Clustering Web Search Engine Results for Improving Information Retrieval: A Survey

Mr. R. Keole1, Dr. P. P. Karde2, Dr. V. M. Thakare3

1Assistant Prof, Department of Information Technology, H. V. P. M. College of Engineering & Technology, Amravati, India
2Professor, Department of Computer science & Engineering, H. V. P. M. College of Engineering & Technology, Amravati, India
3Professor & Head of Department, P.G. Department of Computer science, Sant Gadge Baba Amravati University, Amravati, India

ABSTRACT

Copious material is available from the World Wide Web (www) in response to any user-provided query. Large document collections, such as those delivered by Internet search engines, are difficult and time-consuming for users to read and analyze. It becomes tedious for the user to manually extract real required information from this material. Though the physical characteristics of Web information is distributed and decentralized, the WWW can be viewed as one big virtual document collection. In that regard, the fundamental questions and approaches of traditional Information Retrieval (IR) research (e.g. term weighting, query expansion) are likely to be relevant in Web document retrieval. The Web document collection, massive in size and diverse in content, context, format, purpose and quality, challenges the validity of previous research findings based on relatively small and homogeneous test collections. Keeping information organized is an important issue to make information access easier. Although the information we need is sometimes available on the Web, this information is only useful if we have the ability to find it. With this aim, it is increasingly frequent to use automatic techniques for grouping documents. In this paper the main focus is on web clustering, that is, grouping documents based on the similarity of their contents. In this regard, document representation plays a very important role in web page clustering and constitutes the central point of research. The need is to sort out how to “organize” information. Clustering is useful technique in the field of web mining. Document clustering has been applied to Information Retrieval (IR) for over three decades. Its introduction to Information Retrieval was based on the grounds of its potential to improve the effectiveness of Information Retrieval systems. This paper will focus on the problem of improving the performance of information retrieval from web search engine results.

Keywords: - web mining, Information Retrieval (IR), Clustering, text mining, web search engine.

1. INTRODUCTION

With more than two billion pages created by millions of Web page authors and organizations, the World Wide Web is a tremendously rich knowledge base. The knowledge comes not only from the content of the pages themselves, but also from the unique characteristics of the Web, such as its hyperlink structure and its diversity of content and languages. A considerably large portion of information present on the World Wide Web (www) today is in the form of unstructured or semi-structured text data bases. The www instantaneously delivers huge number of these documents in response to a user query. However, due to lack of structure, the users are at a loss to manage the information contained in these documents efficiently. The www continues to grow at an amazing rate as an information gateway and as a medium for conducting business. Web mining is the extraction of interesting and useful knowledge and implicit information from artifacts or activity related to the www [1, 2].

One main problem in web search engine result is regarding organization of document data. This can be achieved by developing nomenclature or topics to identify different documents. However, assigning topics to documents in a large collection manually can prove to be an arduous task. The work proposes a technique to automatically cluster these documents into the related topics. Clustering is the proven technique for document grouping and categorization based on the similarity between these documents [3]. Documents within one cluster have high similarity with each another, but low similarity with documents in other clusters [4]. Web mining is an application of data mining techniques to discover and extract information from Web. Data mining has the role of searching large volumes of data and extracting Knowledge from data [5]. Clustering is one of the possible techniques to improve the efficiency in information finding process. It is a Data mining tool to use for grouping objects into clusters such that the objects from the same cluster are similar and objects from different cluster are dissimilar[6]. The proposed methodology of clustering can be incorporated within a web based search engine to provide better performance. The hybrid approach combination of Text, hyperlinks and web usage methodology will be more efficient for retrieving documents.
2. WEB MINING TAXONOMY
Web mining involves a wide range of applications that aim at discovering and extracting hidden information in data stored on the Web. Another important purpose of Web mining is to provide a mechanism to make the data access more efficiently and adequately. The third interesting approach is to discover the information which can be derived from the activities of users, which are stored in log files for example for predictive Web caching. Thus, Web mining can be categorized into three different classes based on which part of the Web is to be mined. These three categories are (i) Web content mining, (ii) Web structure mining and (iii) Web usage mining [2,7,14].

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to Web content has been the most widely researched. Issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision, the application of these techniques to Web content mining has been limited.

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting related pages. Web Structure Mining is the process of discovering structure information from the Web. This can be further divided into two kinds based on the kind of structure information used.

