COSC 6337.001: Data Mining

Department of computing sciences

Final Report

Project Name: Improving User Search Experience by Calculating User Intention

Under the supervision of Professor - Dr. Mamta Yadav

Team Members:

Vishnuvardhan Rama – A04256498

Yashaswini Yachamaneni – A04257757

Vujhini Nihal – A04257591

INDEX

S.No.	Content	Page No.
1.	Abstract	
2.	Aims and objectives	
3.	Research Problem	
4.	Cons of Existing Method and Solution	
5.	RELATED WORK	
6.	Methods Used	
7.	PROPOSED SOLUTION	
8.	RESULTS	
9.	CONCLUSION	
10.	FUTURE WORK	
11.	REFERENCES	

Abstract: In Real web world, there is a large amount of web pages which are providing information to the knowledge seekers. This huge collection of the data needs to be processed based on user requirements like search, recommendations etc and many applications has achieved that. But in Web Usage Mining (WUM) personalized web search results is always a trending topic, in this search engine will return search results of web pages based on the user personal data like user log, web content, location etc. Most of the search engines following few concepts of Personalized Web Search (PWS) like keyword suggestions, coloring technique etc. But these are not satisfying the user needs, here in our application we are proposing PWS concept by taking concepts of Web Content and User Log. In our application we predict the search results of web pages based on the user log and content mining. Here for user log we are taking the web browsing history of the user and we are using Term Frequency Inverse Document Frequency for content mining. In this we are achieving a dynamic way of user browsing web pages prediction.

1. INTRODUCTION

In web world, lots of websites are available to share the huge or vast amount of information to the world. Some studies predicted that zettabytes of data we used in 2007, and it is reached 6 zettabytes in 2014 [2]. We can get that information which are websites are providing by search engines, whenever we enter keywords of a topic or data search engine finds the relevant webpage's and return to user, for this process search engines create and update of search index of websites in different ways [1], those indexing algorithms are based on the individual search engines. In current search engines day by day improves the search performance because of rapid growth of the web data, and it is also focuses on web media data like colors, images, videos etc and social media data. By using the web index which are always focus on the web URL's and meta-data will not improve the user's search performance because of the web page data getting by a fixed URL's and displaying information can be dynamically changes. So, we can't estimate the ranking of the webpage's using search index correctly. The ranking of the web pages will be updated based on many concepts like Page Ranking algorithms [6] etc, but these concepts will help to webpage's ranking but not for calculation of user intention. Calculations of user intention by taking properties of his/her own for search results, it is a concept of Personalized Web Search (PWS). Most of the search engines are not considering about the Personalized Web Search, PWS concepts are useful to calculate the user intentions and tailored to user needs. As system need to collect user information and deep data of the web pages for analyzing and produced results to satisfy the user intention to raise a query. Generally, most of the related studies explain that PWS is categorizing in two types, those are *click-based* and *profile-based* [7]. In *click-based,* it is taking the data of user clicks from the search results and history or user log. Based on this, few topics were proposed like coloring concepts, keyword suggestions etc [4]. In profile-based, it is taking the information of user profile like query history, bookmarks, location etc [5]. But in our model, we proposed an architecture which will divide in to two parts those are Web Content Mining and Web Usage Mining. In Web Content Mining (WCM), it is challenging to locate the deep information from the web pages and classify the web pages according to the topic by using text classification algorithm called TF-IDF, based on the classification we filter the web URL's which are relevant. In web usage mining, we monitor the user clicks of the web search results and navigation URLs from the root pages. By this we can estimate the importance of the web pages which are regularly using based on the content.

Information is extracted automatically with the use of web data mining techniques. Web mining contains three different types as

1. Web content mining,
2. Structure mining, and
3. Web usage mining.

Web content mining will help to extract useful information from the web pages such as audio, video, image, and so on. Web structure mining will help to discover the useful information from the hyperlink and analyze both in-links and out-links. Also, it helps to score the web pages. Web usage mining help to analyze the log activities to determine the user behavior and usage patterns.

