International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org, editorijaiem@gmail.com
Volume 3, Issue 8, August 2014 **ISSN 2319 - 4847**

An Improved Heuristic Approach to Page Recommendation in Web Usage Mining

Aditya Kumar
Department of Information and Communication Technology
Manipal Institute of Technology

ABSTRACT

Web usage mining is an application of data mining in order to extract data from a web server log file. The commonly visited navigational paths are extracted in terms of web page addresses from web server visit logs and the patterns are used in various applications including recommendation. In this paper, an existing heuristic strategy is tested and accordingly an improvised technique is designed and developed which mimics human behavior and also adapts to the changing navigational patterns in order to give page recommendations. The method has been tested on real time data from users who visited a popular website of generic content.

Keywords: Web page recommendation, Web usage mining

1. INTRODUCTION

Web usage mining is the process of extracting useful information from server logs. It is the application of data mining techniques to discover interesting usage patterns from web data in order to understand and better serve the needs of web based applications. Usage data captures the identity or origin of web users along with their browsing behavior at a website. One of the greatest challenges of computer science is characterization of human behavior and mining the web is essential to understand it better. Now a days personalized recommendation of products, documents and collaborators has become an important way of meeting user needs. Therefore the analysis of web data is essential to provide web page recommendations.
In web usage mining all the data is taken from the web server logs because collection of data from client side is not feasible. However there are certain limitations associated with web server logs:

- The same IP address may correspond to different user in different times or different IP addresses may correspond to same user at different times.
- A crowded website may have millions of visits each day and as a result the number of meaningful clusters of users is huge. As a result the usefulness of clustering the data decreases.
- Proxy servers can mask individual IP address of the user.
- Sequence of page requests written in log file isn't a reliable indicator of the actual sequence visited because some pages are cached by browser.
- Some page visits are caused by malicious software which is difficult to isolate.

The paper is organized as follows: in the next section literature review is mentioned. In section 3, the proposed algorithm is explained. In section 4, implementation is done. In section 5, the used data is described and the results of testing of the algorithm are summarized. Finally in section 6 main conclusions are drawn and future work is outlined.

2. LITERATURE REVIEW

Numerous researches are available in the literature for web recommendation system using sequential pattern mining. Extracting the beneficial data from raw web log and subsequent transforming of these data into appropriate type for pattern discovery is the objective of data pre-processing. The size of web server log can be decreased to less than 50% of their original size by filtering out the image requests. The natural framework in which web usage mining has been studied widely is that of markovian models, mainly because navigated pages may be easily represented as a discrete sequence of symbols from a finite alphabet. Marko- vian models are also the core of predictive web prefetching algorithms (see for example [1], [2]), in which the forthcoming page accesses of a client are predicted on the base of its past accesses, to the purpose of improving cache effectiveness. Unfortunately, the underlying Markov assumption of a unique generative distribution turns out to be too strong and models of order bigger than 1 have a too high complexity [3]. Mixture of markovian models with a limited number of components have shown better performances [4], [5], however even a mixture of a few components is inadequate for the high heterogeneity of big data sets concerning human behaviour. Conversely, distance based user profiling algorithms tend to be slow and to produce too fragmented results, as the number of significant profiles may easily be huge.

3. PROPOSED METHOD

The aim of the proposed method is to give an effective prediction algorithm that may be used for online recommendations to the users of a website. The problem of page recommendation is handled at page category level instead of a single page level because the page category level avoids the problems of frequent updating of web pages which is encountered at single page level. The proposed method tries to mimic human behavior in an unknown environment crowded by other humans. The dataset used in this project comes from Internet Information Server (IIS) logs for msnbc.com and news related portions of msn.com for the entire day of September, 28, 1999 (Pacific Standard Time). Each sequence in the dataset corresponds to page views of a user during that twenty-four hour period. Each event in the sequence corresponds to a user's request for a page. Requests are not recorded at the finest level of detail, that is at the level of URL, but rather at the level of page category (as determined by site administrator). The categories are front-page, news, tech, local, opinion, on-air, misc, weather, health, living, business, sports, summary, bulletin board service, travel, msn-news and msn-sports. Any page request served via a caching mechanism was not recorded in the server logs and hence not present in the data. Other relevant information:

- Number of users: 989818
- Average number of visits per user: 5.7Initially the dataset was preprocessed to filter dirty data by:
- Removing all single input sequences.
- Removing all sequences with repetitive data.
- Replacing all data entries having 3 or more same page categories simultaneously by 1 entry only.

The database is then clustered according to the length of the input query sequences in ascending order. The first step of the algorithm is to extract all sequences from the reference dataset compatible with the query sequence and suitable to give a prediction. If the selected sequences are the same as the input sequence then the L^{th} element of all the matching sequences are stored as trusted sequences. The most occurring L^{th} element is then recommended to the user. In case the selected sequence does not match the input sequence then it is rejected and the next sequence from the dataset is taken. If no entry from the dataset is same as the input sequence then the size of input sequence is reduced by 1 and the same steps are followed recursively. The recursion stops when the length of input query sequence reaches 1.

