International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0063
ISSN (Online): 2279-0071

International Journal of Software and Web Sciences (IJSWS)
www.iasir.net

A Query Enhancement Technique for Extracting Relevant Information

Mansour Aldawood and Dr. Ahmed Z Emam
Information Systems Department
King Saud University
Riyadh, Kingdom of Saudi Arabia

Abstract: *The size of the contents on the web is exponentially increasing with the facilitating from modern technologies that eases creating and storing them. This evolving of database technology leads the researches and the specialists in a certain field to save their publications easily in these databases, based on their backgrounds and interests. With time passing, these databases size reach to a huge number of the information stored in them, which is making the search for an information is difficult. In Medical filed, a predefined database such as Medline has more than 23 Million scientific papers, which is care about the field of the healthcare. The extraction of a certain information in this huge database requires a technique that guides the searcher, to have the most relevant results that satisfy his need. This paper will introduce a technique that aims to enhance the query entered by searcher of an information by preprocessing the query to increase the degree of relevancy of the results. This technique aims to improve the results of the extraction tools that specialized in extracting information from healthcare database.*

Keywords: *Information Extraction; Medline; Query enhancement; UMLS; information retrieval;*

I. Introduction

The amount of medical information in health care databases has become exponentially huge during the past few years, with the aid of advanced technologies that helps and facilitate the process of easily storing this information. The huge amount of information that resides in these databases has made the search for relevant information in them more difficult than ever for physicians whom seek about a certain information [1]. One of the most famous healthcare databases is Medline [2], which considers the most comprehensive database that specialized in biomedical fields. Its information updated on a daily basis and has more than 24 million records that dated back from the early of the 60s. With this evolving in database size, the traditional searching tools didn't serve the physicians or researches in a proper way since these tools present a big number of results, which makes finding the relevant information taking a lot of time. This time consuming affects directly the decision-making in most cases [3]. This issue leads to the need of specialized extraction tools that mainly extracting a structure information from unstructured repositories based on the keywords that entered by the searchers. Even with existing of extraction tools, the number of results for a certain query about relevant information is still high, which makes finding the most wanted results for physicians or researchers not easy and taking a long time. This led to the need of query enhancer technique that help the physician to search about any information with certain keywords that give the most relevant results to the query, with taking to the consideration the existing extraction tools.

II. Literature Review

The most famous definition of the information extraction concept is the auto extraction of structured information from unstructured source or database such as the databases with the natural language format. The ultimate goal of this technique is to make any wanted information accessible through a certain mechanism and in a timely manner. In addition, most of the researchers believe that the use of information extraction tools will help to develop databases with structured information that will help to search and indexing the information that resides in them more easily. What motivates this technique to be appearing and developed over the past years, is the need for searching in wider pool rather than just searching on a certain keyword in a specific area. This concept has recognized in the late of the 70s and start to be develop widely in the 80s [4]. The Unified Medical Language System is a set of files and tools that contains many standards and vocabulary that related to the field of healthcare and biomedical in order to achieve interoperability between information systems that specialized in the healthcare sector. The most advantage of using UMLS is to have the ability to link between health information, medical standards and medicine names over multiple information systems. UMLS has created during 1986 as databases that about the vocabularies in the biomedical sciences. With the enormous increase of contents that specialized in biomedical resources in predefined databases, the retrieve for this content become more difficult and has a large volume of results. UMLS help solving the issue by enhancing the accessing to these contents by providing a mapping structure between vocabularies related biomedical and thesaurus of biomedical concepts [5]. Medline is the unite state, national library of medicine (NLM) which has over than 24 million scientific papers and journals in science but more concentrates on the biomedical one. Medline [2] has started its development in

