A Query Recommendation System for Efficient Biomedical Information Retrieval

Jagan S.[1]* and Rajagopalan S.P.[2]

1. Department of Computer Science and Engineering, Adhi College of Engineering and Technology, Kanchipuram-631 605, Tamil Nadu, INDIA
2. Department of Computer Science and Engineering, GKM College of Engineering and Technology, Chennai-600 063, Tamil Nadu, INDIA
*jaganshanmugam83@gmail.com

Abstract

In electronic arrangement accessible biomedical information is hastily increasing. Actually, medical information of huge collections for the investigators, health care suppliers and all customer types includes visual and textual data. Improving the provided service through a search engine is much useful by finding a similarity measure among queries. To progress the user search knowledge via search engines has been frequently used by the users to interact information. Suggesting query information which is equal to user query is the Query recommendation and it concerned to improve recovery performance as a method. An outcome with an essential tool for query log mining is the Query Flow Graph (QFG). To represent a subject with general tools are Ontologies. Here, a QRS has been essential in biomedical information retrieval. Dataset like Image CLEF 2005, IMDB and 4 universities are concerned for experiments. Image CLEF 2005 medical image retrieval task presents a high-quality test proposal to assess the image retrieval technologies capability. Consequences obtain the better accuracy.

Keywords: Biomedical information, Query Recommendation, Query Flow Graph (QFG), Ontologies, ImageCLEF 2005, IMDB and 4 universities.

Introduction

In recent times to obtain various information's web has been used extensively. Numerous trainee users have complexity to acquire the preferred information though they used well-organized search engines like yahoo, Google. Dissimilarity exists among search engine and recommendations systems as the users use search engine to know their query in suitable for the essential information.

On the contrary user desires the recommendation system but to find the query from where accurately to obtain query solution with suitable wording they do not know. Hence the query recommendation system assists users to explain their facts which need further obviously in order that search engine can revisit proper answers to meet the required information. Recently to assist reformulate queries and gratify the required rapidly web search engines offered users with query recommendation. It recommends suitable queries for the search engine users only if they not pleased with the initial input query outcomes, hence helping users in civilizing search value. [1] Here exists an issue to improve the outcomes of the search engine to acquire the needed information from web due to the common factors such as: In a search engine, the indexed pages are grown hasty, short and vague queries proposed via web users, unproductive search results organization, user's various goals and web prospects etc.[2]

In study of computer science, Information Retrieval (IR) [3] submits to find text documents to satisfy required information from computers. IR can focus various data kinds and information crisis beyond the core definition specified above. The word "unstructured data" denotes the non-clear data, semantically explicit, simple computer formation and it is the contradictory to structured data, the canonical example of relational database of the sorted companies to preserve product inventories and personnel records. These days, many people employ in information retrieval each day using a web search engine like Google or Bing. Information retrieval is rapid information access form by overtaking conventional database-method searching.

The constant raise in the existing biomedical information sum has resulted in a huge demand on biomedical IR systems. It helped the researches to stay on current literature; numerous existing search systems lean to be either too limiting or too wide accuracy. Intended for this cause here needs search systems, particularly as regards in retrieval performance to improve their accuracy and recall. In IR point of view, there exists an issue in retrieving biomedical information. Absence of broadly distinguished terminology standards is the one major problem with biomedical information. New names and terms are generated each time a environmentalist determines a novel gene or some significant organic entities, and frequently inconsistent typographical/lexical variants exists.[4]

Friendly user interface is the key factor in today's Web search engines. Certainly, search engines permit the users to stipulate queries merely as keywords lists; next the traditional information retrieval systems approach.[5] Keywords represents to wide topics, to technological terms, or still to appropriate nouns to direct the search procedure to the pertinent document collection. In spite of easy interaction method in hunting the Web has been proved to be victorious, a keywords list is not a fine descriptor forever with the user information needs. To devise efficient queries it is not simple for the users forever to search engines is because of the increasing uncertainty in many language.

*** Author for Correspondence**

Queries having uncertain terms can recover documents that are not used for searching. Conversely, users naturally present extremely small queries to the search engine which are uncertain.