Hyperlinks: A Hyperlink is a structural unit that connects a location in a Web page to different location, either within the same Web page or on a different Web page. A hyperlink that connects to a different part of the same page is called an Intra-Document Hyperlink, and a hyperlink that connects two different pages is called an Inter-Document Hyperlink. There has been a significant body of work on hyperlink analysis. Document Structure: In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents.

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. There are three types of log files that can be used for Web usage mining. Log files are stored on the server side, on the client side and on the proxy servers. By having more than one place for storing the information of navigation patterns of the users makes the mining process more difficult. Really reliable results could be obtained only if one has data from all three types of log file.

3. CHARACTERISTICS OF WEB
It is often said that the Web offers an unprecedented opportunity and challenge for data mining. We believe that this is so due to the following characteristics of the Web [1,21,26]:
1. The amount of data/information on the Web is huge and still growing rapidly. Web data is also easily accessible.
2. The coverage of Web information is wide and diverse. One can find information about almost anything on the Web.
3. Data of all types exist on the Web, e.g., structured tables, texts, multimedia data (e.g., images and movies), etc.
4. Information on the Web is heterogeneous. Multiple Web pages may present the same or similar information using completely different formats or syntaxes, which makes integration of information a challenging task.
5. Much of the Web information is semi-structured due to the nested structure of HTML code, and the need of Web page designers to present information in a simple and regular fashion to facilitate human viewing and browsing.
6. Much of the Web information is linked. There are links among pages within a site, and across different sites. These links serve as an information organization tool and also as indications of trust/authority in the linked pages and sites.
7. Much of the Web information is redundant. The same piece of information or its variations may appear in many pages or sites. This property has been explored in many Web data mining tasks.
8. The Web is noisy. A Web page typically contains a mixture of many kinds of information, e.g., main content, advertisements, navigation panels, copyright notices, etc. For a particular application only part of the information is useful, and the rest are noises.
9. The Web consists of surface Web and deep Web. Surface Web is composed of pages that can be browsed using a normal Web browser. Surface Web is also searchable through popular search engines. Deep Web is mainly composed of databases that can only be accessed through parameterized queries using query forms.
10. The Web is also about services. Many Web sites and pages enable people to perform operations with input parameters, i.e., they provide services.
11. Above all, the Web is a virtual society. It is not only about data, information and services, but also about interactions among people, organizations and automatic systems.

12. The Web is dynamic. Information on the Web changes constantly. Keeping up with the changes and monitoring the changes are important issues for many applications.

4. WEB SEARCH ENGINE ARCHITECTURE
A web search engine must present a user with a set of results, given her input. There are five logical tasks a search engine performs for each search. Each task corresponds with a specific component of the architecture presented in Figure below, five tasks every search engine performs for each search [15, 16]:
   1. Accept user input
   2. Process user input
   3. Apply database query
   4. Process results
   5. Display results

Figure below describes the four required components of a web search engine and the crawler for populating the database. The first component is the user interface, which is responsible for accepting user input and presenting the output. Second is the query processor, which generates a reasonable database query from the user input. Third is the database, which is the component that stores the knowledge about each result. Fourth is the scoring module, which processes each result before sending the result to the user interface for display. In addition to these four components, most web search engines have a crawler, which is used to populate and maintain their database.

![Web Search Engine Architecture](image)

**User interface**
The user interface of a web search engine has two tasks:
   1. Accept user input for use by the system
   2. Present results to the user

We define the first task as the *input user interface* and the second as the *output user interface*. The simplest types of interfaces allow only a keyword query, while more complex input interfaces may allow users to choose options from a list, provide extra context information, or even track user actions. The second task of the user interface is to present results to the user. The output user interface affects how easily a user can identify useful documents.

**Query Processor**
User input is a representation of an information need. The query processor must convert this input into a database query for use by the search engine. Unfortunately, users typically don’t enter explicit database queries. The query processor of some search engines has the ability to generate database queries that are different from the query terms entered by the user.

**Database**
A web search engine can only return documents it knows about. The database is the collective local knowledge about the documents on the web. Currently some search engines boast over one billion pages in their index, suggesting a very large database. The query processor formulates queries to the database to produce a reasonable number of documents for scoring. There are several properties of web search engine databases that differentiate them from most databases for non-web online retrieval systems. A web search engine database determines what (local) documents can be returned to a searching user.

