Biosearch: A Domain Specific Energy Efficient Query Processing and Search Optimization in Healthcare Search Engine

Asha Unnikrishnan
M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India, India

B. Senthil Kumar
Assistant Professor, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India

Abstract – Biomedical information extraction is always a difficult process due to its huge set of results and unknown keywords. Some medical terminologies are not aware by the users; often it's very difficult to retrieve such content from the website like PubMed. Several work used concept hierarchies for easy search navigation, however the biomedical query retrieval processing time always higher than the normal data retrieval. Ranking, summarizing and categorization have been proposed together to improve the searching efficiency. Results categorization and result summarization based on the concept hierarchy for biomedical databases is the focus of this work. A natural way to organize biomedical citations is according to their keywords and tags, additionally the conceptual similarity can also be performed in the proposed system. In this paper, a new BioSearch engine is proposed with effective data mining algorithms with less energy for query processing. The proposed system contains Predictive data caching technique for fast data retrieval; this has been performed with the help of facet order and concept hierarchy methods. The proposed system also provides the auto query incremental algorithm to ease the search. Finally the retrieved data's are ranked and summarized using RII (Ranked Inverted Index) algorithm. This helps to summarize and simplify the result to the user view.

Index Terms – Data Mining, Web Mining, Biosearch, Query Processing, Ranked Inverted Index, Ranking.

1. INTRODUCTION

The internet is a collection of hundreds of thousands of web links that often contain high quality information. The basic aim of a user is to select the best collection of information out of this huge repository for his particular information need. A web search engine is designed to collect web links from the World Wide Web and thereafter retrieve information in response to a user query [1]. This paper mainly focuses on creating a hybrid search engine for medical domain, which should have an automatic crawling and indexing process [2]. This paper involves study of different web content extraction and indexing algorithms and its implementation. This paper mainly aim to design and developing a search engine for the medical domain with effective query processing techniques with web crawling architectural model, which can be useful to retrieve valuable pages during a crawling process for effectiveness of medical domain search engine. The proposed paper also contains the existing inverted indexing concepts [3] and makes enhancements as needed with the effective query processing and ranking. The main objective of the study is to develop a medical domain search engine with the analysis of deferent query processing and indexing algorithms to increase the search engine process and covering more and more meaningful information from the World Wide Web. Restructuring the ranked inverted index to gather meaningful information from the particular website for the categorization of the medical terms and the web content for biomedical search engine indexing process without rendering into the entire website is aimed to develop. Enhancing the search experience of the user for medical terms is another aim of the proposal. Using the proposed system, the query processing time will be reduced and retrieving effective and useful links for the given query is performed.

2. PROBLEM DEFINITION

Getting desired results from the search engines is the common expectation of every user in the internet. Due to the massive development of web contents, data search is not fully optimized. This created many internal issues and challenges. Internet users use text based queries as a request to seek information using any search engine. Search engine then tries to infer and retrieve the relevant documents by performing the matching of query to the surrogates of documents and present the likely relevant documents to users in the form of hits list. Search engine performs query processing process by performing document indexing methods. So effective indexing, query processing with energy minimization [4] is an important attribute of the future research. And moreover, the search

engines are common and very few developed with domain specific features [5]. So this will be interesting to have a search engine for biomedical and healthcare domains. Because the query processing may always creates some problem when it is domain specific. User may not have proper knowledge about the health related queries. And several approaches under data mining is used to handle medical data's [6], and a list domain specific search engine and techniques were discussed in [7], So this will be an interesting idea for the further research.

3. PROPOSED SYSTEM

Data retrieval from the web is easier due to the huge development of search engines like Google, Yahoo etc., however, the web data source may contain huge set of resources, which need effective filtering and ranking. It is even tougher for the user when they need biomedical related links or information's from the web. So the proposed system developed a BioSearch engine, which is healthcare specific search engine. This BioSearch facilitates the fast crawling and indexing the healthcare related information's instead of common search.

The BioSearch engine is a medical information retrieval tool that crawls, indexes internet web pages that are related to the medical domain and stores them into a separate repository. It returns list of page links that match user queries using inverted index list. The indexes are generated using the web link category similarity. BioSearch is a tool that enables users to locate information on the World Wide Web for medical related contents. It uses keywords entered by users to find Web links and its short descriptions, which contain the information required by the user. It searches links for specified keywords and returns a list of links and the description about the query from the client. It employs bots or crawlers to search the web. The information gathered by the bio-crawlers is used to create a searchable index of the Net. The proposed search engines maintain very large databases that contain information about the web pages with effective indexing and query processing techniques. They are automatically updated by crawlers that search the WWW for new content and then report their findings to the database. There are three basic types of search engines such as crawler based, human powered and hybrid. The proposed bio-search engine is a hybrid based search engines.

