Students search interest model over an organisation based on web log data

AR. Sivakumaran*
Department of Information Technology,
Sri Shakthi Institute of Engineering and Technology,
Coimbatore, Tamil Nadu, India
Email: ars.sivakumaran@gmail.com
*Corresponding author

P. Marikkannu
Department of Information Technology,
Anna University Regional Centre,
Coimbatore, Tamil Nadu, India
Email: pmarikkannu@gmail.com

Abstract: Browsing information through web has become part and parcel of life. There is hardly any user who does not browse through web. Generally, any user when in search for content only looks at the top ten pages that get displayed in the web search. Therefore, it has been proposed that the information such as the link that is created between both visited and unvisited web pages along with the path that is chosen in the search query needs a novel technique to give the best performance. The operation feature matrix (OFM) is used as one of the novel functionalities that have been used for extracting the data from the web. The automatically identified user profile is the graph-based that is called as the modified page outlook (MPO) graph was proposed that involves a link between the visited and the unvisited web pages.

Keywords: web mining; web log data inter relation; intra-relation; total number of paths; total number of relations modified page outlook; MOP.

Reference to this paper should be made as follows: Sivakumaran, AR. and Marikkannu, P. (2018) 'Students search interest model over an organisation based on web log data', Int. J. Business Intelligence and Data Mining, Vol. 13, Nos. 1/2/3, pp.26–39.

Biographical notes: AR. Sivakumaran received his MTech degree in Computer Science and Engineering from Motilal Nehru National Institute of Technology (MNNIT), Allahabad, UP, India. He is a Research Scholar at Anna University Chennai, Tamil Nadu, India. He is working as an Assistant Professor in the Department of Information Technology, Sri Shakthi Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India. He has published many research articles in various journals. His research area is web mining. He is a Life Member of ISTE.
1 Introduction

On the basis of web resources, for any application of traditional data mining technique, the web mining is described. This has helped for future development considering the specific structures that are available in the web mining. The web resources involve the actual website along with the links that are connected to the various sites. This also contains the path that is chosen by the various users in order to reach the desired websites. The useful knowledge that has been derived from these is often known as the web usage mining. A special challenge includes the details given by the users and also the data that is obtained from the websites. The output that is desired is basically the understanding obtained by the behaviour of the user and the web server logs. Some of the online information search is based upon the shopping, places, foods, online learning about various things. It is of course difficult to search what is to be searched. In order to achieve this goal, the semantic web approaches has been chosen as a promising path that leads to a computational and formal aspects. The research approach has an interdisciplinary subject. The view and thoughts of the lay man should be kept in mind while designing the interface that could result in perfect data mining. The mining system of the users has to be visualised and designed for bring out the correct and nearest search. The dynamism of the domain plays a vital role as there is an often change in the availability of the data and the behaviour of the user. There are various data mining algorithms that are present for the analysing the given dataset. There is also a possibility of change in the content of the data that has been considered or the data might have been updated. The reason behind this can be a long storage of data. As the data is been updated, it becomes necessary to update the pattern. This is done easily because most of the mining techniques are the incremental mining techniques. Figure 1 shows the analysis of the web mining. It is clearly seen that the web mining is classified as the web structure mining, web usage mining and the web content mining. They are further classified into number of various sub division. The phases that are involved in the web mining are pre-processing session reconstruction heuristics; pattern discovery, pattern analysis application. The entities that are involved are raw web log, user session file, rules and patterns and knowledge that are interesting.
2 Related work

Personalisation of web search for the interest of the user includes modification of search algorithm and modification of the query. This is done for the fine search. There are large dataset that exists in a web or internet from where the data are retrieved. The problems are solved by the algorithms that have been discussed (Bansal, 2014; Rajaraman and Ullman, 2012). The web mining temporal explicit and the semantic relations between two or more entities is done using the web search engine. Annotate and mining relation with the concise, temporal and structural information has been done (Xu et al., 2014). A novel technique has been proposed for an unsupervised, independent of language and models of generative machine learning. The probabilistic framework has been discussed and demonstrated with the novel model. With the help of the designed guidelines, the technique has been implemented. This also depends on the interest of the designer team. The results that are obtained show that the improvement is needed in the keywords by the webmaster team (Moens and Vulié, 2014).