**Scoring Module**
Documents returned from the database are ranked by an *ordering policy*. The scoring module determines how documents are scored, and ultimately how they are ranked. Ordering policy refers to the method used by a search engine to produce a ranking of results. There are many factors that can be considered when ranking a document. Most classical IR systems use some version of TF-IDF, where the retrieval status value, or score, of a document is determined...
based on the term frequency of the query terms in each document relative to the number of documents in the collection (database) that contain that term.

**Crawler**

A web search engine is a tool that permits users to locate web pages based on content. This content mapping is done through a database of web pages. Most general purpose search engines populate their database through the use of a crawler, also called a web robot. A crawler explores the web by downloading pages, extracting the URLs from each explored page, and adding the new URLs to its crawl list. Since the web is not static – both pages and links change – and growing at a fast rate, a crawler must make decisions about which pages to examine, as well as which pages to index. Indexing is the, often complex, process of adding a page to a search engine’s database. Indexing involves parsing the document using many advanced algorithms to improve the efficiency retrieval. Tasks of the crawler:

1. Retrieve web pages for use in the database
2. Minimize the resource requirements – this can include CPU, RAM, disk, network, and time
3. Properly balance discovery of new contents with updating of existing contents

The simplest crawler can be thought of as a search algorithm.

**5. LITERATURE REVIEW**

Oren Etzioni was the person who coined the term Web Mining first time [1]. Initially two different approaches were taken for defining Web Mining. First was a “process-centric view”, which defined web mining as a sequence of different processes as resource discovery, information retrieval and generalization [1,10] whereas, second was a “data-centric view”, which defined web mining in terms of the type of data that was being used in the mining process [2]. The second definition has become more acceptable, as is evident from the approach adopted in most research papers [6,11]. Web mining is also a cross point of database, information retrieval and artificial intelligence. The most common way of representing text documents is using the Vector Space Model (VSM) [12], where each document is represented as a feature vector, which length corresponds to the number of unique attributes used for representing documents in the collection. Each vector component, that is, each feature, has an associated weight which indicates the importance of that attribute to characterize or represent the document. Web mining can be categorized into three different classes based on which part of the web is to be mined i.e. Web content mining, Web structure mining and Web usage mining [4,7,9,10]. Oren Zamir and Oren Etzioni [1] in their research listed the key requirements of web document clustering methods as relevance, browsable summaries, overlap, snippet tolerance, speed and accuracy. They have given STC (Suffix Tree Clustering) algorithm which creates clusters based on phrase shared between documents. Fresno and Ribeiro in 2004 presented an Analytical Combination of Criteria (ACC) to represent web pages. It is based on a linear combination of different heuristic criteria within the VSM. The criteria used by ACC are: Title: word frequency in the title of the document, Emphasis: word frequency in highlighted text segments, Position: word positions in a document & Frequency: word frequency in the document. Fresno in 2006 proposed an alternative way of combining them in a non-linear way. In this case, a fuzzy logic based system is employed to define the expert knowledge about how to combine these criteria.

[13] presents a method for extracting news from the Web, based on the visual perception of human users and try to simulate how human beings understand the information found in web news by using a function based object model. The objects of this model can be of four main types: information object, navigation object, interaction object and decoration object.

Researchers initially proposed document clustering for information retrieval and Web search to improve search performance by validating the cluster hypothesis, which states that documents in the same cluster behave similarly with respect to relevance to information needs. In recent years, researchers have used clustering to organize search results, creating a cluster based search interface as an alternative presentation to the list interface. Document clustering is widely applicable in areas such as search engines, web mining and information retrieval. Most document clustering methods perform several pre-processing steps including stop words removal and stemming on the document set [3]. Most of the document clustering algorithm worked on BOW (Bag Of Words) model [7]. Each document is represented by a vector of frequencies (TF) of remaining terms within the document. Some document clustering algorithms employ an extra pre-processing step that divides the actual term frequency by the overall frequency of the term in the entire document set (TF-IDF).