The demand in the IR systems needs the capability to compare big data sets systematically by means of all the knowledge from the issued data to allow the biological data set consequence to be construed. The number of article and journal information issued are rising at a substantial rate, hence it is not possible to maintain up to date for a researcher with the related literature physically on focused topics.

Ontology offers shared words to form a domain of objects and concepts with its properties and relations. To organize information as a framework to be used in artificial intelligence, the Semantic Web, systems engineering, software engineering, biomedical informatics, library science, enterprise bookmarking, and information architecture as knowledge representation method concern the world or part. The domain ontology creation is basic in definition and to use endeavor structural design. [6]

MeSH [7] issued by the National Library of Medicine includes controlled vocabulary and a MeSH Tree. The controlled vocabulary has various terms like descriptor, qualifiers, publication types, geographic, and entry terms. Descriptors and Entry terms are essential in the projected indexing method. In ontology Descriptor terms are the significant approach and headings. Also to the descriptors, the synonyms and its related terms are the Entry terms. MeSH descriptors framework as a MeSH Tree can be viewed as a MeSH Concept Hierarchy. It is a managed vocabulary described by the National Library of Medicine (NLM). MeSH ontology is intended initially to interpret and guide the MEDLINE database article , a corpus of abstracts with biomedical journal articles, however it has used in all types of NLM-produced database books, documents and general library indexing. In other areas also it has widely used such as information retrieval especially for the medical documentation and specialist recommendations

Related Works

Cui et al [8] projected a novel method to query about extension based on query logs. The main proposal was to remove probabilistic relationships among query terms and document terms by means of investigating query logs. These relationships were essential in selecting higher-quality extension terms for novel queries. The investigational outcomes had showed that log-based probabilistic query extension system had improved the search performance and had numerous benefits over the accessible systems.

Ramampiaro and Li [4] applied an information retrieval system as a series in searching biomedical information in an optimal way to merge them. This system consists of expanding extending entrenched IR similarity forms like Vector Space Model (VSM) and BM25 and its underlying scoring formats. It permitted the users to interrupt the ranking in accordance with their relevance view and also it had executed and tested in a prototype named BioTracer, it expands a Java-based open source search engine library. The outcomes in this experiment by with TREC 2004 Genomic Track collection were promising. Our examination had divulged the user search process would had positive results on search results ranking and the concepts can necessary to attain required the user’s information in BioTracer.

Taschwer [9] projected a scheme of Image CLEF medical case retrieval task had text-only retrieval with exploiting the Medical Subject Headings (MeSH) ontology. MeSH terms were extorted to use query extension and term weighting from the query. MeSH footnotes of documents accessible from PubMed Central were added to the corpus. Retrieval outcomes had improved somewhat on full-text retrieval.

Bordino et al [10] utilized the information in query logs to expand semantic similarity a measure among queries. Proposed concept relied on query-flow graph based upon the query log representation. The query-flow graph combined query reformulations from several users: nodes symbolized queries in the graph, and two were attached when they were emerged as part of similar search goal. By means of projecting the graph, query-similarity measure was attained on a low-dimensional Euclidean space. This experiment showed the captured a semantic similarity notion among queries and it had been essential to diversify query suggestions.

Baeza-Yates et al [11] suggested a system of search engine for a given query submitted with a list of related queries. The associated queries were based on earlier published queries by the user to the search engine for tuning and redirecting the search procedure. This proposed method was based on a query clustering procedure with identified semantic similar query groups. The clustering process utilized the users historical preferences content recorded in the search engine query log. This system not only determined the associated queries, but also ranked them in accordance with the condition. Lastly, it had shown the efficiency with the experiments over the search engine query log in this system.

Cheng et al [12] described the systems and investigational outcomes used for the medical retrieval task at ImageCLEF 2005. The competition topics included semantic queries and visual queries together. The content-based concept had four image features and the text-based concept used word expansion was expanded to achieve the mission. The investigational consequences had showed text-based concept had high accuracy rate rather than content-based concept. Additionally, the content-based and text-based concepts effects were better when combined rather than one of the concepts. Summarized that the contemplation on visual image queries can provided further human semantic perception and improved the medical image retrieval efficiency.