A novel approach has been discussed that deals with the collection for the processing of the data that are initiated by the web user. The movements of the web page are the ocular. The movement by the web pages are tracked by the eye tracker tool. Depending up on the eye position of the web user, the data is pinpointed. This indicates that the website key objects are taken acre in the database as it is often viewed by the web users. The website keywords are extracted from the web page by using web mining algorithm (Carmona et al., 2012). The ground work is the data pre-processing work that is done in web mining. There are many web servers related to it (Velásquez, 2013). There is an automatic discovery of the patterns that are present in one or more servers. The pre-processing process involves extraction of unwanted information from the web logs. This is often known as web cleaning. The algorithms for filtering are also used to remove the attributes that are not of any interests. The number of users along with the various
Students search interest model over an organisation based on web log data

sessions has been tracked (Losarwar and Joshi, 2012). The practicalities and the effectiveness have been examined that are present in the web mining. The web mining focuses on quantity, accessibility, completeness, accuracy and flexibility. The information from the website is also self-reported. The motivations of the company might vary depending upon the size and nature of the organisation (Gök et al., 2015).

As the information in the web increase day by day, it is a challenge to the search engine. The search lacks in efficiency as it provides somewhat relevant information (Johnson and Kumar, 2013). The log repositories that are present in the web server are used to keep the records of the knowledge of the source from where the data is being recovered. The pattern of the web usage is defined as the process of the browsing pattern of the web browsers. This is analysed by the behaviour of the navigation (Shah and Jain, 2015). The average time that is spent on the web to search the page indicates the working of the web mining. The web user logs contains all about the behaviour and the need of the user. The accurate web page then displayed (Patel and Parmar, 2014). The quantitative analysis on the web is based on the web mining and the webometrics. There are different methods and applications for the two mentioned above. A key difference between the fields of webometrics and web mining is that the former focuses on the exploratory studies and the later depends on the studies that are focused on the algorithms and the developed methods. The structure is more focused by the webometrics. In the presence of huge data or big data, the combination of webometrics and web mining are equally important (Lorentzen, 2014). The knowledge of extracting the data is known as the information extraction. The meaningful data is considered. This is done automatically. The services are provided in order to retrieve the data after a fine query from the internet. When huge information is extracted, it is known as the web mining. The input plays the most important part. Depending on the input, the data is retrieved from the internet. A complete survey is done based on the different strategies for web mining (Joy, 2015). A detailed study about the web usage has been done. Amongst the various algorithms present, the tree algorithm is used. The tree algorithm is used for an improved frequent pattern (Johnson and Gupta, 2012; Mahajan et al., 2014). The user accessibility pattern during the mining of the log files that are present are known as the usage of web mining (Gupta et al., 2014).

The discovery knowledge obtained from World Wide Web (WWW). The general data that are obtained are collected and put forth with the help of various methodologies that are used in both web mining and data mining. The web structure mining has been explained. In this the various patterns are being extracted from the hyperlinks which can be both inter-hyperlinks and intra-hyperlinks. It is also very important to correlate the web mining with the security in the network. The method of using both the network security along with the web mining is often referred as the hybrid approach. By the use of such an approach, both the cost and the time are reduced (Chuchra et al., 2013). The book deals with the mining for Gems of Information. The mining of the text presented in the web is one of the most common mining used. The ways of text mining along with its various applications has been discussed in detail. Several article and proposal has been given on text mining. The possible application has been explored in the world of production that also includes advanced website warehouse production (Sirmakessis, 2012). The retrieval of the needed data from the web plays a vital role. There is nothing better in retrieving a data in the best possible method and best possible time. There are numerous surveys that have been done on the web mining. By setting up a set of specification, the user on the
web makes it easier to get the desired data. The various clustering methods along with the static report indicating the number of times a website is browsed and the topic often search by the users are clearly observed (Jones and Gupta, 2015). An efficient method for pre-processing of web log has been proposed. This is done towards the web usage mining (Jacobson et al., 2014). One of the important sources of information is the web log file that is used to get the insufficient mining result. Web server log leads to the data processing which in turn leads to the pattern detection and pattern analysis and finally leads to knowledge discovery (Upadhyay and Purswani, 2013).

3 Proposed method: operation feature matrix

The operation feature matrix (OFM) is used as one of the functionalities for extracting the data from the web. The OFM matrix consists of the user intangible index (UII) which provides a link between the relevant pages and the search query.

