Query Expansion For Medical Domain To Augment The Web Search

R.Uma, B.Latha, R.Valarmathi

Abstract: ***There has been a huge upsurge in the data available on the web making it voluminous over the last few years. Most users who search the internet for the medical information have no adequate knowledge about the medical domain. Understanding queries plays a huge role in helping users navigate through the abundant data and find the required information. Query feedbacks generally drift from the topic due to the various vocabulary mismatches or due to the ambiguous synonyms. A system has been developed to bridge the gap between expert users and the laypersons. We develop a system where by users query is enhanced by adding terms provided by experts. The appropriate query term is added from a medical expert vocabulary which is selected by the classifier based on the euclidean distance and a term based methodology for deciphering the query reformulation actions. The proposed system performs well in terms of precision and recall.***

Key Words: Query formulation, Health search, Medical informatics.

I. INTRODUCTION

In the modern and digital world people are highly dependent on information from the Internet. Our daily needs starting from paying bills to purchasing goods are all dependent on Internet. Internet has become one of the basic needs. A lot of medical inputs are also based on the images that are being used for diagnosis of diseases. The image based retrieval systems have proved to be effective with diagnosing diseases.

Many researches have been carried out to retrieve the most effective image and also to retrieve the images in least time. People search and receive responses through web browsers and web servers respectively. A lot of researches have been carried out on the way in which the user enters the request and how the browsers transfer the user requests. This research area involves mining which deals with extracting data from large warehouses. Usually a search engine will retrieve results based on page ranking. That is how many people have visited that particular page based on what query was given to the search engine. Lot of methods such as semantic indexing, query expansion have been used for this purpose. Specific research area includes medicine, sports, telecommunication etc. In medical domain, lot of similarities among the medical terms may be present. Also there may be dependency among such terms based on their usage. Few algorithms have been proposed to deal with such similarities.

Digital text document related to the query are retrieved based on matching of the character strings rather than based on the meaning. The user encounters a selection of documents containing information based on the strings combination presented in the query which may or may not be accurate. This vocabulary problems solution is either sought during the retrieval or storage phase. In the storage phase, as intellectual description of the document is expensive most of the times, the solution based on retrieval phase is concentrated on in this model. Under linguistic level when query expansion is done based on terms we observe that session of query search are illustrated using terms related to the information. The main aim of this model is to provide expert medical terms for lay terms given by laypersons. There are three type of actions provided to the user

- Term addition: During query reformulation addition of new terms that were not present in the previous queries
- Term retention: Retaining of a few terms mostly the core terms for the current information need.
- Term removal: Deletion of terms from a query.

II. RELATED WORK

Query Feeding

Queries are fed as input to most of the search engine. It is necessary to expand the queries as they give the best meaning when they are expanded. It is also flexible to add new words to the query. Expansion of queries can be done both manually and automatically Dwork et al (2001). However manual query expansion requires user intervention to predict what new queries may get added in the future.

Query Enhancement in medical domain

UMLS and MeSH have been used for enhancing the queries in medical domain abdouand Savoy (2008) .UMLS contains synonyms for medical terms.UMLS are utilized in PubMed.Terms from MeSH have been used in TREC Jalali and Borujerdi(2010).Relevance feedback have also been used for expanding the queries in the medical domain.

Layperson seeking Medical Information

When laypeople search for medical information they struggle to extract medical terms to access websites which are trust worthy Griffon et.al (2012).Furthermore they do not reformulate the query with synonyms or health data to

Revised Manuscript Received on 30 September, 2019.

R.Uma, Associate professor, Department of Computer Science and Engineering, Sri Sairam Engineering College,WestTambaram,Chennai, Tamilnadu, India.

Dr. B.Latha, Professor, Department of Computer Science and Engineering, Sri Sairam Engineering College,WestTambaram,Chennai, Tamilnadu, India.

R.Valarmathi, Associate professor, Department of Computer Science and Engineering, Sri Sairam Engineering College,WestTambaram,Chennai, Tamilnadu, India.

increase the relevant results Lau & Horvitz (1999).

Ontology based IR method

Details about a particular domain are stored in the ontology and are organized in the form of clusters. For this purpose semantic based knowledge was used instead of keyword based search. However it did not yield efficient results Can &Baykal(2007).

Semantic Indexing

This method is used to improve accuracy and precision. The index is created based on ontological data. When ranking, the document with ontological data gets higher rates. Using this method performance has been improved when compared to usage of traditional approach Jalali et al (2008).

Global Feature Mapping for bio-medical IR

Cluster of images have been used to extract a set of global content based features. Each cluster has been mapped with a unique alphanumeric code word. The words assigned to an image has been combined with other text related images and an index has been created. This index is searched using textual query Milne et al(2007).