It has great potentials in applications like object recognition, image segmentation and information filtering and retrieval [4]. Most of the clustering techniques fall into two major categories, and these are the hierarchical clustering and the partitional clustering [4]. Scatter/gather described in [11] was an early cluster-based document browsing method that performed postretrieval clustering on top-ranked documents returned from a traditional information retrieval system.
The use of link structure to group document collections comes from citation analysis (Garfield, 1979), a commonly used bibliometric method, where it is assumed that if the author of a document cites another two documents, then these two documents should be somehow related from the author’s point of view. By applying the ideas from bibliometrics to the Web, Larson represents information by means of an input matrix where its element \((i, j)\) contains the number of documents citing both documents \(i\) and \(j\). After that, the raw co-citation matrix is converted to a correlation matrix. A framework presented in [20] uses hyperlinks for topic detection by means of clustering techniques, considering the corpus as a directed graph and document clusters as topics. Of course, though we can consider the fact that clusters have some implicit semantic properties, a topic detection stage would be needed to add semantics to the clusters. As one could expect, documents are represented as graph nodes, while hyperlinks are represented by edges between nodes. A method proposed in [21] models both document contents and link structure in a unified manner, though they employed different representations for each one. The approach relies on the assumption that the probability of a web page to belong to a category can be determined not only by its contents, but also by analyzing the contents of other pages referencing it and the strength of relations with them. On the one hand, textual contents were represented by means of TF-IDF. On the other hand, the link structure was employed by using a directed graph where nodes are documents and arcs are probabilistic links between them.

Kiduk Yang [28] in 2002 proposed a methodology of combining Text, Link and Classification-based retrieval methods to Enhance Information Discovery on the Web. The difference between text database and XML database results in three new challenges was stated in [32]: 1) Identify the user search intention, i.e., identify the XML node types that user wants to search for and search via. 2) Resolve keyword ambiguity problems: a keyword can appear as both a tag name and a text value of some node. 3) As the search results are sub trees of the XML document, new scoring function is needed to estimate its relevance to a given query. A new clustering methodology for XML search results, which clusters results according to the way they match the given query was proposed in [33]. Two approaches to implement the methodology are discussed. The first approach is a conventional one which does clustering after search results are retrieved; the second one clusters search results actively, which has characteristics of clustering on the fly.

A method that addresses XML document classification by considering both structural and content-based features of the documents was presented in [36]. This approach leads to better constructing a set of informative feature vectors that represents both structural and textual aspects of XML documents. For this purpose they integrated soft clustering of words and feature reduction into the process. To extract structural information, they employed an existing frequent Tree-mining algorithm combined with an information gain filter to retrieve the most informative substructures from XML documents. The problem of organizing a user’s historical queries into groups in a dynamic and automated fashion was stated in [37]. Automatically identifying query groups is helpful for a number of different search engine components and applications, such as query suggestions, result ranking, query alterations, sessionization and collaborative search.

A Summation-based Incremental Learning (SAIL) algorithm for Info-K means clustering was proposed in [38]. Specifically, by using an equivalent objective function, SAIL replaces the computation of KL-divergence by the incremental computation of Shannon entropy.

6. THE LIMITATIONS OF WEB SEARCH

With an enormous growth of the Internet it has become very difficult for the users to find relevant documents. In response to the user’s query, currently available search engines return a ranked list of documents along with their partial content. If the query is general, it is extremely difficult to identify the specific document which the user is interested in. The users are forced to sift through a long list of off-topic documents. Moreover, internal relationships among the documents in the search result are rarely presented and are left for the user. Standard information retrieval systems rely on two orthogonal paradigms: the textual similarity with the query (e.g., tf-idf-based cosine similarity) on one hand and a query independent measure of each web page’s importance (e.g., link authority ranking) on the other.
hand. However, these systems generally lack user modeling and thus are far from being optimal i.e. Different users may submit exactly the same query even though they have different intentions. The most famous examples of such ambiguous queries include bass (fish or instrument), java (programming language, island or coffee), jaguar (animal, car or Apple software) and IR application (Infrared application or Information Retrieval application).

7. ANALYSIS OF PROBLEMS IN WEB

The growth of the World Wide Web has enticed many researchers to attempt to devise various methodologies for organizing such a huge information source. Scalability issues come into play as well as the quality of automatic organization and categorization [1,2,7].

Documents on the web have a very large variety of topics, they are differently structured, and most of them are not well-structured. The nature of the sites on the web varies from very simple personal home pages to huge corporate web sites, all contributing to the vast information repository [9,36].

