ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Volume 1, Issue 9, November 2013

Web Data High Quality Search - No User Profiling

P.Kavitha

Associate Professor, Dept. of IT, Bharath University, Chennai, TamilNadu, India.

ABSTRACT: The world wide web contains large amount of information. Gaining knowledge over the information and also over the web user is the only way to extract effective information while mining. This paper deals knowledge management of the user side. In order to extract the exact information that the user needs gaining knowledge is very important. Eliminating the use of user profiling here a new idea is proposed for discovering need of the user and also ordering the links in a rank based manner.

KEYWORDS: Search key, utility, web warehouse

1.INTRODUCTION

For a similar search key may have different views for different user. All their views may not be satisfied during the first time of search. But they can be satisfied by there second time of search done for the same word at the same instant. The main aim is to retrieve the most related pages for the user during the search process. For this the mind of the user has to be read. For this purpose the system has to be trained in such a way that according to the first click done by the user the system must be able to know the main target of the user for doing such a word search.

Let us consider an example of a word search. Let the word to be searched given by the user is “tree”. Some may look for the cultivation of the tree, some may want to know the various types of trees present in a particular area. A professional may need information regarding the various trees structure to organise the data and others may need the tree to just to have a look at its importance in growing. There fore we could conclude that a single word will not be helpful for learning the mind of the user.

The proposed algorithm NUP (No user Profile) gives the knowledge to know the necessity of the user during the search. This requires just the search key from the user to proceed with knowledge extraction, ranking and ordering. The second section is about knowledge management, third section is about the existing related work, fourth to tenth sections is about the proposed method, eleventh section ives conclusion and twelfth section provides the enhancement needed in the proposed work

II. KNOWLEDGE MANAGEMENT

The knowledge has to be gained in two segments while mining. The knowledge from the data that is present in the data ware house or the web warehouse has to be extracted. This knowledge is helpful for knowing about the contents of the information present on the whole. These information are obtained when the data or the link is created and are added to the warehouse. The knowledge gaining process is done even before the mining is done. These knowledge will not change unless the data are modified by the owner of the data.

The next stage of knowledge is extracted from the user while they are performing the mining process. This is dynamically performed. This information dynamically changes

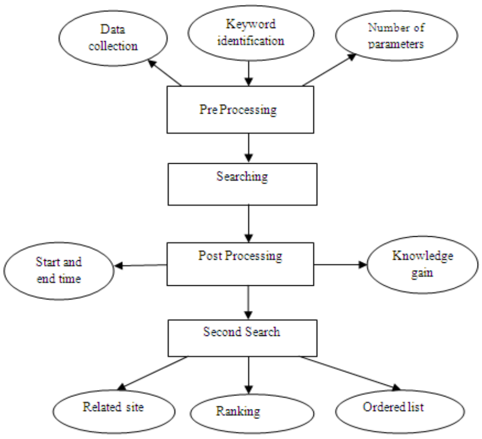

Fig.1.Web mining subtasks

The Fig. 2 gives the clear outline of the procedure done in the proposed system. The stepwise procedure is discussed in the forth coming sections

III. RELATED WORK

Web Usage Mining (WUM) is the process of extracting knowledge from Web user's access data by exploiting Data Mining technologies[1].

A partitioning method was one of the earliest clustering methods to be used in Web usage mining[1]. Incremental algorithm produces high quality clusters[2] the link the system would learn about the information the user is looking for. According to the example the links regarding the cultivation, the information about the types, the data about the structuring of information or data are all formed in to separate groups.

The next time user searches for the same search key he/she would be able to get the most related information in the first set of options. This would increase the effectiveness of the mining process.

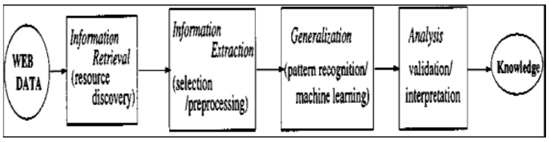
WEB DATA
Information Retrieval (resource discovery)
Information Extraction (selection /preprocessing)
Generalization (pattern recognition/ machine learning)
Analysis validation/ interpretation
Knowledge

Fig.2.Web mining subtasks

The overall web usage mining process can be divided into Four interdependent stages as shown in Fig 2 : Data collection, Pre Processing, Pattern Discovery and Pattern Analysis.

Almost all web sites have searching function, and the search engines under use are confronting the following troubles [6]

3.1. Over abundance: Most of the data on the web are of no interest to most people. In other words, even though there is a lot of data on the web, an individual query will retrieve only a very small subset of it.

3.2. Limited coverage: Search engines often provide results from a subset of the web pages. Because of the extreme size of the web, it is impossible to search the entire web at any time a query is submitted. Instead most search engines create indices that are updated periodically. When a query is posted, only the index is directly accessed.