**Figure 2** Operation feature matrix

|      | \(IWP_1\) | \(IWP_2\) | \(\ldots\) |
|------|-----------|-----------|----------------
| \(SQ_1\) | \((W_{SQ_1,IWP_1})\) | \((W_{SQ_1,IWP_2})\) | \(\ldots\) (\(W_{SQ_1,IWP_m}\)) |
| \(SQ_2\) | \((W_{SQ_2,IWP_1})\) | \((W_{SQ_2,IWP_2})\) | \(\ldots\) (\(W_{SQ_2,IWP_m}\)) |
| \(\cdot\) | \(\cdot\) | \(\cdot\) | \(\ldots\) (\(\cdot\)) |
| \(\cdot\) | \(\cdot\) | \(\cdot\) | \(\ldots\) (\(\cdot\)) |
| \(\cdot\) | \(\cdot\) | \(\cdot\) | \(\ldots\) (\(\cdot\)) |
| \(SQ_m\) | \((W_{SQ_m,IWP_1})\) | \((W_{SQ_m,IWP_2})\) | \(\ldots\) (\(W_{SQ_m,IWP_m}\)) |

The UII is represented by the weighted function which is given as UII = \(f(W_{sq,im})\). Where the \(W\) is represented as the total sum of the sequential query for it is applicable page weight. Form this it has been concluded that there are three components present in the UII. The components include time components, frequency component and the number of actions done. There are different values that are present in the UII. The highest value of UII indicates the interest of the user at its highest. The OFM is given in Figure 2. Each cell in the OFM is filled up with the UII value for each \(SQ_a\) and is related to all the web pages that the users refer to. The parameters that are considered are inter-relation between paths (Inter-RP), intra-relation, overall number of paths (ONOP), overall number of relations (ONOR), and overall number of class (ONOC).
3.1 Algorithm for UII

For calculating the UII, the following steps are considered. This is done for \( a^{\text{th}} \) page and \( b^{\text{th}} \) session.

Step 1 Consider the search query as \( SQ_a \).

Step 2 The number of web pages that has been viewed is considered as operation \( O_a \) for the search query \( SQ_a \).

Step 3 The weight is calculated by the product of document frequency to the term frequency. The total number of web pages over which an operation is done is represented by \( [O_a = WP_1, WP_2, WP_3, \ldots WP_n] \) and the weight of the internet web pages are represented as \( [IWP_a = IWP_1, IWP_2, IWP_3, \ldots, IWP_n] \).

Step 4 The search query weight is calculated by the following formula

\[
w(SQ_a) = \left[ \frac{1}{n} \sum_{d=1}^{n} F_b(SQ_a), \frac{1}{n} \sum_{d=1}^{n} T_b(SQ_a), \frac{1}{n} \sum_{d=1}^{n} A_b \right]
\]

Step 5 The internet web page is calculated by the formula.

\[
w(IWP_a) = \left[ \frac{1}{n} \sum_{d=1}^{n} F_b(IWP_a), \frac{1}{n} \sum_{d=1}^{n} T_b(IWP_a), \frac{1}{n} \sum_{d=1}^{n} A_b \right]
\]

Step 6 Ultimately, the UII is calculated as \( UII = f(WSQ - IWP) \). The UII derived has values that are present in the OFM.

Step 7 End the program.

4 Mathematical model

In this section, the mathematical model is concentrated for the OFM proposed methodology. The art of interpreting the difficulties from any application area to tractable mathematical formulations is often referred as mathematical modelling. As seen in the proposed methodology the UII is represented by the weighted function which is given as

\[
UII = f(W_{sq-\text{tw}})
\]  

(1)

Similarly, the search query weight and the internet web page are calculated by the following formula respectively.

\[
w(SQ_a) = \left[ \frac{1}{n} \sum_{d=1}^{n} F_b(SQ_a), \frac{1}{n} \sum_{d=1}^{n} T_b(SQ_a), \frac{1}{n} \sum_{d=1}^{n} A_b \right]
\]  

(2)

\[
w(IWP_a) = \left[ \frac{1}{n} \sum_{d=1}^{n} F_b(IWP_a), \frac{1}{n} \sum_{d=1}^{n} T_b(IWP_a), \frac{1}{n} \sum_{d=1}^{n} A_b \right]
\]  

(3)

where \( SQ_a \) is the number search queries, \( IWP_a \) is the number of internet web pages. \( F_b \) is the frequency of the page visited and the \( T_a \) is the \( a^{\text{th}} \) time at which the page was browsed.

The weight of the concept which is more important for the path weight computation deals with all the concepts that are taken from the graph. For instance, EDUCATIONAL
ORGANISATION gives more meaning that of the SCHOOLS, COLLEGES, GIRLS and BOYS. The weight of the concept is given as

$$WC_i = \frac{LC_i}{L}$$

(4)

where $LC_i$ is taken as the $i^{th}$ concept in the MPO hierarchy and the $L$ refers to the total length of the hierarchy.