Search engines were introduced to find the relevant information on the web, such as Google, Yahoo!, and AltaVista. However, search engines do not organize documents automatically [37]; they just retrieve related documents to a certain query issued by the user. While search engines are well recognized by the Information Retrieval community, they do not solve the problem of automatically organizing the documents they retrieve. The problem of categorizing a large source of information into groups of similar topics is still unsolved. Typically, the following problems are often mentioned in Web related research and applications:

- To find specific information on the web, users often either browse Web documents directly or use a search engine as a search assistant [2,8].
- Most search engines perform in a query-triggered way that is mainly on a basis of one keyword or several keywords entered. Sometimes the results returned by the search engine don’t exactly match what a user really needs due to the fact of the existence of the homology [8,33].
- With traditional Web search service, query results relevant to query input are returned to Web users in a ranked list of pages. In some cases, we are interested in not only browsing the returned collection of Web pages, but also extracting potentially useful knowledge out of them [5,39].
- To improve the Internet service quality and increase the user click rate on a specific website, thus, it is necessary for a Web developer or designer to know what the user really wants to do, predict which pages the user is potentially interested in, and present the customized Web pages to the user by learning user navigational pattern knowledge [8].

Web data usually exhibits the following characteristics [2,5,8]:

- The data on the Web is huge in amount
- The data on the Web is distributed and heterogeneous
- The data on the Web is unstructured
- The data on the Web is dynamic

The above problems place the existing search engines and other Web applications under significant stress. A variety of efforts have been contributed to deal with these difficulties by developing advanced computational intelligent techniques or algorithms from different research domains, such as database, data mining, machine learning, information retrieval and knowledge management.

8. MOTIVATION

The real motivation is to help in the resolution of this problem by taking one step further toward a satisfactory solution. The intention is to create a system that is able to categorize web documents effectively, based on a more informative representation of the document data and targeted towards achieving high degree of clustering quality by implementing clustering technique/algorithim in web search engine. Hence the motivation behind this work is as follows,

1. The major source of information is from Web [1,5,7]
2. Traditional information retrieval (IR) approaches are hardly appropriate due to enormous size [1, 6, 7].
3. Manually extracting real required information is difficult to read and analyze [5,30].
4. Many search engines gives a long list of ranked documents and most of them are irrelevant [28].
5. Typical queries retrieve hundreds of documents; most of them have no relation with what the user is looking for [8,39].
6. Hence, the main problem is regarding the organization of document data presented on web pages [4, 31].
7. One of the Solutions to this is clustering [1, 3, 4, 6, 33].

9. PROPOSED METHODOLOGY
The limitations and shortcoming stated in existing methods has stated that the work was based on information retrieval based on either Web content – free text, image, records; Web structure – hyperlinks, tags or Web usage – web log data, url viewed, access time, date, referrer. Most of the work carried was based on Web Content which consists of Unstructured Data (Such as Free Text), Semi structured Data (Such as HTML doc), Structured Data (Such as Data in tables or database).

All the approaches mentioned in the literature survey produce clusters in such a way that each document is assigned to one and only one cluster. The methodology mentioned in the literature survey is mainly based on hard clustering and those which are based on soft or fuzzy clustering but the labeling of the clusters is a very daunting challenge of this time. No remarkable effort has been made in this regard to get good result. That is why automatic labeling of the clusters is not so much accurate.

Global search engines serve as de facto Internet portals, local search engines are embedded in numerous individual Web sites, and browsing is the most common activity on the Web, due to the hyper-linked structure that provides access to a large quantity of information in a restricted space.

In the proposed work, emphasis will be given on information retrieval based on Web content – free text, image, records, Web structure – hyperlinks, tags or Web usage – web log data, url viewed, access time, date, referrer. This is to increase the accuracy factor and relevancy in retrieval and clustering algorithm will be applied.

The proposed methodology of clustering can be incorporated within a web based search engine to provide better performance. The hybrid approach combination of Text, hyperlinks and web usage methodology will be more efficient for retrieving documents. The proposed model will be able to reduce the problem of speed while increasing accuracy to some considerable level over the existing approach. Therefore, it will be suitable for web search engine designers to incorporate this model in an existing web based search engine so that web users can retrieve their documents at a faster rate and with higher accuracy.

Browsing and searching complement each other and they are most effective when used together. The goal of a modern Web search engine is to retrieve documents considered “relevant” to a user query from a given collection. Nowadays, a user query is modeled as a set of keywords extracted from a large dictionary of words; a document is typically a Web page, pdf, postscript, doc file or whatever file that can be parsed into a set of tokens.