Contributions of the proposed System:

To achieve better crawling and searching for medical contents from the WWW, the following algorithms and techniques are used.

- Biomedical data's are more complicated to understand, because it contains many keywords which are related to that domain. In order to ease the biomedical data search, a new search engine is created and named as Biosearch, which is domain specific search engine for clinical data retrieval.
- The Biosearch involved with several common problems of search engines like energy effective query processing, crawling and indexing.
- For fast data retrieval an Energy saving keyword Predictive Algorithm is developed. This reduces the time taken to predict the user query. Instead of providing results based on the user query, the system gives a hierarchical and nested structure of links. User can easily navigate to the links from that.
- To reduce the delay, the Predictive data crawling technique is used. This crawl the data and eliminates the auxiliary information's.
- To perform effective data retrieval auto query incremental and indexing methods are used.
- To reduce the size of the result page, a Ranked inverted index algorithm is developed. This will orders the links based on the domain similarity and ranks with access history.

The above contributions are made in the proposed system with additional application features. The following section explains the detailed study of the above techniques and methods.

4. BIOSEARCH

In Biosearch search engine the downloading of web pages is done by several crawlers. There is a URL server that sends lists of URLs to be fetched to the crawlers. The web pages that are fetched are then sent to the store server. The store server then compresses and stores the web pages into a repository. Every web page has an associated ID number called a link ID that is assigned whenever a new URL is parsed out of a web page. The indexing function is performed by the ranked inverted indexer. The indexer performs a number of functions. It reads the repository, uncompresses the links, and extracts them. Each link is converted into a set of word occurrences called hits. The hits record the word, position in link, an approximation of font size and capitalization. The indexer distributes these hits into a set of "barrels", creating a partially sorted forward index. The indexer performs another important function. It parses out all the links in every web page and stores important information about them in an anchors file. This file contains enough information to determine where each link points from and to, and the text of the link. The content crawler from the URL reads the anchors file and converts relative URLs into absolute URLs and in turn into link identifiers. It puts the anchor text into the forward index, associated with the link identifier that the anchor points to. It also generates a database of links which are pairs of link identifiers. The links database is used to compute page ranks for all the links and inverts it before providing the result.

Phases of BioSearch Engine

BioSearch match queries against an index that they created earlier. The index consists of the words in each link and pointers to their locations within the links with appropriate ranking. This is called a ranked inverted index file. This comprises four essential phases:

1. A link processor and crawler
2. A query computation
3. A search and matching function
4. A ranking capability

While users focus on "search," the search and matching function is only one of the four phases. Each of these four phases will be executed and links will be retrieved.

Algorithm: Crawling process with (Predictive data caching)

1. *Begin*
2. While (URL set is not empty)

Begin

a. Take a URL from the set of seed URLs;

b. Initiate a webrequest Wr(URL).

c. Get the web response W←Wr(URL)

d. Download the page source file which carries downloading permissions and also specifies the files to be excluded by the crawler;

e. Determine the protocol of underlying host like http, ftp etc.;

f. Based on the protocol of the host, download the link;

g. Identify the link format like hreg, <a>, www, .com etc.;

Check whether the link has already been downloaded or not; If the link is fresh one Then

a. Read it and extract the links or references to the other cites from that links;

ii. Else

1. Continue;

h. Convert the URL links into their absolute URL equivalents;

i. Add the URLs to set of seed URL list;

End;

In the BioSearch, the query computation has seven steps, though a system can cut these steps short and proceed to match the query to the inverted ranked index file at any of a number of places during the computation. Link processing many steps with query computation. More steps and more links make the process more expensive for computation in terms of computational resources and responsiveness. So, BioSearch system developed with time reduction techniques without compromising the query results.

The next step is RII (Ranked inverted index) it consists of the same process as the forward index, except that they have been processed by the sorter. For every valid term ID, the lexicon contains a pointer into the container that term ID falls into. It points to a link list of link identifiers together with their corresponding hit lists.