3.3. Limited query: Most search engines access on simple keyword-based searching, retrieve or order pages based on popularity of pages.

3.4. Limited customization: Query results are determined only by query itself, often dependent on the background and knowledge of the user. The focus of this paper is to provide an overall view of how to use frequent pattern mining techniques for discovering different types of patterns in a web log database.

IV. ORDERING OF DATA

Just by the single word the intention of the user cannot be read. In this each set of links has the topic of the link as well as the hint about what the particular link is providing. The frequently used keywords would form the search keys of a particular data. This eliminates the commonly used words like the, is, was, about etc., So when ever the search matches the search keys immediately the listing of the link are done in the normal order. This would not give the preference to the intention of the user. Since there is no information about the user is provided. When all the links with a short line of description about the link are displayed, 95% of the user's would read the description before they follow through the link. By analysing the first click done on

V. ESSENTIAL FIELDS

There is said to be a table from where the mining of the information has to be done. This table would contain the following information.

5.1. Title of the web page

This is the main heading of the web page. This would be provided by the data provider.[3] Successful knowledge discovery requires a sufficient document collection.

5.2. Web address of the data

The data should be present in a particular address. This address is provided so that when ever a selection of the particular option is made the particular web address is linked.

5.3. Hint about the data

This gives a small description about the particular link. While a search is made the list of option along with the hint is listed. This hint should not exceed 100 characters.

5.4. Start time

This gains the information about what time the user has traversed through the link. The format of the time is retrieved as "hh;mm;ss".

5.5. End time

This gains the information about what time the user has moved out from the link. The format of the time is same as that of the start time.

5.6. Traversed path

The path traversed by the user is tracked. This field gets an entry only if the user moves through the path provided by the web page. The time duration spent on those paths are also identified.

5.7. Extracting parameters

This field retrieves the parameters provided by each web page on a particular topic. Those are the list of details provided by the developer. The total number of parameters provided in the site are entered.

If the links provided by the page are used it is assumed as the desired web page. It is assumed as the page is being read by the user before moved to the next link in the page.

By list of parameters provided it can find the efficiency of the web page. This specifies the details it can provide about the topic. The quality of the web page is determined by the quantity of information it provide.

The time is mainly used in order to know the time spent by the user. Since the user may go through the description and have made an attempt to traverse through the link. But the page would not have the desired information. Then user may attempt to move back in a few seconds or minutes. There fore a threshold is being set. When ever the time period exceeds the threshold value then it is decide by the system as the desired site by the user. Therefore in the next time of search done for the same search key in the next instant would give a different set of options. These options would be ordered in such a way that the most related sites are present in first few options.

VI. TEMPORARY BUFFERING

In the case of user profiling based web search the user profiles would have a separate table. Each time the user logs in and the activities done by the user is analysed until they log out. This information is saved in a separate memory space for further analysis. These information helps the system to know what the user is actually interested in. Therefore every time the user makes a log in, the options are provided according to the knowledge gained while traversed in the site previously in the same user session.

The main disadvantage regarding the user profile are

- The profiling would occupy large memory space. The user has to remember the username and
- password each time they log in.

[5]User profiles can be categorized into three groups: interviewing, semi-interviewing and no interviewing. Current web information gathering systems attempt to satisfy user requirements by capturing their information needs. For this purpose, user profiles are created for user background knowledge description. Generate a topic profile for the user based on the discovered categories contained in the sub-trees.

To avoid all these drawbacks in the proposed method there is no separate memory wasted for storing the user information. There is just a temporary storage done about the user who is doing the latest search operation. Once the user does the first search and spends time greater than the threshold value immediately the entry is made regarding the search key and the area the user is interested in. This information is helpful while the search is done for the same search key at the next instant.

Thus just a temporary storage is done. When the time elapses for the threshold set for this storage. Then immediately the information of the previous search becomes unavailable to the system. Only the access time and traversed path is being identified. Which are deleted after the session of search is over.

VII. SETTING THE THRESHOLD

The threshold value would be made to compare with the time duration that the user had spent after opening the link.

ET-ST > TV1

Where ET => End time

ST => Start time

TV1 => Threshold value The difference between the start time and the end time would provide the time spent by the user on the particular link. If the time spent is lesser than the threshold it is considered as the site not interested by the user. At this point of time there would not occur any reordering of the options instead the same set of options are shown with the title and the description. The total knowledge is retrieved only by the activities of the user.

Once the condition is satisfied then if the user moves back to look for more options then at that instant the options are re ordered and the sites falling under the group which the user is interested is displayed in the first few options. Even though the user was not provided with a wider range of options to his area of interest the second time of search at that instant would provide with large related site information.

Since there is just a temporary storage done here, there is a threshold set even for this storage to become unavailable. This threshold is compared with the end time [ET] and next search done for the same search key.