In the proposed work, representation is an essential stage for automatic document organization. The way in which documents are organized depends on how they are represented. Different representations can lead to different groupings. This work will focus on web page clustering, that is organizing a dataset in groups of related documents. This relation is based on the similarity among documents. The first thing to do is selecting the features or attributes of the documents to be used for representation and this is called features of the elements employed to characterize the documents. These elements are mainly the words that compound the documents. However, a word is not a feature as it is. Therefore, it needs to preprocess words to convert them into features or terms. A term is basically a preprocessed word. This preprocessing essentially consists of removing punctuation marks, removing stop words, and stemming the words in order to reduce each word to its main part by removing affixes.

In a very large set of web documents containing information of various topics either related topics or mutually exclusive topics, cluster the documents into a number of categories (clusters) such that: (a) the similarity between the documents in one category (intra-cluster similarity) is maximized, and (b) the similarity between different categories (inter-cluster similarity) is minimized. Consequently, the quality of categorization (clustering) should be maximized.

Therefore, the statement clearly suggests that given the large corpus of documents, a solution to the problem of organizing the documents has to produce a grouping of the documents such that documents in each group are closely related to each other, while the documents from different groups should not be related to each other.

The proposed work will implement an efficient approach to improve the performance of information retrieval in web search engine results; this work will focus on mining of the useful information as per the user query from the web documents.

The objectives of the proposed work are:

1. Selection of the application domain or data set to find specific information on the web and also for finding needed information as most search engines perform in a query-triggered way that is mainly on a basis of one keyword or several keywords entered [1,5].
2. Different web documents to be collected from the World Wide Web related to the selected application domain and will be referred as dataset [7].
3. Documents to be preprocessed. This preprocessing essentially consists of removing punctuation marks, removing stop words, and stemming the words in order to reduce each word to its main part by removing affixes [6].
4. Number of clusters will be created for the processed text using clustering algorithm and the web document will belong to the cluster depending on the similarity approach [1, 4].

5. Rules will be generated as per the association in the words of query provided by user. These rules will be added into rule base [8, 39].

6. The clusters are selected as per user query for retrieving information [1,6,7].

7. The specific and needed information will be mined from documents as per selected rules and clusters [5,7,28].

8. Simulation and performance evaluation of the proposed work [28].

10. CONCLUSION

The growth of the World Wide Web has enticed many researchers to attempt to devise various methodologies for organizing such a huge information source. Search engines were introduced to help find the relevant information on the web. However, search engines do not organize documents automatically; they just retrieve related documents to a certain query issued by the user. While search engines are well recognized by the Information Retrieval community, they do not solve the problem of automatically organizing the documents they retrieve. According to the limitations and shortcoming stated it has been found that the work was based on information retrieval based on either Web content, Web structure or Web usage. Most of the work carried was based on Web Content which consists of Unstructured Data (Such as Free Text), Semi structured Data (Such as HTML doc), Structured Data (Such as Data in tables or database).

In the proposed work emphasis will be given on information retrieval based on Web content, Web structure & Web usage. This is to increase the accuracy factor and relevancy in retrieval and clustering algorithm will be applied. The proposed methodology of clustering can be incorporated within a web based search engine to provide better performance. The hybrid approach combination of Text, hyperlinks and web usage methodology will be more efficient for retrieving documents. The proposed model will be able to reduce the problem of speed while increasing accuracy to some considerable level over the existing approach. Therefore, it will be suitable for web search engine designers to incorporate this model in an existing web based search engine so that web users can retrieve their documents at a faster rate and with higher accuracy. As future work, intention to investigate the usefulness of the knowledge gained from these query groups in various applications such as providing query suggestions and biasing the ranking of search results.

REFERENCES


[27] Jingwen Tian, Meijuan Gao, and Yang Sun, “Study on Web Classification Mining Method Based on Fuzzy Neural Network”, Shenyang, China August 2009


[38] Jie Cao, Zhiang Wu, Junjie Wu, Member, IEEE, and Hui Xiong, Senior Member, IEEE,” SAIL: Summation-bAsed Incremental Learning for Information-Theoretic Text Clustering”, IEEE Transactions On Cybernetics, VOL. 43, NO. 2, April 2013.