The ranked inverted indexing phase is a web content mining process, where the structured information's are used to rank. Starting from a collection of unstructured files or links the indexer extracts a large number of information such as the list of links, which contain a given term and the case, the typeface, the number of all occurrences of each term within every link. The indexer stores the information extracted in a structured archive (i.e. the index) which is usually represented using a Ranked Inverted Index (RII) File. RII is the most useful format for the index of a BioSearch due to its relatively small space occupancy and the efficiency involved in resolution of keywords based queries. RII index consists of an array of posting lists where each posting list is associated to a distinct term t and contains the link identifiers of all the links containing t as shown in Table 3.1. Sometimes, the position of the term within the page is also included in the posting list in the index. Since the link identifiers of each posting list are sorted, they are stored with a simple difference coding technique. For e.g. consider the posting list ((Cancer ; 78, 2,3,4,5,6,7,9,13,53,12) indicating that the term caner appears in 78 links having the link identifiers 2,3,4,5,6,7,9,13,53,12 respectively. The above list can be written as ((Lung cancer; 2-9, blood cancer (9-13), where the items of the list represent the difference between the successive link identifiers.

The indexer of existing search engine collects information from the web links gathered in the web repository and creates an index which is bound to be a global large sized index. This global large sized index is stored in the form of ranked inverted indexed (RII) files. This index is so large in size that the search in response to user's query has to traverse through large number of entries.

In this work, hierarchical ranked indexes are being proposed that are created at multiple levels of hierarchy of links clustering. Thus, instead of creating a single global index, the index is created at different levels. At first level of link hierarchy, the similar links are clustered into clusters on the basis of links similarity. As described in clustering based indexing, the links in the same cluster are assigned the closer link identifiers and the link level index consists of the posting

list which contains the terms and the link identifiers of the links which contain the given term. This index has a structure similar to the global large sized index which is followed in the existing search engines. It is a large sized index and is a detailed index. Then at the next level, the similar clusters are clustered into mega clusters as discussed earlier. The index at this level contains only the terms and hence it is a small sized index. At the last level of hierarchy, the similar mega clusters are clubbed together to form super clusters as shown in the hierarchical clustering.

The index at this level has a structure similar to that of index at mega cluster level. This is also a small sized index which contains only the terms on the basis of which matching of the terms are performed by the searcher component during searching. The hierarchical ranked indexing structure for a search engine index speeds up the search process by directing the search to a specific path from the higher level index to the lower level indexes. The proposed architecture and the structure of the indexes are given in next section with a view to reduce the search space and search time by maintaining the index at various levels.

It may be noted that the hierarchy of indexes has been created. The index formation starts at the lowest level where each cluster has a separate index consisting of the terms contained in the similar links contained within that cluster. This is called as descriptive index which is stored in the form of inverted files. The index consists of an array of posting lists as in case of global large size index.

5. RESULT AND ANALYSIS

The implementation and experiments of the BioSearch over several datasets are performed and the results are described below. Some links may provide complete extraction availability and some links are limited to the extraction process. Proposed system is implemented in Asp.Net with C#.

Data Sets: Pages from biomedical websites are the primary sources of datasets used for the experiments. The following datasets are used in the experiments to compare the performance of BioSearch with the existing methods. The following dataset contains some medical websites available for clinical data searches namely webMd.com, medicinenet,com etc.,

The input of the experiments can be any kind of web URL or page source. For example if a need to obtain an information from webMd website, keywords such as disease, medicine, pathology information's are matched can be given as input and information can be obtained from the source site. The extraction of links, tags, and values with query terms are matched and implemented. Extraction process is performed by the admin as pre-fetching process, the followings are the URLs used as a dataset source.

Table 1.0 dataset information's

Parameter	Value
Total main websites used	20
Website links	https://www.webmd.com/default tm http://www.medicinenet.com/ et

Like the above, several links have been used for experiments. The system initially extracts the page source by applying the extract html function which is coded using C#.net. The web page source and the links are extracted from the medicinenet.com website. The web source may have more unwanted page contents, which the user doesn't wants. The data units in some composite text nodes are separated by blank spaces created by consecutive HTML entities like " " or some formatting HTML tags such as <SPAN>. Second, the links from the website may contain some general and unwanted sub links. After elimination, the important keywords form the link will be passed for the content extraction process. The extracted URL and the description about the URL is given the figure 1.0.