Nowadays, most people are using the internet and several websites are viewed by billions of users. A massive amount of information or data will be created once the user accesses the network. Then, it will store it in weblogs. In the files of weblogs, the web page will be recorded once the user searches for the same types of pages. This kind of sequence will help to track and analyze the access behavior of users. With this information, it is easy to predict the users of web pages who might visit next based on the access sequences that are previously visited. This will minimize the search time for users.

2. Aims and objectives

In this we discuss about the main aims and objectives of the application. The aims and objectives of the project is mentioned following.

This work proposed for enhancing the user experience in web search results by calculating the user intention means tracking the preferred internal web URLs from the search engine results page only.

It can be mine data of the webpage content data dynamically and user web usage in the browsers.

We can classify the data of the webpage's by using the TF-IDF algorithm.

User can save the computation time for finding the relevant we URL's internally from the root pages of the search engine returned results.

3. Research Problem

In recent days, technologies are increasing day by day with advanced feature which makes people's life easier and helps the organization with better process. Apart from this, due to emerging technologies with feature advancements, there is a massive amount of unprocessed data or information. This will take more time to view or extract the required information or data. Also, the web has been transformed into one of the primary tools, so it is difficult to analyze and track the access patterns of users. In addition to this, most of the time, web users will face issues when the information and data get overloaded. The main reason is due to a massive number of resources or information on the web. So, it is important to determine the behavior of web users on web site by enhancing the user experiences. It is not possible to capture the activities and user intention using existing approaches.

4. Cons of Existing Method and Solution

The existing method of semantically enhanced clustering is used to determine the user behavior when users access the web page. The clustering success is highly correlated by measuring the similarities over the navigation sequences. It may be difficult to manage and store the entire previous session of user access or visit due to a large amount of data. This will take more prediction time by comparing the entire session when the cluster has several sessions. Sometimes, the use of clustering may lead to severe challenges because of its lack of ability in recovering the data corruption from the database.

To overcome the issues, here going to propose a different kind of approach based on web mining which helps to capture and track information in an effective way and stores the entire collected web page accessed by web users. This kind of process will help to make a better prediction. The usage of web mining will obtain a user profile through an application.

5. RELATED WORK

In literature survey few concepts were proposed for personalized prediction of user web usage intention and web search. *Kilic et al.* proposed concept of prediction of web pages recommendation by clustering techniques. They focused on the session series of the web pages and its session timings. Based on these concepts they applied KNN algorithms for prediction of the recommendation of the websites based on the visible website. *Shou et al.* introduced UPS architecture which means User customizable Privacy-preserving Search, this is for prediction of the personalized web search results based on the user taxonomy repository. This architecture is designed in two ways, those are client profile and networking. Networking will re-rank web results based on the user taxonomy repository.

Dhandi et al. proposed a framework of Web Usage Mining (WUM) which is used to monitor the behavior of users while searching and visiting web pages. By using this we can analyze the discovery patterns of the user search. In this framework there is a important concept for WUM is Pattern Discovery, for this they were used association rule mining algorithm like Apriori. In this research they were tried to provide a clear understanding of the data preparation and knowledge discovery process in the web search personalization concept.

6. Methods Used

6.1. Web Mining

Web mining is the process of extracting valuable and useful information from the web server which contains a large repository of data. It is the pattern recognition technique in which the techniques are applied to find the pattern through exploration, identification, and deployment of the pattern. It is one of the functions of data mining. It can be classified into three types of content mining, usage mining, and structure mining. Improving the structure and organization of data on the internet can help the process of extracting the data. The required data is generated for the search based on the web server logs created based on the user search and needs. For web mining the data are collected from the database, pre-processed, the clustered to analyze the pattern for the data discovery using pattern analyzing tools. In this paper, the general concept of web data mining is discussed, and its application is explained in detail.

6.2. Web Page Prediction

Web logs are the data collected based on the information of the different user's search information on different web servers. The data web usage mining is followed by the discovering of data, analyzing of those data to find the pattern for extracting data and visualization to prepare the pattern of data analysis. This tool is used to analyze the behavior of the user search and their interest to predict the future traffic pattern for that website. The table and graphics along with the plot matrix are used to express the web log data. The merits and demerits of web log mining are listed below

Merits

It helps in achieving desired results for the user search.