Sisode and Patil[13] appraised and compared various query log processing methods for information retrieval. Furthermore the clicked snippets concept was better in understanding users' interaction procedure along with search engines to discover the suitable information need.

de Campos et al [14] expanded a query expansion method on Bayesian networks. By a learning algorithm, a Bayesian network had been constructed to represent some relationships between the terms materializing for a known document; this network was then used as a vocabulary. The attained outcomes were reported based on three basic test collections in this method.

Btihal El Ghali et al [15] extracted the user query environment to utilize in its query recommendation method. Here, three different query recommendation methods were proposed, and compared on the extorted environments quality, by means of calculating the Average Internal Similarity (AIS) in each constructed environment. The effects showed that the document information had significant manipulate on the similarity among queries over the existing information for all projected concepts. The final experiment had comparison among the three methods, and it showed that AIS had higher value for short and long queries through TLM concept using Language Models which was based on ordinary terms and appropriate documents.

Methodology

Figure 1 shows the graphical representation of an information retrieval system.

Dataset Used

Universities Dataset: This data set has WWW pages gathered from computer science departments of different universities in January 1997 through world web knowledge base project in CMU text learning group. The 8282 pages are physically classified as: Student (1641), faculty (1124), staff (137), department (182), course (930), project (504), other (3764). The class other is a group of pages that were not deemed the "main page" to represent an instance previous six classes.

Image CLEF Dataset: In Image CLEF 2005, medical image retrieval task includes 25 queries for assessment, the queries combined visual image and semantic textual in retrieving desire images. The visual queries use image example for finding similar images on all topics of at least one image example. The semantic textual queries permits user query by a sentence of some semantic concept which are tough to obtain directly from images. The purpose of this task is to observe the visual feature to improve the query effect. [12]

In Image CLEF 2005, the evaluation of medical image collection has gray and color images. In color images, users are attracted by the color change more than the objects positions. Hence, in color image query the effective feature is diverse from a gray image query. An evident feature image is either the image is color or gray value. For instance, when the user queries an image, it determined either color or grayscale.

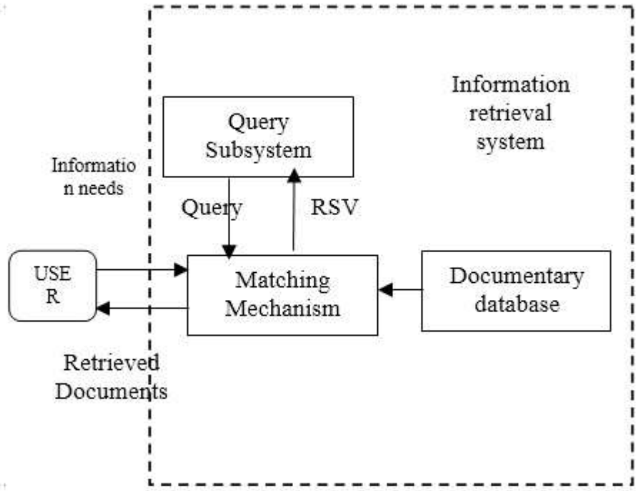

Figure 1: Graphical Representation of an Information Retrieval System

IMDb Dataset: An online information database with films, television programs, and video games is the Internet Movie Database (IMDb). It also includes the featured actors, production crew, and fictional characters. As an online subset which contains open available dataset with downloading and processing offline. From the IMDB an English dataset has 25k training and 25k testing examples with positive or negative sentiment labels. Also here exists further 50k unlabeled reviews and all these reviews are convoyed with its related URL movies.

