Newsletter* 2, no. 1 (2000): 1-15.

[2] Dai, Honghua, and Bamshad Mobasher. "A Road Map to More Effective Web Personalization: Integrating Domain Knowledge with Web Usage Mining." In *International Conference on Internet Computing*, pp. 58-64. 2003.

[3] Kazienko, Przemyshaw, and Maciej Kiewra. "Integration of relational databases and Web site content for product and page recommendation." In *Database Engineering and Applications Symposium, 2004. IDEAS'04. Proceedings. International*, pp. 111-116. IEEE, 2004.

[4] Lu Chen, Qiang Su, "Discovering User's Interest at E-Commerce Site Using Clickstream Data," 10th International Conference on Service Systems and Service Management (ICSSSM) IEEE 2013, pp. 124-129, July 2013

[5] Omer Adel Nasser, Dr. Nedhal A. Al Saiyd, "The Integrating Between Web Usage Mining and Data Mining Techniques," 5th International Conference on Computer Science and Information Technology, pp. 243-247, 2013.

[6] Yanduo Zhao, "The Review of Web Mining in E-commerce," Proceeding ICCIS'13 Proceedings of the 2013 International Conference on Computational and Information Sciences, pp. 571-574, 2013..

[7] Anupma Surya, Dilip Kumar Sharma, "An approach for web page ordering using user session", Proceedings of 2013 IEEE Conference on Information and Communication Technologies(ICT), pp. 1009-1013, 2013.

[8] Kenneth Wai-Ting Leung, Dik Lun Lee, and Wang-Chien Lee, "PMSE: A Personalized Mobile Search Engine," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 4, pp. 820-824, April 2013.

[9] Ting, IHsien. "Web mining techniques for on-line social networks analysis." In *Service Systems and Service Management, 2008 International Conference on*, pp. 1-5. IEEE, 2008.

[10] Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." *Foundations and trends in information retrieval* 2, no. 1-2 (2008): 1-135.

[11] Yang, Qiang, Haining Henry Zhang, and Tianyi Li. "Mining web logs for prediction models in WWW caching and prefetching." In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 473-478. ACM, 2001.

[12] Spiliopoulou, Myra, Carsten Pohle, and Lukas C. Faulstich. "Improving the effectiveness of a web site with web usage mining." In *Web Usage Analysis and User Profiling*, pp. 142-162. Springer Berlin Heidelberg, 2000.

[13] Wichian Premchaiswadi, Walia Romsaiyud, "Extracting WebLog of Siam University for Learning User Behavior on MapReduce," 2012 4th International Conference on Intelligent and Advanced Systems(ICIAS), vol. 1, pp.149-154, 2012

[14] Nagarnaik, Paritosh, and A. Thomas. "Survey on recommendation system methods." In *Electronics and Communication Systems (ICECS), 2015 2nd International Conference on*, pp. 1496-1501. IEEE, 2015.

[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, Oct 2014.

[16] Venu Gopalachari, M., and P. Sammulal. "Personalized collaborative filtering recommender system using domain knowledge." International Conference on Computer and Communications Technologies (ICCCT) IEEE, pp. 1-6, 2014.