The analysis can be easily performed on a large data set with high accuracy level.

Demerits

Different analyzing tools and algorithms are required based on the manners of the data so selecting tools required a skilled person to work on it

6.3. Behavior Prediction Using Session Analysis

The web log files created are used to extract the session of the user using their IP address. The pre-processing is applied to the web log data to remove the unwanted information to identify the user and their session data. In this paper, the algorithm for cleaning data, identifying users, and identifying sessions are explained. This analysis of the weblog will provide various information regarding usage of the website, website traffic, and problems that occurred during the usage of the website at regular intervals of time. This analysis is based on the frequency of users visiting the web pages. These data are useful to regulate the website traffic and organizing the website.

This analysis will help in discovering the search and navigation pattern of the users. This kind of behavioral analysis is used to understand the behavior and buying pattern of the users. Though this analysis is helpful in improving the website administration it may expose the privacy of the user. The personal information received from this site is misused by unauthorized hackers

6.4 Web structure mining:

This is the technique of using graph theory to examine the nodes and connection structure of a website. This can be used to determine two things: the structure of a website in terms of how it connects to other websites, as well as the document structure of the website itself, or the connections between each page.

6.5. A statistical techniqueTF-IDF (term frequency-inverse document frequency):

Assesses how pertinent a word is to a document within a collection of documents.

A word's frequency in a document and its inverse document frequency over a group of documents are multiplied to achieve this.

It is highly useful for scoring words in machine learning algorithms for Natural Language Processing and has a wide range of applications, with automated text analysis being the most essential one (NLP).

For document search or web search and information retrieval, TF-IDF was developed. It operates by escalating according to the frequency with which a word appears in a document but is counterbalanced by the quantity of documents in which the word appears. Therefore, even though they may appear frequently, words like this, what, and if rank low because they aren't important to that document.

6.6. Crawling Internal URL's

To index the content of websites so they may be searched, website crawling is the automated retrieval of online pages by a software process. The crawler examines a page's content for links leading to more pages that it can fetch and index.

7. PROPOSED SOLUTION

Personalized Web Search

In this chapter we discuss about the main flow and architecture of the application. Our proposed system we can divide in to two processes. First one is Client side which means user side flow of the application and another in Server side which is internal process. In fig. 1, we describe our project architecture. Client-side approach is a end user process, whenever end user search a query with some keywords this process will start. But for this approach application requires a predefined data, for achieving that data we require second approach. In second approach means server-side approach we build the internal links for the root pages with classification algorithm.

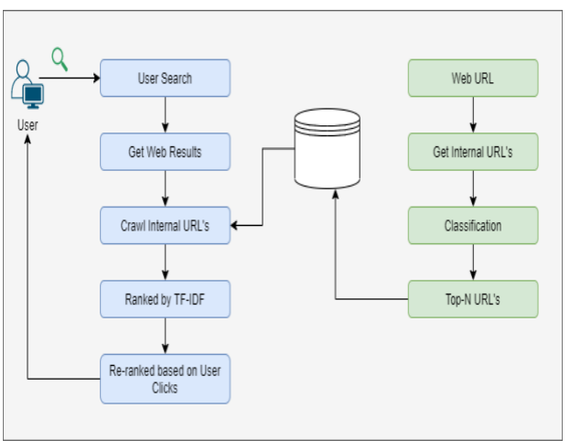

Fig 1. System Architecture

7.1 Server-Side Approach

In this process first we will select a URL for crawling internal URLs of the selected webpage. Admin will search and get few URL’s by using the existing search engines like google, Bing etc. Collection of webpages for a different topic with the help of the search engines is to avoid the huge collection of the website visitation. We can use advanced coding APIs in Java and Python for crawling the search engine web pages.

Algorithm 1: Crawling Text

Input: URL

Output: Content

```
site = Clean(URL)
if true:
        WebPage = API.connect(site).get();
        text = WebPage.body().text();
end if
return text
```

In next process, after we select a webpage system will crawl internal or sub webpages of the URL. If need to calculate the top best sub webpage's first, we need to get the internal URL's list from the main web page. Based on the Meta and Html tags we can get the internal URLs of the main URL. We collect these sub-URLs for ranking process. For ranking process, we need collect the webpage content from the individual sub-URL. For collection of the text data from the webpage we use Crawling Text algorithm.