The IMDb web site provides a dataset of information on cast, crew, titles, technical details, and biographies. This dataset has been organized into compressed text files sets in which each one has specific aspect of information. A description has been provided in the dataset within the file tools and movie-database-faq at the top of each file. The unrefined IMDb dataset has not an easy in progression, hence we recommend the cleaned data set to use above dataset when possible. [16]

Query Expansion: Investigation on routine query extension was already been under way by 1960 if initial appeals were enlarged on the statistical evidence grounds. The proposal was to attain related documents furthermore by expanding queries on the co-occurrence terms. Then, the co-occurrence index terms was the only condition relevance feedback absence. Though, this kind of routine query expansion has not been successful. The effective retrieval of the expanded queries were often no greater or less than the original queries.

The log-based query expansion systems have three significant properties. First, the early retrieval phase is not required any longer as the correlation term can be pre-computed offline. Second, the correlation term can be reflected most users preference, as query logs includes query sessions from several users. For instance, when the majority users use “windows” for searching information regarding Microsoft Windows product, “windows” this term will have stronger correlations with the words like “Microsoft”, “OS” and “software” other than the terms like “decorate”, “door” and “house”. Hence the query expansions will consequence in superior ranks in documenting about Microsoft Windows. Alike idea has been employed in different accessible search engines, like Direct Hit. [18]

Query Recommendations: The modern majority search engines have some automatic query recommendation to suggest novel queries that was related to the existing user’s mission and query- log massive information had been suggested for this purpose. Here query recommendations were obtained as a query flow graph application. Query recommendation has been a best method to help the users in satisfying their needs by means of query suggestion to the existing users via managing query log processing files, historical navigation patterns, updating the records of query processing, dynamic and static log data, clicked snippets, and so on.

Recommending appropriate past queries to an input query, each queries has been represented with a weight term vector and a document-weight vector. Next, the similarities among the input query and past query are measured by term vectors and document vectors. In the subsequent step, the related interest query is calculated by their rank to the input query, by normalizing and combining the term-based similarity with the document-based similarity. Finally, the past queries extorted from the search engine’s log are classified in accordance with the new user query. The purpose of the measuring similarities among the queries, symbolized as a document vectors to search for queries to have general relevant documents. In the other way, to compute similarity among queries with term vector representation is to search for queries for significant common term numbers to concern the related queries based on their terms they contain.

The query-flow graph: A query-flow graphis a directed graph G = (V,E,w) where:

- V = Q∪{s,t} is the set of distinct queries Q submitted to the search engine plus two special nodes s and t, representing a starting state and a terminal state of any user search task;
- E ⊆ V × V is the set of directed edges;
- w: E → (0...1] is a weighting function that assigns to every pair of queries (q,q′) ∈ E a weight w(q,q′).

In the query-flow graph, two queries q and q′ are linked through an edge when there exists at least one session of the query log in which q′ follows q. The weight w may depend on the application; in the following the weight may be considered to be the transition frequency in the query log. The edge probabilities to each transition with related data are used in segmenting physical sessions into missions [17] and chains. Physical sessions are termed as activity sequences before a timeout of 30 minutes of a single user, whereas missions and chains are termed as activity sequences related topically. This is significant for applications in improving the user search experience.

Result and Discussion

Table 1 to 3 and figure 2 to 4 shows the precision using Imageclef 2005 dataset, IMDB dataset and 4 Universities dataset respectively.

Table 1
Results for Precision using Image clef 2005 dataset

Recall	Query	Query Expansion	Query Recommendation
0	0.63	0.79	0.89
0.1	0.59	0.73	0.87
0.2	0.44	0.71	0.85
0.3	0.41	0.68	0.77
0.4	0.37	0.66	0.72
0.5	0.37	0.45	0.7
0.6	0.32	0.45	0.68
0.7	0.32	0.45	0.68
0.8	0.32	0.45	0.68
0.9	0.32	0.45	0.68
1	0.32	0.45	0.68

