Personalization and Clustering of Similar Web Pages

Smita Gupta
Assistant Professor
CS & IT Deptt
Moradabad Institute of Technology,
Moradabad-244001, Uttar Pradesh, India

Anurag Malik
Associate Professor
CS & IT Deptt
Moradabad Institute of Technology,
Moradabad- 244001, Uttar Pradesh, India

ABSTRACT

Over the last decade, clichéd information age has justly arrived. Moreover, the evolution of the Internet into the Global Information Infrastructure, together with the massive popularity of the Web, has also enabled the ordinary citizen to become not just a consumer of information, but also a part of it. In order to make user trouble free, it is required to save his/her time and effort. So some way is needed to give the relevant information to the user in a quick way and also enables to manage the whole lot of data without troublesome. Through this paper, the authors have used tf-idf (term frequency inverse document frequency approach) technique along with the concept of web mining to attain the required solution. Web mining is the application of data mining techniques that aims in discovering the patterns from the Web. Among its different ways, like Web usage mining, Web content mining and Web structure mining, here, efforts are only being made in the field of web content mining. In this work, a windows application is developed which act as a data analysis tool. This application is using the API of Bing search engine. The proposed algorithm is applied on the snippets (short description provided below each search result) of web search results to find those web pages that contains maximum number of query words. Moreover, it also aims at managing the information more easily on client's machine by using simple grouping technique.

General Terms

Web mining, Web Content mining.

Keywords

Term frequency-inverse document frequency (tf-idf); static clustering; Mining methods and algorithms; Information Retrieval.

1. INTRODUCTION

In the last few years, there is a tremendous growth in the amount of online resources. Moreover, the progress of the Internet into the Global Information Infrastructure, together with the immense popularity of the Web, has also allowed the ordinary citizen to become not just a customer of information, but also a part of it. Distribution of this enormous information led to use various accessibility techniques [4]. In order to make user trouble free, it is required to save his/her time and effort.

A user can utilize a search engine [15] and find, in most cases, data that is useful for their needs in a reasonable amount of time. Online searching is one of the most important and valued activities on the Internet and search engines are the gateways to access its information. But certain problems exits even there:

- ❖ The problem with many search engines is that relevance is hard to determine quickly and still more valuable assets are wasted while trying to find the most relevant data.
- ❖ Web information has enlarged from quantity to types, showing the trends of exponential growth, so the search engine cannot index all the pages.
- ❖ The web information has changed dynamically, so the search engine cannot be sure to update in time
- ❖ Search engine requires hardware owing more storage capacities.

Conventional document retrieval systems return long lists of ranked documents that users are forced to go through to find relevant documents. The majority of today's Web search engines (*e.g.*, Excite, AltaVista) follow this paradigm.

Moreover, data representation and data management are also vital problems in the world of online documents as the user's satisfaction is not confine to the single document rather he is interested in seeing many similar results for the purpose of increased reliability. But, currently, the situation is something different. User gets loads of results that may satisfy his search to the fullest but these results are not in proper form or in other words that there is deficiency in the arrangement of various pages. Due to the properties of the huge, diverse, dynamic and unstructured nature of Web data, Web search [2] has encountered a lot of challenges, such as scalability, multimedia and temporal issues etc. As a result, Web users are always drowning in an "ocean" of information and facing the problem of information overload when interacting with the web.

Here, in this paper, the authors have focused on search results personalization as well as static clustering of similar web pages. This similarity is based on the query. Search result personalization [13] is the dynamic process of enhancing web search results with related user preferences with the aim of providing personalized results. This technique of personalizing and clustering has led to emerge out many ways that helps users in quickly and accurately retrieving and organizing the interested information.

Through this work, the aim is to provide such kind of facility to the user that decreases his/her effort in searching the information. In other words, web personalization is achieved by this work that eases the comfort level for any web searcher. Web personalization [3] is becoming necessary now a day as it includes the steps that mould the information or services provided by any web resource according to the needs of a particular user as shown in figure 1. It gives emphasis on discovering the priorities and interest of individual users in order to provide personalized results.