	Bid	Bname	Burl	Bdec
1	1	Blood Cancer	www.medicinenet.com	Blood cancer is...
2	2	Heart Attack	www.medicalnewstoday.com	Heart muscle does not have enough blood ...
3	2	Heart Attack	www.ncbi.nlm.nih.gov	A heart attack occurs when blood flow to ...
4	2	Heart Attack	www.en.wikipedia.org	Myocardial infarction...
5	2	Heart Attack	www.nhlbi.nih.gov	A heart attack occurs if the flow of oxygen-rich bl...
6	2	Heart Attack	www.emedicinehealth.com	The heart is a muscle like any other in the body...
7	57	allergies	http://www.medicinenet.com/dis...	An allergy is a reaction of the immune system. Th...
8	58	arthritis	http://www.medicinenet.com/dis...	Arthritis, according to the Centers for Disease Con...
9	59	asthma	http://www.medicinenet.com/dis...	According to the Centers for Disease Control and ...
10	60	cancer	http://www.medicinenet.com/dis...	The term cancer refers to many different diseases...
11	61	cholesterol	http://www.medicinenet.com/dis...	Cholesterol is a waxy, fat-like substance that occ...
12	62	depression	http://www.medicinenet.com/dis...	Depression refers to: Economics: a period of eco...
13	63	diabetes	http://www.medicinenet.com/dis...	Diabetes has more than one meaning. As such, t...
14	64	digestion	http://www.medicinenet.com/dis...	Digestion is the process by which an organism br...
15	65	eyesight	http://www.medicinenet.com/dis...	Eyesight is the vision that is granted by the eyes ...
16	66	heart	http://www.medicinenet.com/dis...	
17	67	migraine	http://www.medicinenet.com/dis...	A migraine is a form of headache. Migraines tend ...
18	68	neurology	http://www.medicinenet.com/dis...	Neurology is the medical specialty concerning dis...
19	69	pregnancy	http://www.medicinenet.com/dis...	Pregnancy describes the condition of a mother ca...

Figure 1.0 extracted links with its description

In the above, data are extracted from the source by means of providing the key terms to the other website for example "cancer", this will extract the description about the disease. From the given link the system extracts the page source, which contains the html and other tags. The following example shows the extracted data from the given URL. The eliminated links form the extracted url list are given below figure 3.0.

```
http://www.medicinenet.com/image_collection/article.htm

http://www.medicinenet.com/quizzes_a-z_list/article.htm
```

Figure 2.0 extracted links

The extracted results will be summarized by highlighting the key terms. Based on the details such as description, links, the relevant contents can be identified. In case of cancer, the description of cancer and related links for the cancer is summarized.

6. PERFORMANCE ANALYSIS

The performance of the proposed methods is compared in three different ways. General data set evaluation presents the performance on the first three data sets, which exhibit a variety of properties and have been used in previous work by others. The other two evaluations focus on specific properties of the query result pages. Noncontiguous search query evaluation compares the performance for query result pages in which the existing systems are used. This experiment tests on two URL data sets with 9 diseases. The medicine net and webMD websites each of which contains 50-60 web links in average per disease. The website tests on multiple entries like the above. More than 5 results of these pages that contain a single data record are used for extraction. For each web database, 10 result pages are collected after manually submitting 10 different queries via its query interface.

Table 2.0: Comparative Study of Existing Data Extraction and search tools with BioSearch

Techniques and features	**BioSearch**	**InvIndex**
Middle out similarity	available	×
Single Result Page extraction	possible	possible
Non-contiguous Data Regions	possible	possible
Inverted Indexing	possible	×
Structure based priority	possible	×
Data summarization	possible	×

Two common measures, Recall and Precision, to evaluate the performance of this approach. Recall is the percentage of the number of data records that have been correctly extracted over the total number of data records on a result page. Precision is the percentage of the number of data records that have been correctly extracted over the total number of data records that have been extracted.

$Pr = Cc / Ce$

$Rr = Cc / Cr$

Where, Cc is the count of correctly extracted and aligned search results,

Ce is the count of extracted search results, and

Cr is the actual count of search results in the query result pages.

The number of search results in different query result pages varies from a few to hundreds. Consequently, pages with many search results will dominate the record level metrics. To use a page-level metric, namely page- level precision defined as,

$Pp = Cp / Na$

Where, cp is the count of correctly extracted pages, which means that all the search results in the pages are correctly extracted and aligned; Na is the count of all the pages from which search results are extracted. To assume that each input page contains at least two search results and data extraction is performed on all input pages.