Based on the content we can mine the webpage data that is how much relevant to the topic. This process is nothing but Web Content Mining. For this mining we are taking best classification algorithm for content classification called TF-IDF. In the algorithm 2, explained clearly about the TF-IDF algorithm. Based on the TF-IDF scores system will store the top 10 URLs in the database. This process we are doing server side because of there is a time taking process for calculation of the TF-IDF scores and ranking process for sub-URLs at user side while searching. So, we are process this procedure at server side for selected URL's.

Algorithm 2: TF-IDF

Input: Document, keywords

Output: Score

let,

TF = Total times Words Appear in the content

W = Total words in Document

TST = Total No. of Statements

M = Matched Statements

$$\text{Score} = \frac{\text{TF}}{\text{W}} * \log \frac{\text{TST}}{\text{M}}$$

7.2 Client-Side Approach

In this process, user can search the webpage's what s/he wants by providing the keywords. According to given keywords, system will get the webpages from the existing search engines like google, Bing etc., because of to avoid the huge collection of the website visitation. After getting search web results from the search engines, system will match with stored URL's which are processed with classification algorithm. If any web URL is matched, then system will display crawled URL's along with root URL's. In this situation Web Usage Mining will perform, while displaying the crawled URL's system will

collect the history of the browser, and system will re-rank the crawled URL's according to the frequency of the webpage usage.

Application based Prediction of Web Page Access

1. Enhance user navigation by caching and pre-fetching the web page access
2. Enhance the design of web sites
3. E – commerce web sites
4. Recommending dynamic web pages
5. Personalizing web pages for user's groups or individual user
6. Generation of dynamic hyperlink

8. RESULTS

In our experimental results of Improving User Search Experience by Calculating User Intention, we calculated TF-IDF scores of webpages. In table 1, we demonstrated our results. We design and develop our implementation in the Windows operating system having 8 GB Ram and 1 TB HDD. We developed a web application using Django framework. We can see some web page results executed in the local server for demonstrating our system.

In this example we have taken keywords 'java tutorial', for this we have taken root URL called 'www.javatpoint.com/', in following table list out the TF-IDF scores of the internal links of top 10.

Sno	Internal URL's	TF-IDF Score
1	https://training.javatpoint.com/java-training.jsp	0.0478
2	https://www.javatpoint.com/java-tutorial	0.04
3	https://www.javatpoint.com	0.0323
4	https://www.javatpoint.com/hadoop-tutorial	0.0323
5	https://www.javatpoint.com/selenium-tutorial	0.0257
6	https://www.javatpoint.com/compiler-tutorial	0.0239
7	https://www.javatpoint.com/os-tutorial	0.0208
8	https://www.javatpoint.com/angular-7-tutorial	0.019
9	https://www.javatpoint.com/cpp-tutorial	0.0187
10	https://www.javatpoint.com/aws-tutorial	0.0181

Table 1. Top 10 internal links

We have represented the TF-IDF score data in graphically in the following graph.

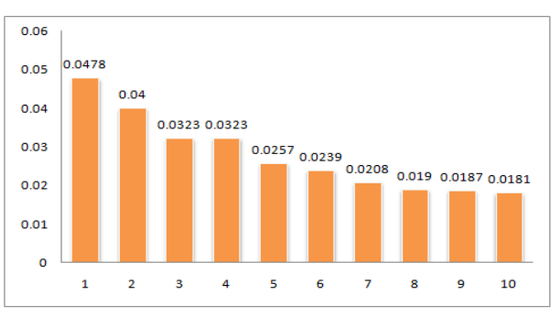

Fig 2. TF-IDF score graph

In our application in user part, we design a search bar, this is for searching the data from the web. For this we have used Google search engines. We can retrieve the data means we URL's using through google with help of python API. We can see the execution screenshot in the fig 3. For this we are using the python API to crawl the Google Search Results. This application we design in the Python Django Framework, coming to database we have used PostgreSQL for soring all the internal URLs and URLs which are ranked high with TF-IDF algorithm.