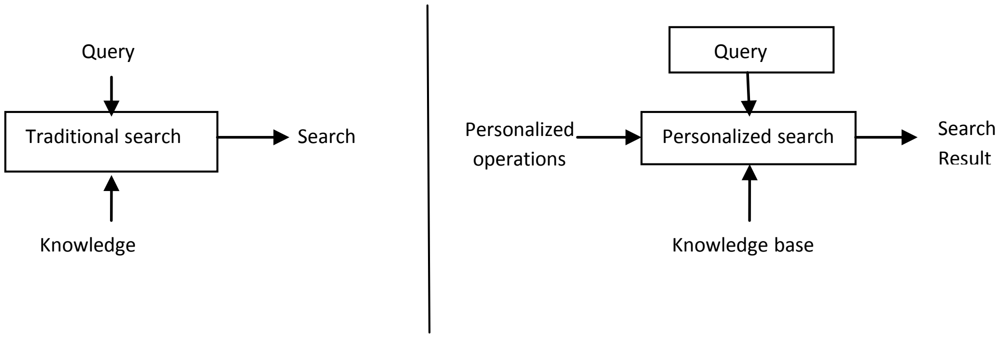

Figure 1: Traditional versus personalized search

This work introduces the new way of finding and arranging the relevant information that provides an appealing effect to the searcher without putting much stress. As especially the browsing activities beyond search are outside the reach of a search engine, client-side solutions are favorable. Moreover, as all user data is kept locally, user privacy is not violated. So, therefore, a client-side search personalization has been set up.

Among the three kinds of web mining, web text mining is of great use for this work. Web content mining is a great way to get the content you need from the web by using minimum amount of work. Web text mining continuously gaining its importance due to the tremendous growth in the web contents. Web text mining is a fundamental step in extracting knowledge from vulnerable amount of data. This step is of great use due to the nested structure of the web.

2. RELATED WORK

2.1. Objective

The ultimate goal of this work is to give emphasis on how to collect relevant documents about the user's interests and how to organize it in such a way that requires minimal user intervention. The motivational factor behind this work is to rearrange the web search results provided by the search engine to facilitate adaptive personalization [18] to the user and thereby comforts his/her search. Whenever a user sends a query to any search engine (like Google, yahoo etc.), then search engine returns a thousands of results to the user. Two levels of reduction can then be possible on the given number of search results.

First level personalization is achieved by using tf-idf approach. This technique is applied by calculating the frequency of major term present in the user's query in the result set provided by the search engine. For the second level personalization, user can ask to see top-k retrievals in the cluster formed in first level, where k cab be defined by the user. Thereby, reducing the irrelevant documents or web pages from the final list. At any point of time, user can see the actual results provided by the search engine. Additionally, this research also focuses on how to cluster the similar documents in a folder on client's machine.

2.2. Methodology Used

The basic methodology works at the level of client side. Basically, various personalization techniques [6] require implicitly or explicitly collecting visitor response. However, this work rank the web documents on the basis of its usefulness for the user based on the query. Thus, the ranking is *query-dependent.* Principal elements of Web personalization include modeling of Web objects (pages, etc.) and subjects (users), categorization of objects and subjects, matching between and across objects and/or subjects, and determination of the set of actions to be recommended for personalization and clustering of relevant documents.

The general architecture of the related work has been shown in figure 2. The algorithm involved may take the following steps:

1. A *start set* of documents matching the whole query is fetched from a search engine.
2. The start set is then augmented by its term's frequencies present in the query. So that the resulting set (in a clustered form) is reduced to only contain relevant documents.
3. Now all the results are again ranked in descending order of total frequency obtained from combining term's frequencies for each result.
4. Next level of reduction in the number of results is achieved by explicitly asking about the top- k retrievals from the user, where k is an integer number specifying the number of results.
5. At any point of time, user can neglect the level of personalization.
6. Finally, all the downloaded documents are clustered in a folder to the client's side based on the query name.

2.3. IMPLEMENTATION

In this work, a windows application is developed which act as a data analysis tool. This application is using the API of Bing search engine [5]. The API returns more than 40 results for a single query, but due to processing speed's reason, the authors have used only 40 search results to implement the concept only. The proposed algorithm is based on tf-idf (term frequency inverse document frequency approach). This algorithm is applied on the snippets (short description provided below each search result) of web search results to find those web pages that contains maximum number of query words.