4. IMPLEMENTATION

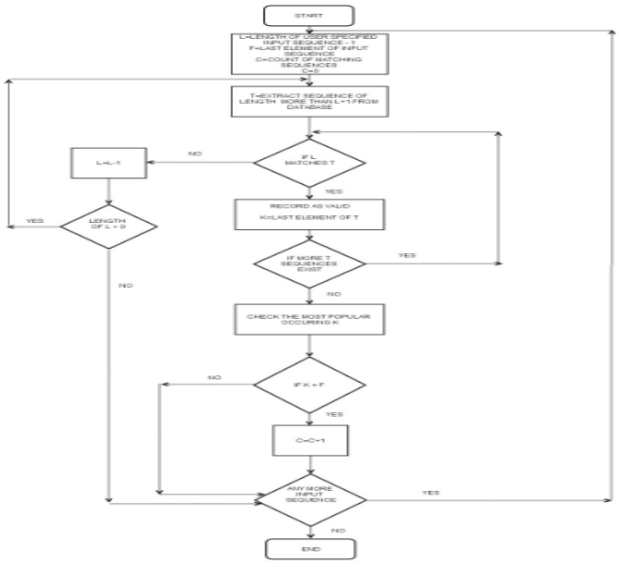

Figure 1: Flow chart for improvised heuristic approach.

According to [6] input sequence from the web server log of an existing website is chosen and only the first 1,000 sequences of size 5 are taken as input. The length of the input sequence is reduced by 1, removing the latest entry and storing it for comparison with the predicted recommendation. The input sequence is searched for in the database and similar patterns are selected. The outputs of all the patterns are taken and stored. The most frequently occurring output of the pattern is taken as the recommendation and compared with the actual existing value recorded earlier. If the recommended output matches with the original value, then increment the value of the counter which is initially set to 0. Finally the success rate of the algorithm is recorded. The same is tested for input sequences of size 6 and 7. An input sequence from the web server log of an existing website is chosen and only the last 5 terms of the input sequence are taken as input. The length of the input sequence is

reduced by 1, removing the latest entry and storing it for comparison with the predicted recommendation. The input sequence is searched for in the non clustered database and similar patterns are selected. The outputs of all the patterns are taken and stored. The input pattern is also searched in the middle of the dataset sequences. The most frequently occurring output of the pattern is taken as the recommendation and compared with the actual existing value recorded earlier. If the recommended output matches with the original value, then increment the value of the counter which is initially set to 0. Finally the success rate of the algorithm is recorded. Next, the same is tested for input sequences of size 6 and 7.Again the same are repeated but now the clustered dataset is taken and all the observations are recorded.

5.RESULTS

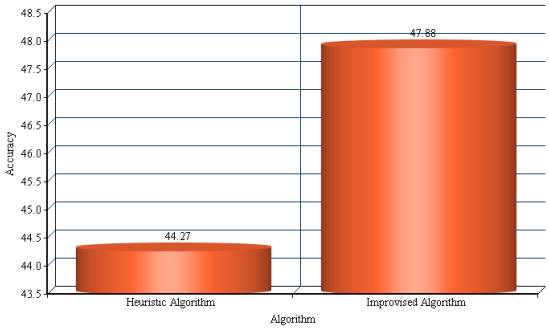

Figure 2: Comparison Of Heuristic Algorithm and Improvised Heuristic Algorithm

Initially a comparison of automated input sequences(see Figure 2) were taken using both the algorithms. The testing was done on msnbc database with first 40,000 entries. Here we find that the results of the improved algorithm are slightly more better(3.61% that means 1446 more entries were predicted correct in improved algorithm than in the one mentioned in the paper).

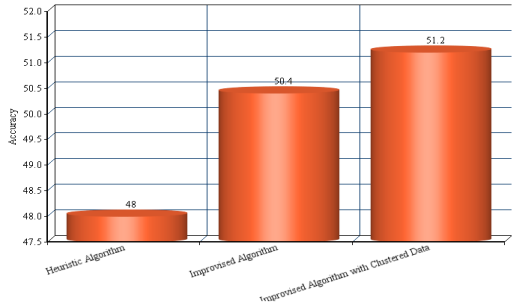

Figure 3: Comparison Of Algorithms with Input Data Sequence Of size 5

Input data sequence of size 5 was taken. 1,000 input sequences were taken and checked on 20,000 data sequences(see Figure 3). It was found that accuracy of data sets was 48% according to paper algorithm. The accuracy of the improved algorithm was 50.4%. Another accuracy was found out using the improved algorithm on the same data set but now with data clustered according to the length of the data sequences. In this case the accuracy was found to be 51.2%.An excellent style manual for science writers is [7].