1964 as a medical literature analysis and retrieval system (MEDLARS).The Porter stemmer algorithm according to his founder Porter [6] is a "process for removing the commoner morphological endings from words in English". In other word, this algorithm is responsible for retrieving English word to its original form to aid the developers in the field of information retrieval. Porter stemmer is a rule-based algorithm that aims to remove to the endings of English words, for example, the word relational will be relate according to porter stemming.The growth of contents in predefined databases has exponentially increased, therefore the need for tools that extract the exact what we looking for in large databases has become necessary. One of these tools called BioIE that has developed in 2005 by Divoli and Attwood [7], which is a rule based extraction tool that extract an informative sentence which are related to a protein families and their structure. In the same year, Mitchell and his team [8] stated that BioIE has the ability to perform in classifying sentences that related to disease more than a support vector machine (SVM) since it has 56% precision while in SVM 48% but when using other factor such as the sentence that related to the structure it precision become less. In 2006, Schuhmann and his colleagues [9] has developed a tool called EBIMed to be efficiently retrieve sentences or abstracts from the Medline database with ability to analyze these phrases. The extracted abstracts that used to create a table, which has an overview about the protein, gene ontology, drugs and species of the same biological context. Furthermore, in 2004 a BioRAT has presented by Corney and his team [10] as an information extraction tool. This software aims to extract a biomedical information and be able to locate and analyze either abstract or full papers. Additionally, in 2007, Hearst and his colleagues [11] has developed a web based search engine called BioText. From the chosen name, it indicates its purpose which is to help the biologists to have a new method to access the recent scientific papers by searching on articles illustrations, figures and captions. BioText is a web based application and still ongoing research papers with an interface that's designed carefully to serve the purpose of this engine and to accomplish its functionality. After a period of time, BioText developers [12] added more functionality to their search engine by allowing the users to have the ability to search over full text, abstracts, figure captions and tables. Later in 2008, Gladki and his colleagues [13] has represented a web based tool that extract the abstracts of a biomedical information in the PubMed database. Their goal was to design a tool that can find the right and true correlations of a search by the users. They present a software that has a useful functionality such as searching by author name and using logical operators such as AND, OR and NOT. Over the years, the researchers and developers start to think differently about the way that the search engines must perform. One of these ways, as it is presented by Wang and his team in 2010 [14] which using the fuzzy search technique in their search engine. They present a web-based tool called IPubMed that extract and search for publications from the Medline database. Their goal of this tool is to have the ability to retrieve instant exact feedback to searcher query plus to have the approximate result of the same query as a fuzzy result. IPubMed is the tool that utilized for the proposed system.

III. Research Question

As it mentioned in the introduction of this paper, the number of results for a certain query about relevant information is still high in the existing of extraction tools, which makes finding the most wanted results for physicians or researchers not easy and taking a long time. Which leads to the need of query enhancer technique that help the physician to search about any information with certain keywords that give the most relevant results to the query. This paper has two questions to verify and prove which are, how using UMLS biomedical Meta-thesaurus concepts and ontology will improve the searching quality for a physician and searchers. Second question is what the appropriate tool that integrate with UMLS to enhance the query performance. Moreover, it will verify about applying a proposed tool as an extraction tool with the proposed system (EQI- Enhanced Query for IPubMed) will improve the precision and quality of the search.

IV. Methodolgy

The proposed methodology starts by utilizing the existing extraction tool called IPubMed that responsible of query handling, document analysis and indexing to improve search results and ranking the most relevant information, abstracts or sites. The second step is using the Unified Medical Language System (UMLS) APIs to retrieve synonyms keywords that related to the healthcare domain. The third step is applying porter stemmer to reduce the entered keywords to their original form then removing the stop words from the query. The last step is using the results keywords from UMLS as a query in EQI and PubMed Search engines to compare their precision and recall using confusion matrix technique.