The whole algorithm is demonstrated by developing a windows application that has default namespace and assembly name as "search it" and developed in C# in .net platform. The approach used is relatively different to various other approaches as in this various comparatively easy concepts are applied for finding the relevant pages based on frequency. Thereby, reducing the amount of time in processing and managing the whole lots of web data. Namespace used is:

Namespace ="http://schemas.microsoft.com/LiveSearch/2008/03/Search"

Functional operations of this work include:

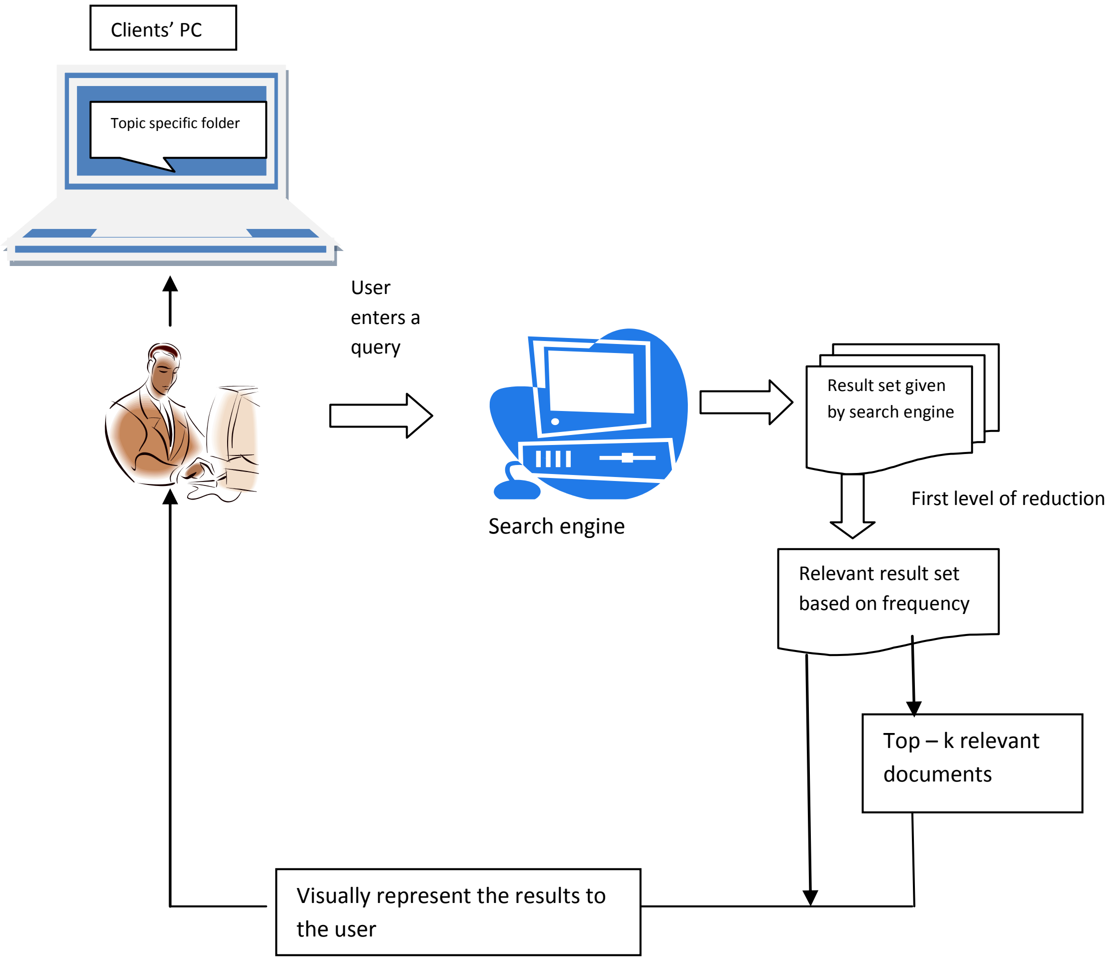
Clients' PC
Topic specific folder
User enters a query
Search engine
Result set given by search engine
First level of reduction
Relevant result set based on frequency
Top – k relevant documents
Visually represent the results to the user

Figure 2: General Architecture of the Application

1. Converts the Results into a partial result set/subset so that sorting can be applied on the Results.
2. Removes the frequently used words/proverbs/pronouns etc.
3. Sets the Frequency of the words in the Subset property.
4. Sorts the result set based on the Frequency of words in the short Description named as snippet of each URL
5. Displays the Result Set with an extra node showing the frequency.