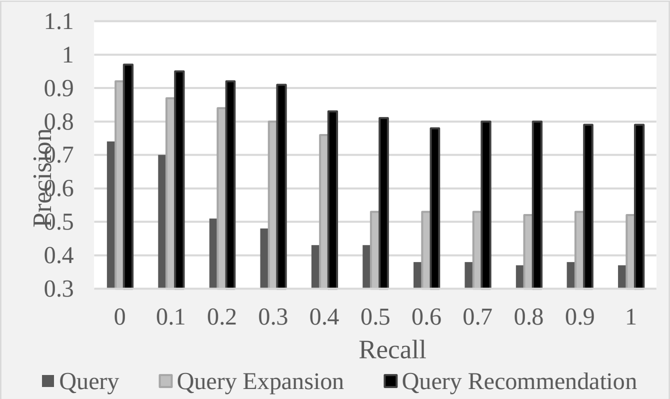

Figure 2: Precision using Image clef 2005 dataset

It is obtained from table 1 and figure 2 that the query recommendation obtain better precision by 60.1% than

Query and by 26.68% than Query Expansion by using Image clef 2005 dataset.

Table 2
Results for Precision using IMDB dataset

Recall	Query	Query Expansion	Query Recommendation
0	0.74	0.92	0.97
0.1	0.7	0.87	0.95
0.2	0.51	0.84	0.92
0.3	0.48	0.8	0.91
0.4	0.43	0.76	0.83
0.5	0.43	0.53	0.81
0.6	0.38	0.53	0.78
0.7	0.38	0.53	0.8
0.8	0.37	0.52	0.8
0.9	0.38	0.53	0.79
1	0.37	0.52	0.79

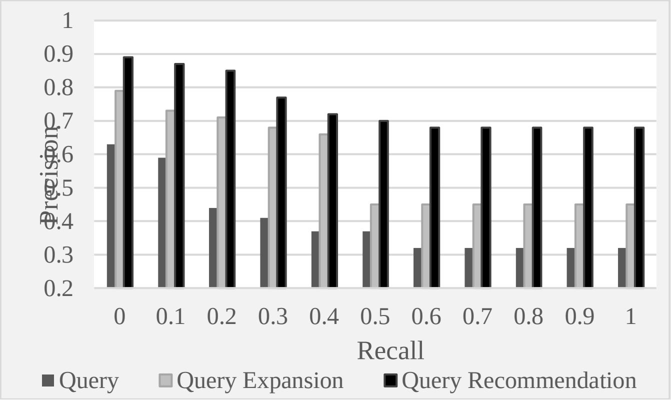

Figure 3: Precision using IMDB dataset

It is obtained from table 2 and figure 3 that the query recommendation obtain better precision by 57.58% than Query and by 23.95% than Query Expansion by using IMDB dataset.

Table 3
Results for Precision using 4 Universities dataset

Recall	Query	Query Expansion	Query Recommendation
0	0.69	0.87	0.9
0.1	0.67	0.82	0.87
0.2	0.48	0.81	0.87
0.3	0.45	0.77	0.86
0.4	0.41	0.72	0.77
0.5	0.41	0.49	0.78
0.6	0.35	0.51	0.72
0.7	0.36	0.5	0.76
0.8	0.35	0.49	0.76
0.9	0.36	0.5	0.76
1	0.35	0.49	0.76

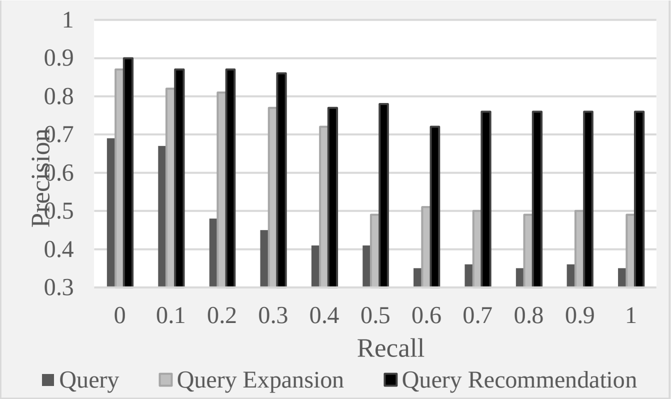

Figure 4: Precision using 4 Universities dataset

It is obtained from table 3 and figure 4 that the query recommendation obtain better precision by 57.41% than Query and by 23.3% than Query Expansion by using 4 Universities dataset.