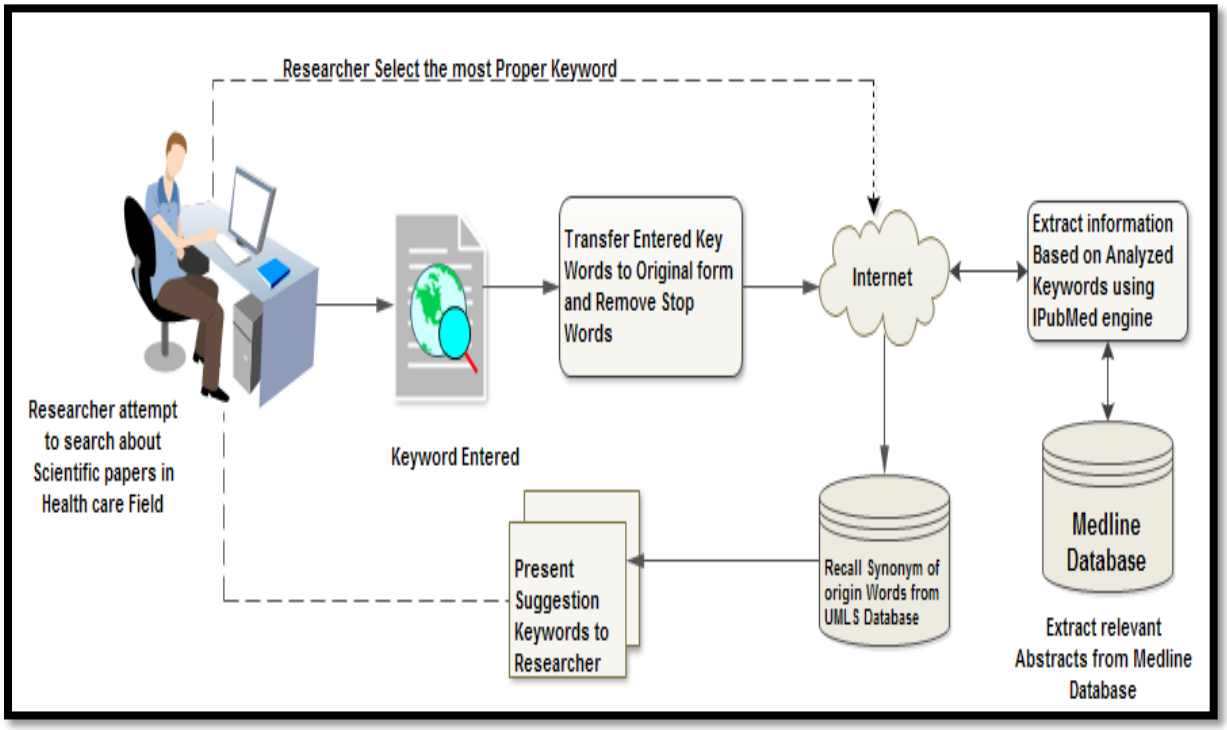
Table 1: Synonyms keywords for stomach cancer query from UMLS

Malignant neoplasm of stomach	Malignant neoplasm of stomach stage IV	cellular diagnosis, gastric cancer	FH: Stomach cancer
Stomach Carcinoma	recurrent gastric cancer	stage, gastric cancer	Stage I Gastric Carcinoma
Stomach Neoplasms	Gastric Adenocarcinoma	Gastric Fundus Carcinoma	Stomach Problem
Carcinoma in situ of stomach	intestinal adenocarcinoma of the stomach	Gastric Body Carcinoma	Endoscopy of stomach

V. Proposed System

The proposed system running through two phases, which are the preprocessing phase for analyzing the entered keywords, and the second phase is the extracting phase, which responsible for extracting the information form predefined repository.

Figure 1: Architecture of the proposed system



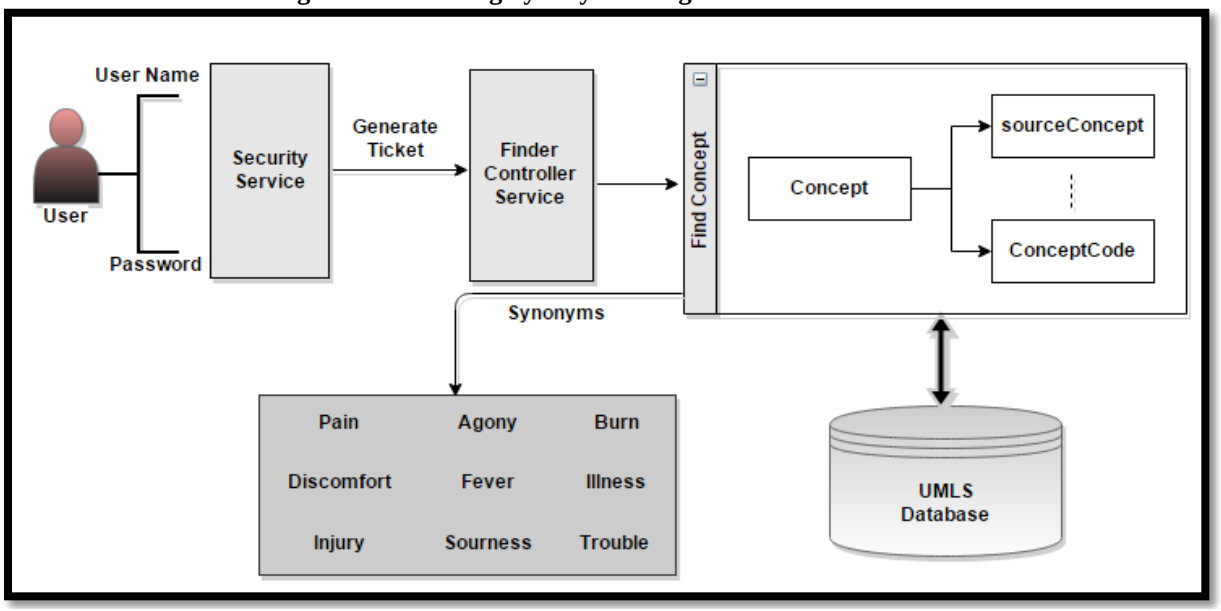
The first phase will be responsible for analyzing the entered keywords that entered by the physician before integrate them with the extraction tool. Firstly, the physician or researcher will type keywords through a web-based interface to have an information based on the entered keywords. The preprocessing client will remove the stop words from the query to end up with the terms that mostly wanted and which will yield to better results, also will do the stemming of the entered keyword to retrieve it to its origin form using a stemmer algorithm to improve the precision of the extracted information. After that, the client will start communicating with the UMLS database to have synonym words. The client will send the entered keywords to the UMLS database through its APIs to check whether these keywords exist in in UMLS repository or not. If these keywords do not exist, it will return an empty result to the client, which in its turn will present a result to the physician, indicates that his or her searching words do not have any synonym words and the physician should return to enter a new keyword. If these keywords have a synonym word, then it will retrieved from UMLS database and send to the client in order to present them to the physician. In this stage, the synonym words will presented to the physician by the analysis client and in his or her turn will choose the most relevant words that is close to what he or she looking for. After selecting the relevant word, the analysis client will send these words to the extraction tool for retrieving the information based on the entered keywords. The second phase is responsible for extracting the information from the repository. It utilizes a predeveloped extracting tool that has integration with Medline repository. This extraction tool received the analyzed keywords from the client in the preprocessing phase, and extract the information from the repository to present them to the physician.