Chiefly, the whole work of web personalization and clustering of similar pages includes the following phases:

i) Collection phase
ii) Relevance phase
iii) Retrieval phase
iv) Clustering phase

Collection Phase

This phase includes collection of web data from the World Wide Web. More specifically, list of URLs are obtained for a given query by using a Bing search engine's API. The web information retrieval process starts by creating a document collection phase that fetches the URLs of web search results provided by Bing search engine's API. This API returns more than 40 results. But due to processing speed, only 40 results are used in this project. This phase is also responsible for moldings the web data so that it can be used in next phase (i.e. relevance phase). In executing this phase, there is no need to keep track of any database as authors are dealing with the dynamic set of result set, which keeps on changing at every transaction. There may be various reasons for this dynamic behavior. First and most of course reason is that most of the search engine's rating is based on user click. So in no more than few seconds, this ranking may be affected, and the result set may get changed.

This phase can also be known as preprocessing phase as it may include removing the stop words and stemming the words to their base form using Porter Stemmer [14]. Each cleaned document is then converted to an IDF vector or 1/0 bit vector as required and used for further processing. Query normalization is also a sub-concept that is used in this. In this technique, instead of first calculating the frequency of most frequently words like connectors, verbs (i.e. is, am, are, what….), the query is passed to first normalization. After normalization, it will not contain any frequent terms, thereby, eliminating the need of calculating the idf (inverse document frequency) for all documents. Hence, helps in increasing the processing speed.

Relevance Phase

In this phase, tf- idf [7] approach is used half way. This means that term frequency is calculated in the very similar way as in tf-idf approach. To calculate term frequency (tf), each URL's snippet [10] is scanned to count the frequency of the words in the normalized query. But unlike tf-idf approach, inverse document frequency is not calculated here. As the only purpose of inverse document frequency is just to eliminate the count of highly frequent terms. But its biggest disadvantage is that this approach also increases the processing time as there are so many word connectors; verbs etc are in a webpage that results in wasting processing time. So in order to reduce processing time, the query is already being normalized (i.e. all its connectors, verbs, adjectives are removed from the query itself), that ultimately reduces the need to calculate the inverse document frequency. Additionally, to reduce the effect of various suffixes like advantages in place of advantage, the authors have also introduced a way. In this, all the words in the URL's snippets are checked after removing all the specified suffixes in the coding part. To implement this, the authors have used the concept of regular expressions that count the word after eliminating any suffixes in the word.

Traditionally, a similarity measure [9] is required to determine the angle of similarity between the document vectors and the query vector when they are represented in a *V*-dimensional Euclidean space in a vector space model [8], where *V* is the vocabulary size. In this work, Implementation of any similarity measure is not required here as it is assuming that the fetched results that are obtained from Bing search engine, have already some similarity with the query. Thus, none similarity measure is implemented. This elimination helps in reducing the processing speed of this work.

Top-K Relevant Phase

This is the easiest phase which just helps in retrieving the top most k relevant URLs. If a user is interested in seeing only the top most k results for its query, where k is an integer number, then it is made possible in this phase. This phase is primarily related to show only the requested number of results to the user. The advantage of this phase is it restricts the user's mind to the specific set of results; thereby avoid to deviating him from the useful results.

If a user clicks on a URL, then a new window is open for that page. Until the new opened window is not closed, none modification can be made on the main window. If a hyperlink is clicked on the opened page, then back and forward buttons are given to control the traversing. The mere objective of this phase is to provide only a set of results to the user that contains higher frequency of query terms. In other words, it returns a result set that contains as much as relevant information to satisfy the customer needs.