BST – ET < TV2

` Where BST => Back Search Time

ET => End time

TV2 => Threshold value 2

The difference between the end time and the back search time would give the time gap of the next search done for the same search key . If the threshold is larger than the time gap then the machine analyses that the next search for the same search key is done immediately. If the threshold value is smaller than the time gap then the next immediate search is not done by the same user. This is the analysis done by the system.

VIII. RANKING AND REORDERING

When it comes to the point of reordering, they are done only based on the rank that are being generated. The ranking is done only after identifying the user desired web page. The desired web page is identified by the previous equations that are discussed. Once that is identified all the related website is considered. The total number of parameters provided by the developer are given in the extracted parameter field. The site which has more number of parameters takes up the first rank

Once this ranking is performed they display the result from the first rank to the last rank during the next search operation done by the user. This helps to identify the site that provide wide range of details about a particular topic.

IX. TRAINING THE SYSTEM

Here the system is trained in such a way that it should analyse two factors.

9.1. Users utility of the web page

How much time user spends time on the particular link is analysed. This helps to find the most interesting web site. During this analysis the system would be able to find which is the area of interest of the user. With out getting the user profile information the area of interest is found. Once this area of interest for the particular search key is found and the list of search option is re ordered. The group of site falling under the area of interest of the user is given the first priority and the listing is re generated.

9.2. Users next idea of searching

This analysis is done in order to know whether the same user wants the listing for the search key. To analyse this point the time gap is considered that is between the first search completion and the next search. According to this information the system may conclude whether the user is the same user. Do they need still more options for the same search key. The time gap should be too small so that no other user is getting the search result. This information helps in making the temporary buffer to be unavailable to the system. Minimal wastage of memory space.

These two factors help the system to gain knowledge over the entire process done through out mining.

[4]The two techniques of knowledge discovery can be applied for fully meeting the analyst needs: association rules and sequential patterns. Association rules mining provides the end user with correlations among references to various pages and sequential patterns can be used to determine temporal relationships among pages. Furthermore, mining sequences with time constraints allows a more flexible handling of the visitor transactions.

A number of clustering approaches have been proposed, all of which use web server logs to generate a model of user actions that is then grouped with a clustering algorithm.

X. STEPS INVOLVED

The various steps done in these type of mining are

10.1. Pre Processing

These two activities are done before the search process

- The data that are present are sorted to the respective key words
- Then they are made into groups according to the area they fall in spite of similar search key.
- Total number of parameters provided by each link is identified

10.2. Searching

This is done by the user once the user types the search key They found match with the data search key words All the

- matched data are listed with their
- respective link address, title and the hint.

10.3. Post processing

After the search result is displayed the following activities are carried out

- The start time and the end time are noted.
- They compared with the threshold and the corresponding activity is carried out as given in the previous sections.

10.4. Second time search for same search key

When the second time of search done for the same key word the time threshold is analysed

Only the related sites are ext•racted

Ranking based on the number of parameters is done. •

Ordering the link according to the rank and displayed.

XI. CONCLUSION

The proposed method is much effective than that of the existing machine learning and the user profiling method. An efficient personalised search is made in the system is done by this method. The user's area of interest is given the highest priority. Which inturn provides a efficient set of options for the web search made by the user.

XII. FUTURE ENHANCEMENT

The proposed method is implemented with a limited amount of data and information. This can be implemented by using the data available in the World Wide Web. With increase in the set of information in the web this type of personalisation is very much essential.

REFERENCES

[1] Dipa Dixit And Jayant Gadge, A New Approach For Clustering Of Navigation Patterns Of Online Users, International Journal Of Engineering Science And Technology, Vol. 2(6), 1670-1676, 2010.

[2] V.V.R. Maheswara Rao, V. Valli Kumari And K.V.S.V.N Raju, An Advanced Optimal Web Intelligent Model For Mining Web User Usage Behavior Using Genetic Algorithm, Proc. Of Int. Conf. On Advances In Computer Science 2010.

[3] John M. Pierre, "Mining Knowledge From Text Collections Using Automatically Generated Metadata," Interwoven, Inc,2002.

[4] F. Masseglia, P. Poncelet, M. Teisseire, "Using Data Mining Techniques On Web Access Logs To Dynamically Improve Hypertext Structure," Acm Sigweb Newslettervolume 8 Issue 3, October.

[5] Xiaohui Tao, Yuefeng Li, And Ning Zhong, "A Personalized Ontology Model For Web Information Gathering" Ieee Transactions On Knowledge And Data Engineering, Vol. 23, No. 4, April 2011.

[6] Xiaohui Tao, Yuefeng Li, And Ning Zhong, "Constraint Based Frequent Pattern Mining For Generalized Query Templates From Web Log" International Journal Of Engineering, Science And Technology Vol. 2, No. 11, Pp. 17-33, 2010.