Similarly, the weight of the path length depends purely on the context. Some of the user prefer shortest path, whereas some do not. Hence, it often becomes necessary to reduce the path length. The weight of the length of the path is given as

$$WPL_i = \frac{1}{C}$$

(5)

where $WPL_i$ is the weight of the length of the path and $C$ is determined as the elements that are present on the path. The elements may refer to both concept as well as the relation between them.

The other measures are Inter-RP, intra-relation, ONOP, ONOR, and ONOC. The average number of relations that is considered per concept is $\mu$ and the average path that has been considered is $\rho$.

$$\mu = \frac{ONOR}{ONOC}$$

(6)

where $\mu$ is the average number of relations per concept, ONOP is the ONOP chosen and ONOC is the ONOC that are present.

$$\rho = \frac{ONOP}{ONOC}$$

(7)

where $\rho$ is the average number of paths per concept, ONOR is the ONOR and ONOC is the ONOC.

5 Illustration taken of an organisation

The set of various process that are involved in the creating a best graph-based that are present in the knowledge layer. The layer that collects the information of the data from the data layer is called as the knowledge layer. The automatically identified user profile is the graph-based that is called as the modified page outlook (MPO) graph that involves a link between the visited and the unvisited web pages. The proposed MPO graph shows the shortest distance to get the details wanted by the users. The user behaviour examination is a significant portion of personalised web search since it contains particulars concerning user navigation through the web. Such user behaviours are responsible for identification of shortest search path or a path that gives clues to the appropriate evidence.
Figure 3 comprises of various models such as a user search queries, existing search engine, and different types of browsers, WWW, user behaviours, pre-processor, server and MPO graph. With the help of the existing search engine, the WWW is browsed. The server is directly connected to pre-processor, MPO graph, path weight computation, modified page recommendation and user behaviour. Path weight computation various components such as the weight of the concept which is represented by $W_C$, weight of the path length which is represented by $W_{PL}$, and weight of the personalised, which is represented by $W_P$.

6 construction of MPO graph’s for an organisation

Considering sample such as the engineering, higher secondary, arts, medical, etc., these are obtained under the page of the organisation. The extracted information from the web regarding the organisation is given in a specific domain. The index words such as the Arts, Science, secondary, higher secondary are mapped with the certain concept in the domain. Once the mapping process is over, it has been found that the terms such as {SCHOOL, COMPUTER SCIENCE, MATHS BIOLOGY} are related to the words such as {GIRLS AND BOYS}. This means that they are semantically related to the input index. Finally the entire set produces is equal to that of the following {SCHOOL, COMPUTER SCIENCE, MATHS BIOLOGY, GIRLS, BOYS, ENGINEERING AND MEDICAL} that are used to search the related web pages, this includes web pages that are included and excluded. Approximately, there are ten pages that come in the search of it. This is very much known as the automatic divisive method for a group.
The index words are given a vector respectively. Also their weights are allotted depending upon the pages that are retrieved. The computation of the cosine is done for both the words that are present in the visited pages and also in the unvisited pages. In order to provide a neat relation between a MPO pages has been proposed and implemented. This is done using the Java. The words that are taken as the indexed words are represented in the form of entity. Their relation is represented in the form of edges. Consider the situation when the user clicks onto the entity. The pages that are related to the indexed words are displayed and the path is also noted. As known the indexed pages are for both visited and unvisited page. Both of these are highlighted with different colours to know there differentiation. Those pages that are not visited by the web users are therefore identified and the when users are recommended with it. Sometimes those web pages are the ones that give the user their best results. Thus, the systems proposed by us are created in such a way that they take care of those when pages that are not much concentrated upon and give it to the web users.

The sample MPO graph that is created for the educational organisation is given in Figure 4. Here it has been noted that the educational organisation is further divided into two depending upon the age of the students. If the students are above 18 then they are to be in college and those below 18 are in the school. The highlights are generated wherever the educational organisation is described such as the {SCHOOL, COLLEGE}. They are further divided into subclasses such as {KINDERGARDER, MIDDLE CLASS, ARTS, COMPUTER SCIENCE, MATHS BIOLOGY, MEDICAL, ENGINEERING}. The subclasses of the SCHOOLS and COLLEGES are grouped and given a tag. The path length weight is also accountable for personalisation of the users.
While considering the concepts in the graph, there are various hierarchy under the educational organisation. The users can browse any of the sections under the educational organisation. Hence when the user types the word ‘educational organisation’, all the related pages have to pop up. There are various sub categories that are present under each category by itself.