Table 3.0: Performance Analysis

Search Results	Websites			
	Medicinenet.com		WebMD.com	
	InvIndex	**BioSearch_RII**	**InvIndex**	**BioSearch_RII**
Extracted search links	500	330	350	340
Correctly Extracted search links	480	325	330	335
Record lev Precision (%	96	98.4	94.2	98.5
Record lev Recall (%)	90.2	95.4	96.3	99.7
Page lev Precision (%	80.5	91	93.8	95.1

As per theoretical comparison and proof from the current experiment setup, the comparison study has developed. The

proposed BioSearch shows better results, as a well known data record extraction system.

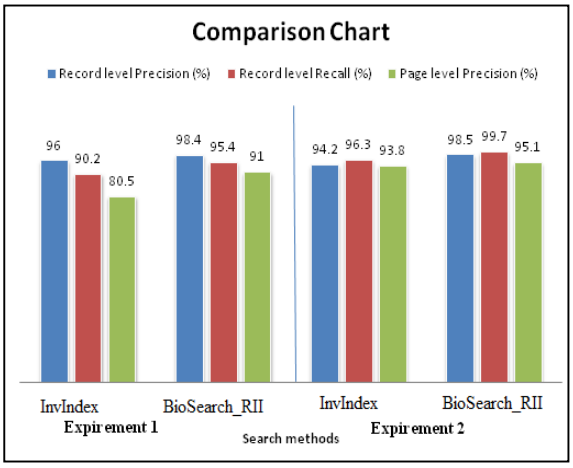

Figure 3.0 the accuracy comparison in terms of precision

The above chart shows the graphical representation comparing the performance of BioSearch with the existing InvIndex approach. As per Table 3.0, this approach has much better experimental results than existing approach InvIndex, and in almost every domain this approach significantly outperforms InvIndex. The precision and recall of this approach are both high across all domains, approaching 100%. This approach can also extract query result pages with single data records.

7. CONCLUSION

This paper presented a domain specific search engine with optimal query processing algorithm which is named as BioSearch, for extracting structured data from a collection of query result records obtained from the web pages. BioSearch first discovers the data regions from multiple pages and merges the data region that contains similar data results. Finally it summarizes the data values in search results by the following methods: query expansion, query pre-fetching. The proposed system is used to search clinical information's, which are very complicated to search in the normal search engines. This includes static pages as well as pages created by the web server in response to each user query submitted to it using web page forms. Data extraction from different medical related websites and summarizing with the relevant links using ranked inverted index is the main outcome of the proposed system. This allows extracting the desired links from the web pages and brings to the user based on their query. The medical terms are gathered and made as a cluster. Every cluster contains the relevant links and the desired description about the query. BioSearch extracts desired data from various search results pages. The experiments on several collections of search results, drawn from many well-known data rich sites, this indicates that BioSearch is extremely good in extracting and summarizing medical data from the web page sources. Another desirable feature of the proposed system is that it does not completely fail to extract any data even when some of the assumptions made by optional tag are not met by the input collection. In other words the impact of the failed assumptions is limited to a few attributes. BioSearch prone to provide higher accuracy compared with the existing methods. Finally the energy effective query search is gained using fast indexing methods.

REFERENCES

[1] Sahoo, Pradeep, and S. P. Rajagopalan. "An Efficient Web Search Engine for Noisy Free Information Retrieval." *International Arab Journal of Information Technology (IAJIT)*(2015).

[2] Duda, Cristian, Gianni Frey, Donald Kossmann, and Chong Zhou. "Ajaxsearch: crawling, indexing and searching web 2.0 applications." *Proceedings of the VLDB Endowment* 1, no. 2 (2008): 1440-1443.

[3] Sagayam, R., S. Srinivasan, and S. Roshni. "A survey of text mining: Retrieval, extraction and indexing techniques." *International Journal of Computational Engineering Research*2.5 (2012).

[4] Catena, Matteo, and Nicola Tonellotto. "Energy-Efficient Query Processing in Web Search Engines." *IEEE Transactions on Knowledge and Data Engineering* 29.7 (2017): 1412-1425.

[5] McCallum, Andrew, Kamal Nigam, Jason Rennie, and Kristie Seymore. "A machine learning approach to building domain-specific search engines." In *IJCAI*, vol. 99, pp. 662-667. 1999.

[6] Senthil Kumar, B., and Dr Gunavathi R. "A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis." *IJARCCE* 5 (2016): 463-467.

[7] Kumar, B. Senthil, and Asha Unnikrishnan. "A Survey on Effective Query Processing Techniques in Web Search Engines." *Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org* 7.9 (2017).