Retrieving Synonyms using UMLS API Design

Unified Medical Language System (UMLS) database has rich synonyms for the words that related to healthcare fields. This proposed system uses this database facility through its provided API that will explained in details in the next illustration. As it shows in figure 2, the communication with the API requires prerequisites requirements that permits the establishment of transfer the information between the proposed system and UMLS database, which are essentially the user name and password. The user name and password validity is essential to create the one time per eight hours ticket that communicates with UMLS database. This ticket generated by a security service that firstly check for the validity of the user name and password, and then generates the single time ticket. After the ticket generated, the communication with UMLS database will be open with its synonym words through the provided API. The provided API by UMLS terminology services has many features that used for useful needs. In this proposed system, retrieving the synonyms of the entered keywords is the important feature that this system seeks. A function used to retrieve the synonyms that provide the choice for searching in the database by either

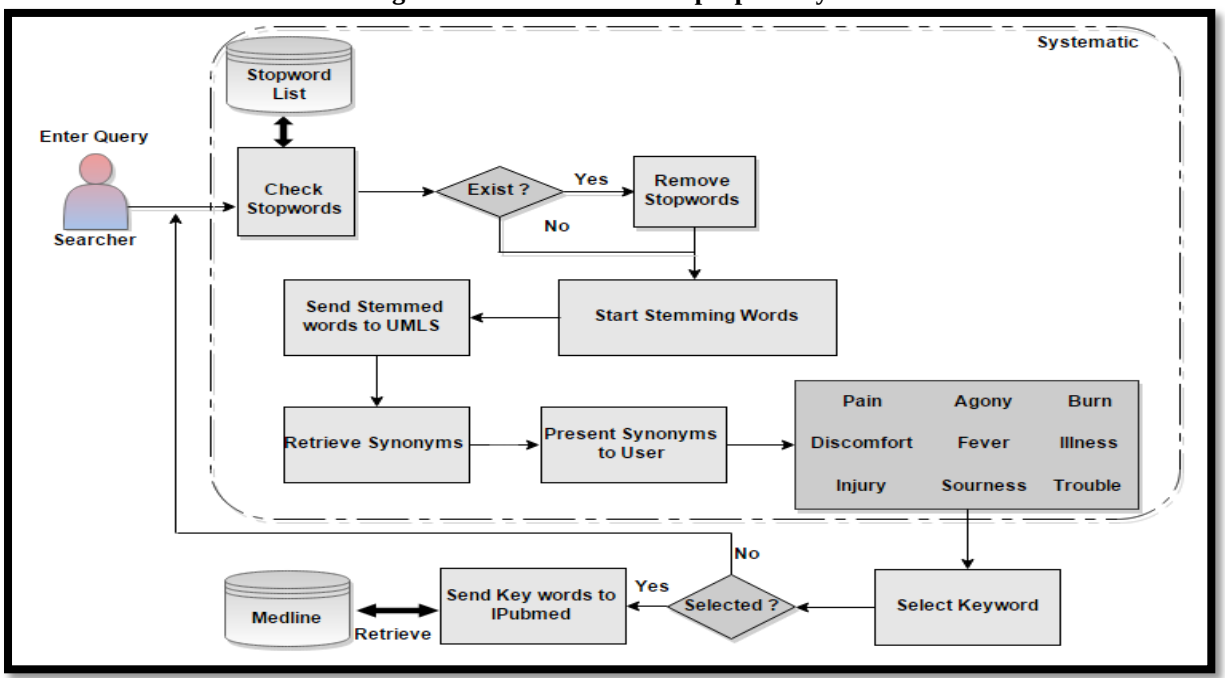
exact, approximate or normalized words. Moreover, this function has the ability to retrieve either the synonym concept, the source of the concept or the code of the concept. After this customization of the search, the result will be stored in an array and presented.