7 Results and discussions

For each class and its subclasses the weight of the concept varies. Considering the equation (4) the following has been obtained.

\[
\begin{align*}
\text{EDUCATIONAL ORGANISATION} & = WC_1 = LC_1/L = 1/5 = 0.2 \\
\text{SCHOOL} & = WLC_2 = LC_2/L = 2/5 = 0.4 \\
\text{COLLEGE} & = WLC_3 = LC_3/L = 2/5 = 0.4 \\
\text{GIRLS IN SCHOOL} & = WLC_4 = LC_4/L = 4/5 = 0.8 \\
\text{GIRLS IN COLLEGE} & = WLC_5 = LC_5/L = 4/5 = 0.8
\end{align*}
\]
Similarly, the weight of the length of the path is also calculated. Considering there are three paths that are present. Consider, the user is on the page of EDUCATIONAL ORGANISATION from Figure 5 and has found the link of COMPUTER SCIENCE from this page. Now this can be the direct answer to the query of the user. The second path may be through the HIGHER SECONDARY and the third path may be through the page SCHOOLS. Therefore total three paths are available. The weight of the length of the path may be computed by using the equation (5).

\[
\text{Direct path } = \frac{WPL_1}{C_1} = \frac{1}{1} = 1
\]

\[
\text{Longer path } = \frac{WPL_2}{C_2} = \frac{2}{5} = 0.2
\]

\[
\text{Longest path } = \frac{WPL_3}{C_3} = \frac{3}{5} = 0.6
\]

It has been proved that \( WPL_1 > WPL_2 \). Hence, the longer the paths are they are rated with low rank.

<table>
<thead>
<tr>
<th>Source page</th>
<th>Destination page</th>
<th>Path travelled</th>
<th>Length of the path</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDUCATIONAL ORGANISATION</td>
<td>Computer science</td>
<td>C1 → C4</td>
<td>Shortest path</td>
</tr>
<tr>
<td>HIGHER SECONDARY</td>
<td>Computer science</td>
<td>C1 → C3 → C4</td>
<td>Long path</td>
</tr>
<tr>
<td>SCHOOL</td>
<td>Computer science</td>
<td>C1 → C2 → C3 → C4</td>
<td>Longest path</td>
</tr>
</tbody>
</table>

Figure 6 shows all the primitive measures versus the number of documents. These kinds of relation between pages are collected and the semantic relations between are considered. The MPO graph that is developed personalised web search is thus made different and unique by itself when compared to the existing approaches. There are many measures that are taken. They are Inter-RP, Intra-Relation, ONOP, ONOR, and ONOC. From equations (6) and (7), the average relations and the average paths are calculated. When the average path is greater than unity, then the path obtained becomes a tree.

Figure 7 shows all the primitive measures versus the number of documents. However, it has been observed that the relation seems to be growing faster than the concepts.
Similarly, when the paths are considered it is more or less the same. The only thing is that there is more exponential growth in the path than the concept. Therefore as the number of documents increases, there is more connection in the graph and the paths get evolved faster. The average number of paths crosses unity when the total number of document crosses 100. The average relation that is obtained when there is a growth in the rate of connections that are being established that goes more than the 100.

Figure 7  Variation of the different the primitive measures (see online version for colours)

Figure 8  Search path distribution (see online version for colours)

Figure 8 shows the Search path distribution. By evaluation, the total number of search path that has been detected is around 156. The numbers of search path involving 2 pages are 34, the numbers of search paths involving 3 to 5 pages are 45 and the number of search paths involving more than 5 pages is 67. Figure 9 shows the number of search with different levels. As the level increases from document 1 to document 4, the ranking decreases.
Figure 9  Number of search with different levels (see online version for colours)

![Graph showing number of searches with different levels.]

Figure 10  User action distribution obtained at search path (see online version for colours)

![Graph showing user action distribution.]

Figure 10 shows the user action distribution obtained at search path. Few of the searched pages are the ones that were desired by the user. Hence, it is clearly seen how many web pages were obtained that were related to the search. Was it two pages or above two pages or above five pages.

8 Conclusions

The value of UII has helped us in establishing the link in the search query. The identification of the unvisited web pages is also considered. A novel technique has also been proposed for both visited and unvisited web pages. The development of the MPO graph is relevant to the web page needed by the user. Also, the shortest path has been considered. The shortest path gives the highest rank. By choosing a longest path, there is lots delay, much use of the resource. By performing a search analysis for identifying user having similar interest also makes the search difficult, since it has to be more precise. The future enhancement may be grouping of users with similar interest to make the search simpler and time saving.
Students search interest model over an organisation based on web log data

References