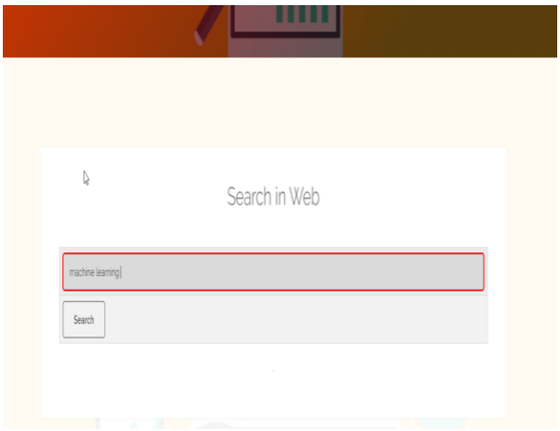

Fig 3. Search page

After user search request we can see the web results using Google we can see in fig 3. The search results sequence and pattern are not customized. The structure and ranking getting from Google search engine in our application.

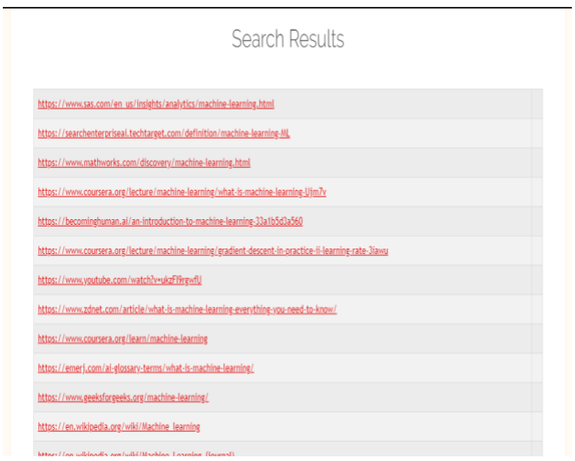

Fig 4. Search Results

In previous results we can see that there is no crawling of the internal URLs for root pages. Because there is no URL is matched with classified results. After getting search web results from the search engines, system will match with stored URL's which are processed with classification algorithm. If any web URL is matched, then system will display crawled URL's along with root URL's. In this situation Web Usage Mining will perform, while displaying the crawled URL's system will collect the history of the browser, and system will re-rank the crawled URL's according to the frequency of the webpage usage. For getting of the web history, we are using Chrome browser history.

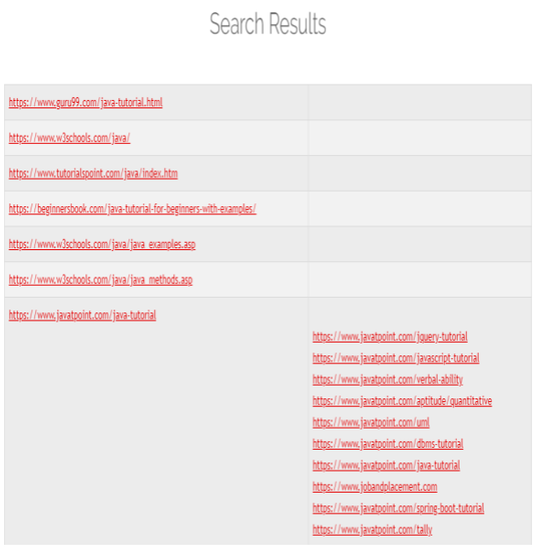

Fig 5. Search Results

Benefits of Web Page Prediction

1. The use of prediction in web page access will reduce time.
2. Pre-fetching the web page through prediction will minimize network latency.
3. It reduces the search time.

Based on this, here proposed session discarding strategies. Applying session discarding strategies will help to discard the old session to handle and store the new session. The discarding session contains three strategies such as first come first leave, least frequent ones leave, and older than the time frame left.

- In the strategy of first come first leave, the first old session will be discarded and the session which originates first will be discarded first.
- In the strategy of least frequent one leave, each session will be associated with the count of frequency and then the session will have a count of least frequency.
- In the strategy of older, than some time frame left, the time frame will be taken and sessions older than some days will be discarded once the records of session extended their capability. For example, a session will discard if the session is older than 30 days.