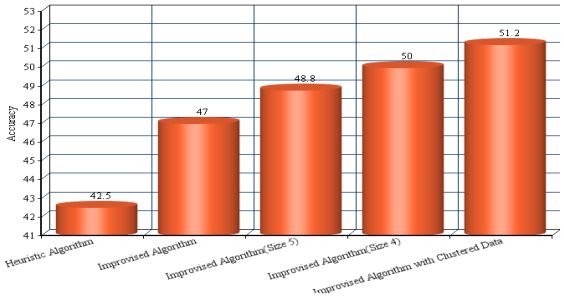

Figure 4: Comparison Of Algorithms with Input Data Sequence Of size 6

Input data sequence of size 6 was taken. 1,000 input sequences were taken and checked on 20,000 data sequences(see Figure 4). It was found that accuracy of data sets was 42.5% according to paper algorithm. The accuracy of the improved algorithm was 47%. The next accuracy was found with reducing the input size to 5 and it was found out to be 48.8%. The

next accuracy was found with input size 4 and it was 50%. Finally, accuracy was found out using the improved algorithm on the same data set but now with data clustered according to the length of the data sequences. In this case the accuracy was found to be 51.2%.

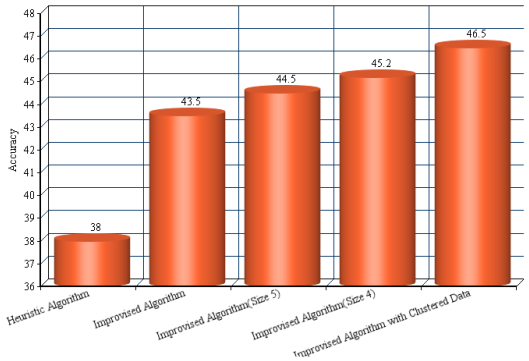

Figure 5: Comparison Of Algorithms with Input Data Sequence Of size 7

Input data sequence of size 7 was taken. 1,000 input sequences were taken and checked on 20,000 data sequences (see Figure 5). It was found that accuracy of data sets was 38% according to paper algorithm. The accuracy of the improved algorithm was 43.5%. The next accuracy was found with reducing the input size to 5 and it was found out to be 44.5%. The next accuracy was found with input size 4 and it was 45.2%. Finally, accuracy was found out using the improved algorithm on the same data set but now with data clustered according to the length of the data sequences. In this case the accuracy was found to be 46.5%.

6. CONCLUSION AND FUTURE WORK

Web Log Mining is a very difficult task due to many intrinsic limitations of web logs and many uncontrollable sources of variation. Due to the high number of meaningful user profiles of a typical highly rated website, model or distance based method tend to make too strong and simplistic assumptions or to become excessively complex and slow. In this project an improved heuristic approach for page recommendation was designed and developed and compared with an existing heuristic approach. Improvement was done on the basis of the size of the input sequence and the clustering of data on the basis of length of input sequences. The overall accuracy increased by 8.5% to 9% for the improved approach. Future work is about studying estimation strategies for parameters, extending suggestions to more than one category and trying possible extensions to log files with a different structure and granularity.

REFERENCES

[1.] A.Nanopoulos, D. Katsaros and Y. Manolopoulos (2003), "A Data Mining Algorithm for Generalized Web Prefetching", IEEE Trans. On Knowl. and Data Eng. ,vol. 15, no. 5, pp. 1155-1169.

[2.] T. Palpanas and A. Mendelzon (1999) "Web Prefetching Using Partial Match Prediction", in Proceedings of the 4th International Web Caching Workshop, San Diego, California.

[3.] S. Jespersen, T. B. Pedersen and J. Thorhauge (2003) "Evaluating the markov assumption for web usage mining", in WIDM 03: Proceedings of the 5th ACM international workshop on Web information and data management , New Orleans, Louisiana, USA, November 7-8 2003.

[4.] Cadez, D. Heckerman, C. Meek, P. Smyth and S. White (2003) "Model based clustering and visualization of navigation patterns on a Web site", Data Mining and Knowledge Discovery, vol. 7, no. 4, pp.399-424.

[5.] M. Girolami and Ata Kab´an (2004) "Simplicial Mixtures of Markov Chains: Distributed Modelling of Dynamic User Profiles", Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference, David S. Touretzky, Sebastian Thrun, Lawrence K. Saul, Bernhard Schlkopf Eds, MIT Press, 2004, pp.9-6.

[6.] A.Maratea, A. Petrosino (2009) "An Heuristic Approach to Page Recommendation in Web Usage Mining", in 2009 Ninth International Conference on Intelligent Systems Design and Applications.

AUTHOR

Aditya Kumar received B.E. degree in Information Technology from Manipal Institute of Technology in 2014.