Figure 2: Retrieving Synonyms using UMLS API Architecture



The main goal of the proposed system is to enhance the query that enters the search box for yielding to better and relevant results for the user query. In this proposed system, the deletion of stop words, stemming of the keywords in its original form and retrieve the synonyms from UMLS database is an automated process. The searcher, who will utilize this system, will just enter the query that satisfies his need to find any information related to his field, which is in our case in the healthcare field. The query will start the refining process by firstly check for the stop words if the query has them. Once the stop words detected, the system will delete them and send the results to the second refining process, which is stemming process. In the steaming process, the query will be diving into tokens of word to be easily reduce them to their original form. After the stemming process completed for each token, the query will combined as whole query. Thirdly, the query will sent to the UMLS database to check if there are synonyms words in the query. If the synonyms exist, it will retrieved to the searcher as suggestion keywords. As it shows in Figure 3, the user will select one of the suggested keywords if it`s satisfy his or her information needs. These selected key words will sent to as keywords, to the predefined search engine, which is IPubMed in order to extract the needed information from the Medline database. IPubMed engine is a well-developed engine that responsible of retrieving the most relevant information that related to healthcare.

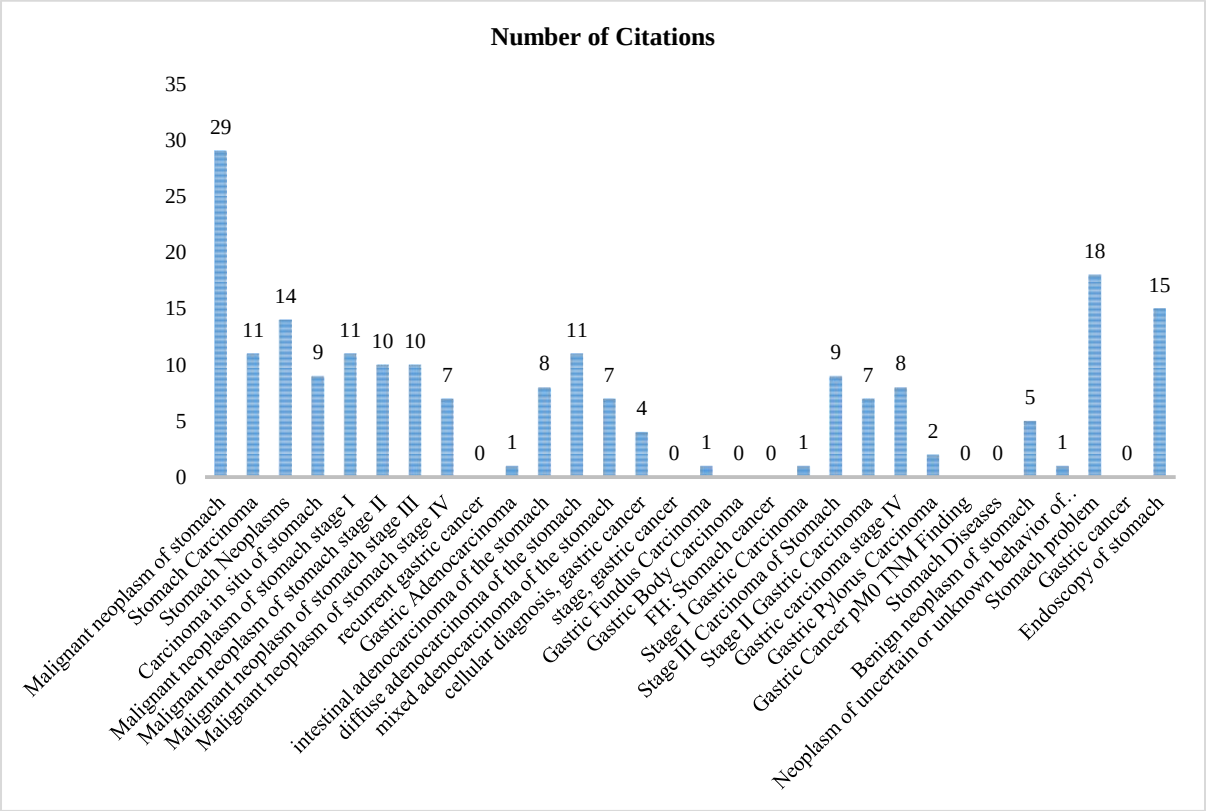
Figure 3: Process flow of the proposed system