Instead of using web page URLs, here used web page content and each web page that is associated with semantic web pages. Hence, each set of keywords shows concept and session sequences.

9. CONCLUSION

In web world, we are very much habituated for the surfing in web for different purposes like information gathering, e-commerce, social connectivity etc. For this collection of data from the web is become very easy now a days due to the search engines like google, Bing etc. But in Web Usage Mining (WUM) personalized web search results is always a trending topic, most of the search engines following few concepts of Personalized Web Search (PWS) like keyword suggestions, coloring technique etc. But these are not satisfying the user needs, here in our application we are proposed PWS concept by taking concepts of Web Content Mining and Web Usage Mining. In our application we predicted the search results of web pages

based on the user log and content mining. Here for user log, we are taking the web browsing history of the user and we are using Term Frequency Inverse Document Frequency for content mining. In this we are achieved a dynamic way of user browsing web pages prediction. Our results proved concept of PWS in the dynamic way.

10. FUTURE WORK

In our proposed work we focused on the personalized web search results, and we have taken two concepts like web content and user log. But for personalized web search results we can take few more to get better results. In future we need to focus on the personalized web results based on the location and user activities. There are some web pages which are designed with high security those pages are so difficult to crawl, we need to work on those pages to get the internal URLs.

Few pages with honeypot trap which give us dangerous internal URLs we need to work on eliminating them in further version of application.

11. REFERENCES

[1] Saraiva, P.C., de Moura, E.S., Ziviani, N., Meira, W., Fonseca, R., Ribeiro-Neto, B.: Rank-preserving two-level caching for scalable search engines. In: Proceedings of the 24th annual international ACM SIGIR on Research and development in information retrieval, New Orleans, LA, September 2001, pp. 51–58 (2001)

[2] F. Zhao, J. Zhou, C. Nie, H. Huang and H. Jin, "SmartCrawler: A Two-Stage Crawler for Efficiently Harvesting Deep-Web Interfaces," in IEEE Transactions on Services Computing, vol. 9, no. 4, pp. 608-620, 1 July-Aug. 2016, doi: 10.1109/TSC.2015.2414931.

[4] S. Qi, D. Wu and N. Mamoulis, "Location Aware Keyword Query Suggestion Based on Document Proximity," in IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 1, pp. 82-97, 1 Jan. 2016, doi: 10.1109/TKDE.2015.2465391.

[5] Y. Tang, H. Wang, K. Guo, Y. Xiao and T. Chi, "Relevant Feedback Based Accurate and Intelligent Retrieval on Capturing User Intention for Personalized Websites," in IEEE Access, vol. 6, pp. 24239-24248, 2018, doi: 10.1109/ACCESS.2018.2828081.

[6] J. Berkhout, "Google's PageRank algorithm for ranking nodes in general networks," 2016 13th International Workshop on Discrete Event Systems (WODES), Xi'an, 2016, pp. 153-158, doi: 10.1109/WODES.2016.7497841.

[7] L. Shou, H. Bai, K. Chen and G. Chen, "Supporting Privacy Protection in Personalized Web Search," in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 2, pp. 453-467, Feb. 2014, doi: 10.1109/TKDE.2012.201.

[8] Kilic, Sefa & KARAGOZ, Pinar & Toroslu, Ismail. (2013). Clustering Frequent Navigation Patterns from Website Logs by Using Ontology and Temporal Information. 10.1007/978-1-4471-4594-3_37.

[9] M. Dhandi and R. K. Chakrawarti, "A comprehensive study of web usage mining," 2016 Symposium on Colossal Data Analysis and Networking (CDAN), Indore, 2016, pp. 1-5, doi: 10.1109/CDAN.2016.7570889.

[10] Amit Pratap Singh 1Research Scholar, Dr. R. C. Jain Director Samrat Ashok Technical Institute, Vidisha, Madhya Pradesh, India, “ A Survey on Different Phases of Web Usage Mining for Anomaly User Behavior Investigation” ., Barkatullah University, Bhopal, M.P, India.