Clustering Phase

No doubt, clustering is a concept that bounds all similar entities into a same group. In this phase also, clustering means the same but the scenario is different. Here clustering is done on the client side by grouping all results in a folder that are downloaded by the client based on the query. This means that whenever a client saves a page, then it will automatically in the query named folder in C drive. Various other clustering approaches are made till now that performs dynamic clustering of web search results [19] [20].

This phase includes the offline clustering of similar web pages. This clustering phase accepts a list of documents and returns a collection of clusters (folders on client's machine). Each cluster consists of a variable number of documents. A cluster is created by locating all similar documents within a given list of documents according to the query.

3. EXPERIMANTAL SETTINGS

Various different queries are executed on "search it" as well as four other top most search engines- Google, yahoo, Bing, AltaVista. And then queries are matched for their result set to check how many results are matched in between my application "search it" and all other four search engine. The authors have carried the same task of matching for 6 different queries. Additionally, the authors have also matched the results out of different number (i.e. out of 10, 20, 30 results).

3.1 Analytical Results

The authors have performed two different experiments on the same application. First experiment is showing in figure 3. In this, a graph is drawn for different values of threshold values for a single query. As it can be easily see that as the value of threshold increases, number of results decreases.

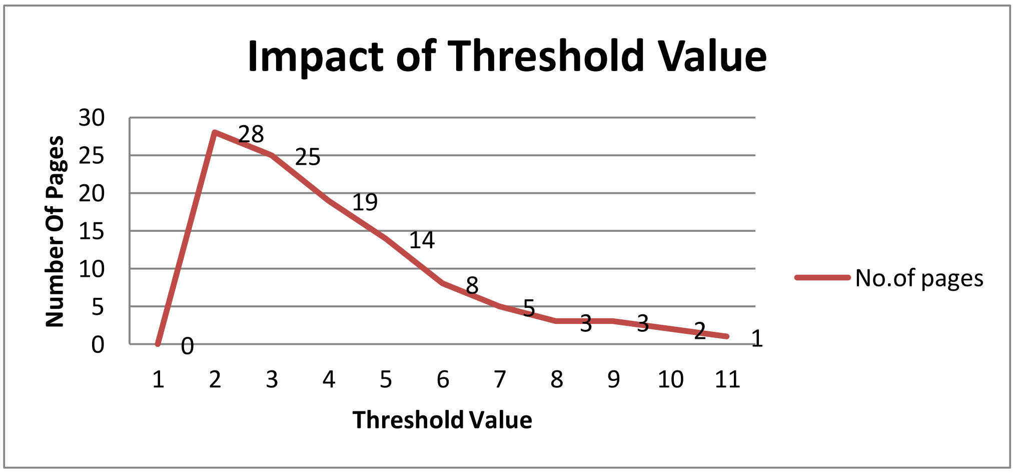

Figure 3: Graph Showing the Relation between the Threshold Value and the Number of Results

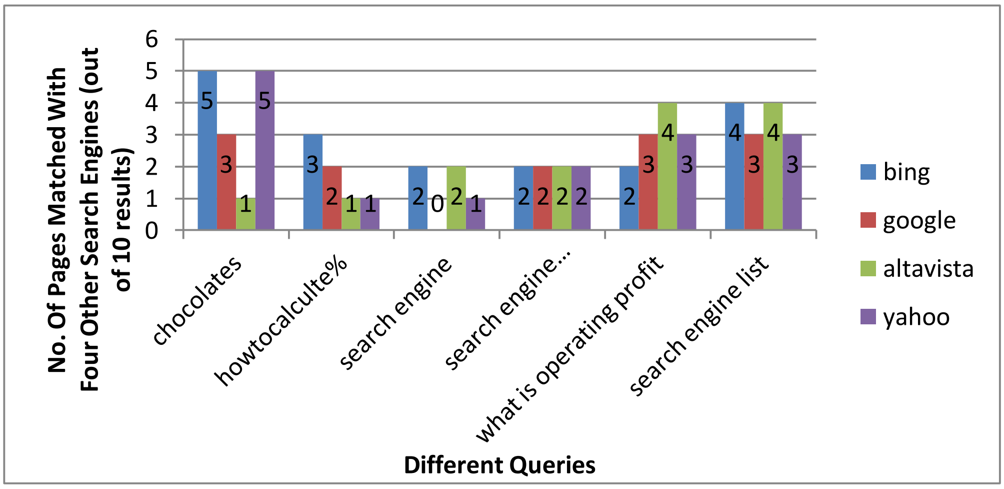

Figure 4: Graph Showing Comparative Analysis with Other Search Engines for First 10 Results