VI. System Validation and Results

As Figure 4 illustrates, the query "Stomach Cancer" has used in the proposed system API to retrieve the synonyms words from UMLS database. Thirty-one results have received from the query with different synonyms that used in the healthcare field regarding stomach cancer. These keywords will integrated into the EQI search engine to count the number of scientific papers that has the synonymous keywords and the original keywords "Stomach cancer" in the same paper.

Figure 4: Number of Citations per Synonym word



Confusion matrix technique used to facilitate measuring the performance of the proposed system, since it offers a classification that leads to measuring the precision, recall, error rate, accuracy and F-measure of the presented technique. Confusion matrix techniques divide the current class that aimed to measure to four classifications as shown in Table 2.

.Table 2: Four classes of confusion matrix

Confusion matrix		Predicted Class	
		+	-
Actual class	+	**True positives (++)**	**False negatives(+-)**
	-	**False positives(-+)**	**True negatives(--)**

A set of results (hits) has retrieved when applying the most ranked synonym words which is" Malignant neoplasm of stomach" on the two search engines. One-hundred hits has tested from the two search engines and the following results based on this token sample. After analyzing the two results of the two search engines, which are EQI and the PubMed on 100 extracted documents for the queries, which are, query1 "Malignant neoplasm of stomach", query2"Stomach problem" and query3"Stomach Neoplasms". The next tables show the results of four classes of the confusion matrix as shown in Table 3.

Table 3: Performance results for the search engines

PubMed	**Performance classes for PubMed (Total=100)**			
Query/Class	TP	FN	FP	TN
Query1	37	9	22	32
Query2	28	6	19	47
Query3	33	14	23	30
EQI	**Performance classes for IPubMed (Total=100)**			
Query/Class	TP	FN	FP	TN
Query1	19	7	28	46
Query2	11	5	26	58
Query3	21	6	24	49

Table 4 Equations of precision, recall, error rate, accuracy and F-measure

Measure	Equation
Accuracy	(True positives+ True negatives)/Total Documents
Error rate	(False positives+ False negatives)/Total Documents
Recall	(True positives)/ (True positives+ False negatives)
Precision	(True positives)/(True positives+ False positives)
F-Measure	2*((Precision* Recall)/ (Precision+ Recall))