Conclusion

These days the World Wide Web growth is increasing with the size and popularity along with large scale congregation volumes of web data. Hence it was complex to extort the related information to be used in broad application range. Query logs record the queries and the actions of the users of search engines with expensive information regarding the interests, the preferences, and the user's performance with feedback to search- engine. Query recommendation had been extensively approved by search users as significant way to find information efficiently. In accordance with search performance survey report, 78.2% users has changed their queries when satisfactory outcomes was not obtained with the current query and Users had adopted query recommendation function in clarifying their information needs devoid of taking attempts in entering new queries. Results showed that the query recommendation obtained better accuracy by 60.1% over Query and by 26.68% over Query Expansion through Imageclef 2005 dataset. Better results would be attained by using some other datasets.

References

1. Zahera H., Hady G. El and El-Wahed W., Query Recommendation for Improving Search Engine Results, World Congress on Engineering and Computer Science (WCECS), San Francisco, USA **(2010)**

2. Goyal Poonam and Mehala N., Concept based query recommendation, Proceedings of the Ninth Australasian Data Mining Conference, Australian Computer Society, Inc. **(2011)**

3. Dong L., Hybrid query expansion on ontology graph in biomedical information retrieval **(2012)**

4. Ramampiaro H. and Li C., Supporting biomedical information retrieval: The biotracer approach, In Transactions on large-scale data-and knowledge-centered systems, Springer Berlin Heidelberg, 73-94 **(2011)**

5. Fonseca B.M., Golgher P.B., de Moura E.S. and Ziviani N., Using association rules to discover search engines related queries, In Web Congress, Proceedings, First Latin American, IEEE, 66-71 **(2003)**

6. Ontology Information Science, Available from: http://en.wikipedia.org/wiki/Ontology_%28information_science %29 **(2011)**

7. Logeswari S. and Premalatha K., Biomedical document clustering using ontology based concept weight, In Computer Communication and Informatics (ICCCI), 2013 International Conference on, IEEE, 1-4 **(2013)**

8. Cui H., Wen J.R., Nie J.Y. and Ma W.Y., Probabilistic query expansion using query logs, In Proceedings of the 11th International Conference on World Wide Web, ACM, 325-332 **(2002)**

9. Taschwer M., Text-Based Medical Case Retrieval Using MeSH Ontology, In CLEF (Working Notes) **(2013)**

10. Bordino I., Castillo C., Donato D. and Gionis A., Query similarity by projecting the query-flow graph, In Proceedings of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 515-522 **(2010)**

11. Baeza-Yates R., Hurtado C. and Mendoza M., Query recommendation using query logs in search engines, In International Conference on Extending Database Technology, Springer Berlin Heidelberg, 588-596 **(2004)**

12. Cheng P.C., Chien B.C., Ke H.R. and Yang W.P., NCTU_DBLAB@ Image CLEFmed 2005: Medical Image Retrieval Task, In CLEF (Working Notes) **(2005)**

13. Sisode M.R. and Patil U.M., A Survey on Query Recommendation Techniques and Evaluation of Snippet based Query Recommendation **(2014)**

14. de Campos L.M., Fernandez J.M. and Huete J.F., Query expansion in information retrieval systems using a Bayesian network-based thesaurus, In Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., 53-60 **(1998)**

15. El Ghali B.T.I.H.A.L., El Qadi A.B.D.E.R.R.A.H.I.M., El Midaoui O.M.A.R. and Ouadou M., Query recommendation based terms and relevant documents using language models, *WSEAS Trans Inf Sci Appl*, **12**, 112-119 **(2015)**

16. Demir D., Kapralova O. and Lai H., Predicting IMDB movie ratings using Google Trends, Dept. Elect. Eng., Stanford Univ., California **(2012)**

17. Boldi P., Bonchi F., Castillo C., Donato D., Gionis A. and Vigna S., The query-flow graph: Model and applications, In CIKM **(2008)**.