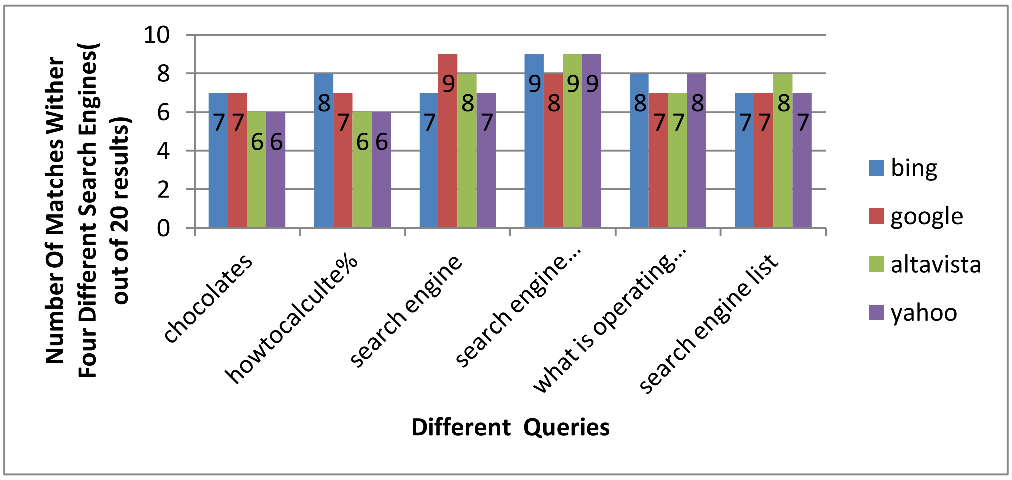

Figure 5: Graph Showing Comparative Analysis with Other Search Engines for First 20 Results

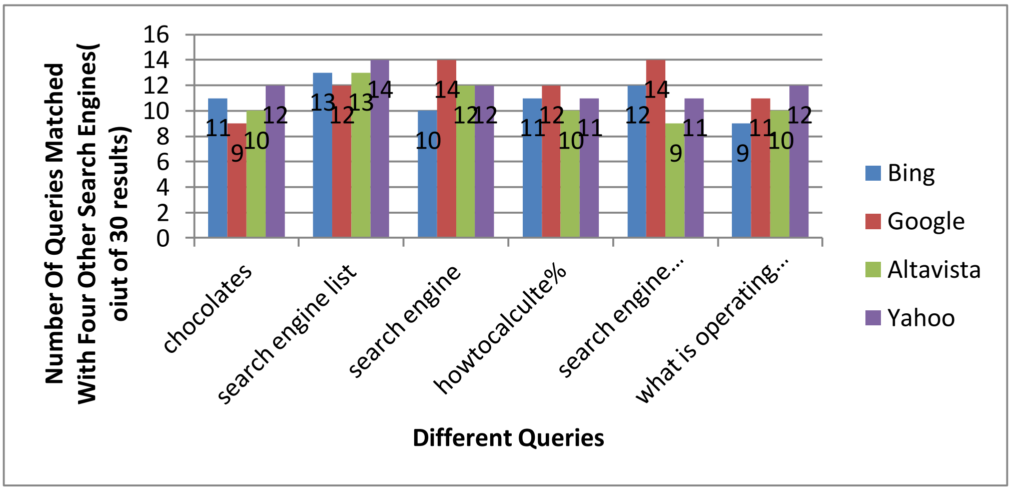

Figure 6: Graph Showing Comparative Analysis with Other Search Engines for First 30 Results