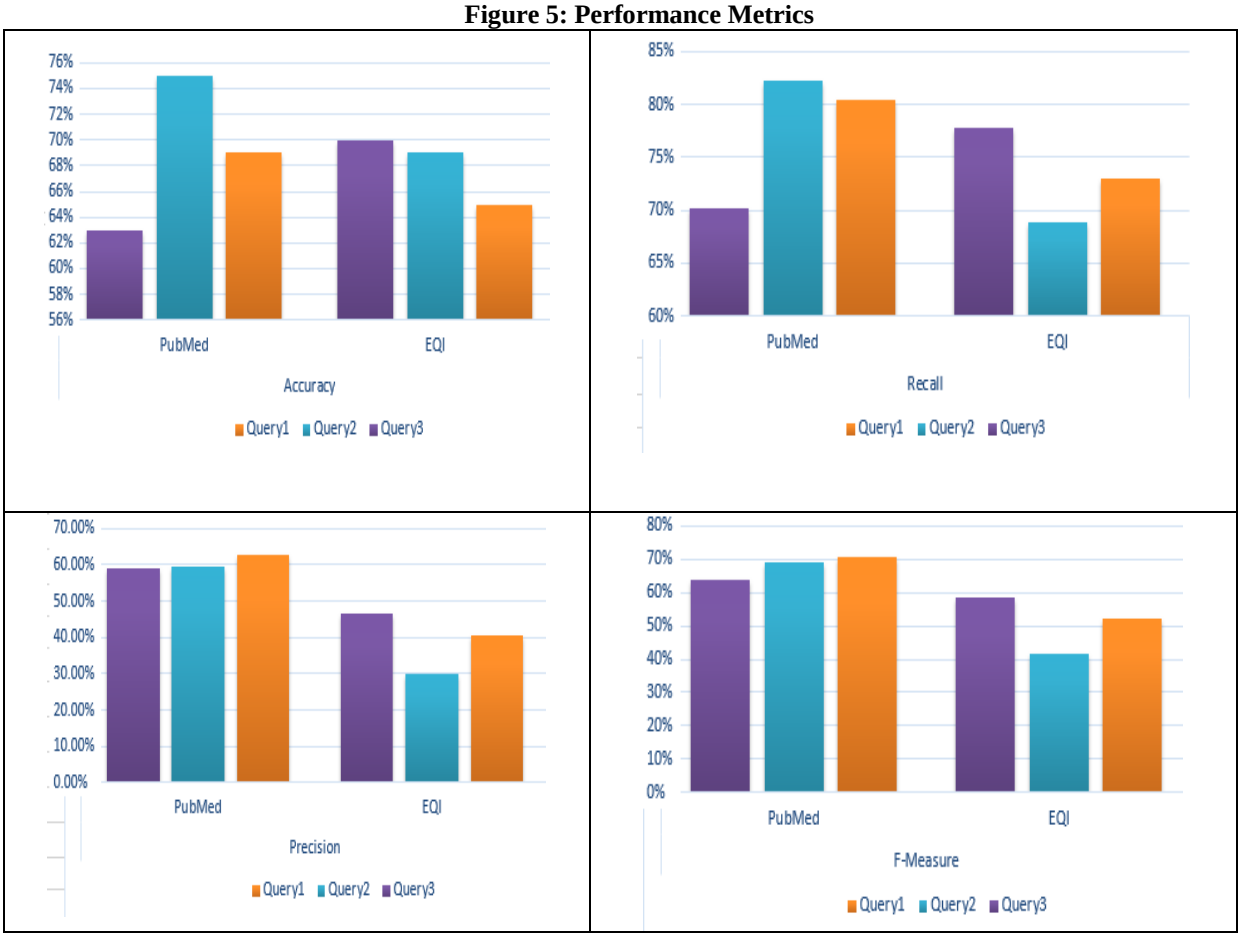
From the literature review, the suggested performance evaluation metrics for similar system are precision, recall, error rate, accuracy and F-measure can achieved as shown in Table (4). Table (4) shows the equation that used for measuring the precision, recall, error rate, accuracy and F-measure. The results will be as shown in table5.

Table 5: Calculation of precision, recall, error rate, accuracy and F-measure

EQI	Accuracy	Error rate	Recall	Precision	F-Measure
Query1	65%	35%	73%	40.4%	52%
Query2	69%	31%	68.8%	29.7%	41.4%
Query3	70%	30%	77.8%	46.7%	58.3%
PubMed	Accuracy	Error rate	Recall	Precision	F-Measure
Query1	69%	31%	80.4%	62.7%	70.4%
Query2	75%	25%	82.3%	59.6%	69.1%
Query3	63%	37%	70.2%	58.9%	64%

To evaluate the efficiency of the proposed system, the generated four three evaluation metrics as described in table 4 are measured. We have implemented this evaluation model using the proposed system and the comparison among the three quires by each metric is shown in figures (5).

Figure 5: Performance Metrics



VII. Conclusion and Future work

According to the results mentioned earlier in the previous chapter, a set of recommendations has assembled in the matter of the proposed system. Using the proposed system has a value to the searcher of physicians whom facing a difficulty in finding the right query that leads to the most relevant information. Therefore, in any

information retrieval system the enhancement of the query is must in order to have relevant results. Enhancing the query is not enough, the ranking mechanism is important as well for retrieving the most relevant information. Because with existing of a good enhancing technique, but with a less quality ranking technique, the results will not be as expected to the user. The information retrieval system that has used in the two systems, which are EQI and PubMed, uses a different mechanism of ranking the results; therefore, the results of the recalled elements in PubMed are much higher than the recalled elements in EQI. Regarding that, the two search engines reading their results from the same database. Moreover, measuring of the relevancy of the results for EQI is depending on the term frequency that mentioned in the abstract of a document, which means if the document has the term of the query repeated many times, the document will ranked at the top and so on. On the other hand, PubMed uses different kinds of algorithm that measuring the relevancy that combining more than one factor to rank the results, the factors such term frequency of the whole document, the date of the publication. Therefore, the precision of PubMed is higher than EQI. The accuracy and error for the two search engines considered reasonable, since they have different way of retrieving the information from Medline, with the regards that the results that collected from PubMed considered more reliable to the query than EQI, since it uses term frequency for the whole document for scoring and ranking the document.

As a future work of the proposed system, the proposed system could be used more components for enhancing the query than the ones were presented. For instance, adding the tokenization component that responsible of split the query into a group of words in order to for stemming them or treat them as query individually. This approach will take one word at a time and prepares it for synonyms as the following illustration shows. Moreover, automatic annotation to enhance the query can added, which is simply caring about the metadata for the query not just the query itself. For instance, if a searcher wants to search about a paper with knowing the author of the paper only, a result of suggestions author lists can be helpful for the searcher. The synonyms that retrieved from UMLS database could cause a noise to the proposed system. Because not all the synonyms words, that has been collected from UMLS are usable as a keywords in the scientific papers. Moreover, from 200 citations that been ranked according to their relevancy to the user query, the synonyms were not used and this causing a noise to the proposed system. As future improvement, a filter technique could added to the query enhancement system to filter the unusable keywords and to reduce the number of suggestions to the most relevant ones. The Porter stemmer that used in this proposed system is a general technique that reducing the keywords in their original form based on the English dictionary words. An enhancement could added to the system by adding a porter stemmer that specialized in reducing the medical keywords to their root form. This enhancement could raise the quality of the retrieved suggestions to the user who seeks about the relevant documents.

VIII. References

[1] Ebbert, J., Dupras, D. and Erwin, P. (2003). Searching the Medical Literature Using PubMed: A Tutorial. Mayo Clinic Proceedings, 78(1), pp.87-91.

[2] Ncbi.nlm.nih.gov, (2014). Home - PubMed - NCBI. [online] Available at: http://www.ncbi.nlm.nih.gov/pubmed [Accessed 25 Dec. 2014].

[3] Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A. and Mazo, I. (2004). Extracting human protein interactions from MEDLINE using a full-sentence parser. Bioinformatics, 20(5), pp.604-611.

[4] Pazienza, M. (1997). Information extraction. Berlin: Springer.

[5] Selden, C. and Humphreys, B. (1997). Unified Medical Language System (UMLS). Bethesda, Md. (8600 Rockville Pike): U.S. Dept. of Health and Human Services, Public Health Service, National Institutes of Health, National Library of Medicine, Reference Section.

[6] Porter, M. (1980). An algorithm for suffix stripping. Program: electronic library and information systems, 14(3), pp.130-137.

[7] Divoli, A. and Attwood, T. (2005). BioIE: extracting informative sentences from the biomedical literature. Bioinformatics, 21(9), pp.2138-2139.

[8] Mitchell, A., Divoli, A., Kim, J., Hilario, M., Selimas, I. and Attwood, T. (2005). METIS: multiple extraction techniques for informative sentences. Bioinformatics, 21(22), pp.4196-4197.

[9] Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M. and Stoehr, P. (2007). EBIMed--text crunching to gather facts for proteins from Medline. Bioinformatics, 23(2), pp.e237-e244.

[10] Corney, D., Buxton, B., Langdon, W. and Jones, D. (2004). BioRAT: extracting biological information from full-length papers. Bioinformatics, 20(17), pp.3206-3213.

[11] Hearst, M., Divoli, A., Guturu, H., Ksikes, A., Nakov, P., Wooldridge, M. and Ye, J. (2007). BioText Search Engine: beyond abstract search. Bioinformatics, 23(16), pp.2196-2197.

[12] Hearst MA, Divoli A, Ye J, Wooldridge MA. Exploring the Efficacy of Caption Search for Bioscience Journal Search Interfaces. 2007. in Proceedings of BioNLP 2007, a workshop of ACL 2007.

[13] Gladki, A., Siedlecki, P., Kaczanowski, S., & Zielenkiewicz, P. (2008). e-LiSe - An online tool for finding needles in the "(Medline) haystack". Bioinformatics.

[14] Wang, J., Cetindil, I., Ji, S., Li, C., Xie, X., Li, G. and Feng, J. (2010). Interactive and fuzzy search: a dynamic way to explore MEDLINE. Bioinformatics, 26(18), pp.2321-2327.

[15] Chen, C. (2013). FEATURE SELECTION BASED ON COMPACTNESS AND SEPARABILITY: COMPARISON WITH FILTER-BASED METHODS. Computational Intelligence, 30(3), pp.636-656.