Here, threshold value implies the maximum frequency (number of matching characters) that must be fulfilled by all the satisfying URLs. So it is clear from the graph that as the number of matching words increases, the number of URLs in the result set decreases. Second experiment is carried out for 6 different queries between 4 search engines and developed application "search it". For carrying out the analysis, search engines like Bing, Google, Altavista and Yahoo are taken into consideration. Figure 4, 5, 6 are showing the relative comparisons for first 10, 20, 30 results respectively. Figure 4 shows that how many URLs are matched for a given query in each of the four search engines. For example, for a query "chocolates", "search it" application returns the results set in which 5, 3, 1, 5, URLs are matched with the top 10 URLs of the Bing, Google, Altavista, yahoo search respectively. Same is true for figure 5 and figure 6 for top 20 and top 30 URLs.

In all the three graphs, it can be easily observe that as the value of total number of results(web pages) increases in respective graphs, number of matched results also increases between the "search it" application and any other search engine . So that the web pages that contains maximum number query words occupies highest position in the result set of the "search it" application. In other words, we can say that content wise rich pages come before the all other pages, or sometimes irrelevant pages. Therefore, web content mining proves to be a successful tool in extracting the content wise rich web pages from the web. Generally, Web mining basically deals with the mining of large, heterogeneous and hyperlink online database. Also, being an interactive medium, graphical user interface isa key component of many web applications. So various issues are needed to handle sensitive and inexact queries that ultimately emerges the need for personalization. Thus, web mining [16], even if it is considered to be a particular application of data mining, lead to a separate field of research.

4. CONCLUSION

The first phase, collection phase, aims at fetching the web data from the Bing search engine and arrange in such a way that can be easily utilized by the second phase to find out the relevant URLs. This phase carried out all the necessary preprocessing steps such as normalization of query etc. The output of this phase is a query table that comprises of all the URLs row wise and all the words of a normalized query.

The second phase, i.e. the relevance phase, emphasis on how to evaluate the output of first phase. In this phase, all the URL's snippets are scanned to count the frequency of each query term. All the calculated frequencies for each URL are then summed up to calculate the total frequency of that page and ultimately its position in the list of relevant documents. The output of this phase results in a list of relevant documents, arranged in descending order according to their total frequency.

The retrieval phase helps the user in just showing the first k results from the list of relevant URLs obtained in the previous phase. This phase is also known as top-k retrieval phase. The only objective of this phase is to provide only a set of results to the user that contains higher frequency of query terms. Or in other words, a result set that contains as much as relevant information to satisfy the customer needs.

Lastly, clustering phase aims at managing the informational resources on the client side. This phase is also known as static clustering or offline clustering of similar web pages. It groups the similar web pages that a user wants to save on its machine, in a query named folder. So every time, when a user will save a web page by using this application, it will automatically save in folder, whose name is same as that of the query for that web page.

Overall, this dissertation aims at providing a content level personalization of web search results to the user. Till now, it has been seen that different search engines have different strategies to rank their documents, where contents of the web page does not play any significant role in ranking. Thereby, efforts are made to use the concept of web content mining in web search results with the only aim of reducing the user's effort and time in finding the useful information.

5. FUTURE DIRECTIONS

Various advancements can be applied in this field, as web mining is an upcoming field of modernization. This concept may be implemented at higher level by applying various tests to evaluate in WWW environment. This personalization approach can be implemented in the server side that can be elaborated to design an adaptive search engine [11]. Server side implementation can be much more successful. This higher implementation may removes various limitations of this work, such as lesser number of fetched results, fetch main content's of webpage, query length etc.

Through this application, the authors have just tried to implement a small concept on small set of data. Various other expansion can be applied to reaches it to a new height.

Query expansion is another concept that can be added to it. Just as several other search engines that present a list of options before a user about to complete his/her query, that module also enables to enhance the performance of this application.

Visualization techniques can be applied on web search results just to make it attractive and user friendly. Information visualization [17] provides a unique way to handle conceptual information by taking advantage of their visual perception capabilities. By showing information pictorially, human beings understand and process information more easily and strongly by using their perceptual abilities. Human mind have a very great tendency to perceive visual information more quickly that makes information visualization a powerful and necessary tool for information discovery. Various visualization techniques have been developed so far that aims at displaying the search results in such a way that enhance the way of representing the search results just as in TouchGraph [12]. Cloud mining [1] can also be integrated with web mining.

6. REFERENCES

[1]. Ajay Ohri, 2010, "Data mining through Cloud Computing". http://knol.google.com/k/data-mining-through-cloud-computing#.

[2]. Andrei Broder , 2002, "A taxonomy of web search" , IBM Research , SIGIR Forum, Fall 2002, Vol. 36, No. 2

[3]. Bamshad Mobasher, "Data Mining for Web Personalization", Center for Web Intelligence School of Computer Science, Telecommunication, and Information Systems DePaul University, Chicago, Illinois, USA

[4]. Giles, L. and S. Lawrence, 1999, "Accessibility and distribution of information on the web." Nature, 400.

[5]. API Basics , http://www.bing.com/developers/s/APIBasics.html

[6]. Personalization is not Technology: Using Web Personalization to promote your Business, http://www.boxesandarrows.com/view/personalization_is_not_technology_using_web_personalization_to_promote_your_business_goal. Accessed by Christian Ricci on 2004/01/12

[7]. Scoring and Ranking Techniques - tf-idf term weighting and cosine similarity, http://www.ir-facility.org/scoring-and-ranking-techniques-tf-idf-term-weighting-and-cosine-similarity. , Published Mar 31, 2010 by Michael Dittenbach

[8]. Information Retrieval and Data Mining, Part 1 – Information Retrieval, http://lsirwww.epfl.ch/courses/dis/2007ws/lecture/week%20

10%20Vector%20Space%20Model.pdf. Accessed by Karl Aberer, EPFL-IC, Laboratoire de systèmes d'informations répartis Information Retrieval – 1, 2007-8

[9]. Cosine Similarity and Term Weight Tutorial, An Information Retrieval Tutorial on Cosine Similarity Measures, Dot Products and Term Weight Calculations, http://www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html, by Dr. E. Garcia 2006

[10]. How does Google Pick Snippets for Your Pages to Show in Search Results?, http://www.seobythesea.com/2007/12/how-does-google-pick-snippets-for-your-pages-to-show-in-search-results/. Accessed by By Bill Slawski, on December 18, 2007

[11]. Martin-Bautista, M. J., Vila, M., and Larsen, H. L. 1999 , "A Fuzzy Genetic Algorithm Approach to an Adaptive Information Retrieval Agent," Journal of the American Society for Information Science (50:9), pp. 760-771

[12]. M. Angelaccio, B. Buttarazzi, M. Patrignanelli, 2007, "Graph Use to Visualize Web Search Results: MyWish 3.0", 11th International Conference Information Visualization (IV'07), © 2007IEEE

[13]. Mulvenna, M., Anand , S.S., B¨uchner, 2000, " A.G.: Personalization on the net using web mining", Communication of ACM 43(8) 122–125

[14]. Porter, M.F., 1980, "An Algorithm for Suffix Stripping Program", 14 no. 3, pp. 130-137.

[15]. Rainie, L. and J. Shermak., 2005, "Search engine use shoots up in the past year and edges towards email as the primary internet application." Technical report, Online Activities & Pursuits, Pew Internet & American Life Project.

[16]. Raymond Kosala, Hendrik Blockeel, 2000,"Web Mining Research: A Survey", In ACM SIGKDD

[17]. S. K. Card, J. Mackinlay, and B. Shneiderman. 1999, "Readings in Information Visualization: Using Vision to Think". Interactive Technologies Series. Morgan Kaufmann Publishers

[18]. Shady Elbassuoni, (2007), "Adaptive Personalization of Web Search", JUNE 2007 (elbassmasters)

[19]. Xiaohui Cui, Thomas E. Potok, Paul Palathingal , 2005, "Document Clustering using Particle Swarm Optimization", Applied Software Engineering Research Group Computational Sciences and Engineering Division Oak Ridge National Laboratory Oak Ridge, IEEE

[20]. Y. Wang, M. Kitsuregawa, " Link-based Clustering of Web Search Results", In Proceedings of The Second International Conference on Web-Age Information Management.