III. SYSTEM METHODOLOGY

The proposed system considers a specific domain. For example sports, medical or literature. For a particular domain, how the user will prompt to search and what the user will search related to that domain is analyzed. Based on this the semantic words are extracted. Words that best captures the users intent can be used to expand the original query. Let us consider the medical domain and implement the system based on the terms, queries and words that are used in the medical domain.One of the major failures of search engine was identified to be the language gap. For example if we consider the medical domain, then how the people will use a search engine or any retrieval system to obtain results for their queries depends on the users level of expertise in that particular domain. An user who has medical knowledge can accurately retrieve their expected results with more precision. However an user who is not aware of the medical terms, if given a query regarding a disease or if they themselves ask a query then their level of expressing it may not be very accurate. To overcome this gap a better system has been proposed.

Division of Queries

Based on the expertise level of the user, queries can be divided into two categories. As in case of medical domain if the user is not familiar with the terms used then they can use the help of experts to construct their queries. The expert users will be preferring the original query which exactly describes what the user want from the IR system .

Synonym Mapping

The synonym of a word is mapped to various links. These links are provided based on how accurately they describe the terms and how frequently they are used by other experts to analyze the query submitted by user. Constructing domain specific synonyms by consulting with the domain experts can help in identifying the exact page by reformulating the query in a search. Precision value is computed to determine the relevancy between the total number of links available and the number of relevant links obtained henceforth. Precision is calculated with the number of correct information which the system returns .

Overlapping

A compare and contrast is done on the synonym mapping obtained above. Different candidates can be considered for synonym mapping and the overlapping, measures the similarities and differences between the mapped words. An online platform is built to analyze the effectiveness of the approach. Here the queries entered by the users are analyzed, mapped and subjected to overlapping. Then the minimal optimal overlap value can be obtained based on the greedy approach. Graphs can be constructed based on the number of terms used and the number of candidates present. Let us consider a scenario where we have four candidates namely C1, C2, C3 and C4. The various candidates being C1 is Layman with Query Enhancement,C2 is Layman without Query Enhancement,C3 is Expert with Query Enhancement,C4 is Expert without Query Enhancement. These candidates are compared with the terms that are retrieved based on synonym mapping.

IV. RESULTS AND DISCUSSION

The proposed system ensures user friendliness apart from ensuring accuracy and precision in retrieving the relevant records or links related to bio-medical domain. The same can be extended to other domains following the same processes as above. As we have considered the four candidates namely C1, C2,C3 and C4 ,we will construct the comparative graphs among these four candidates with respect to the terms that are retrieved under each candidate. Initially we will divide the users into two categories non expert group the layperson and the expert group. The corresponding user count is also displayed in the graph.

The performance measure is calculated in terms of Accuracy, Precision, Recall and F-Score. The graph is plotted for Layman and Experts, each with and without Query Enhancements. After applying our Information Retrieval System (with Query Enhancement), the performance measures for Experts and Layman were greater than that without Query Enhancement.

Accuracy = True Positive+True Negative/ True Positive +False Positive+ False Negative +True Negative

Precision = True Positive / True Positive + True Positive

Recall = True Positive / True Positive +False Negative

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Table 1 Precision, Recall and F-measures for comparison of methods

Type of User	Performance measures			
	Accuracy	Precision	Recall	F-Score
C1	0.727	0.732	0.86	0.79
C2	0.445	0.525	0.667	0.587
C3	0.84	0.906	0.922	0.913
C4	0.811	0.828	0.856	0.841

Figure 1 shows the candidates and their corresponding Accuracy, Precision, Recall and F-Score in obtaining the correct terms.

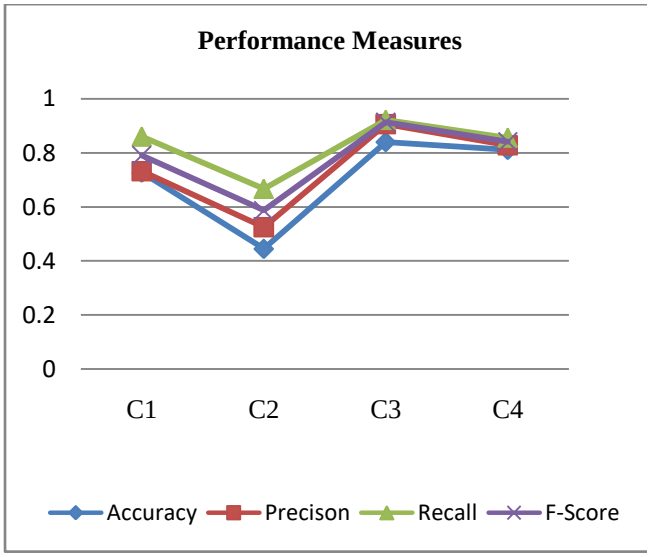

Figure1. Performance Measures

.

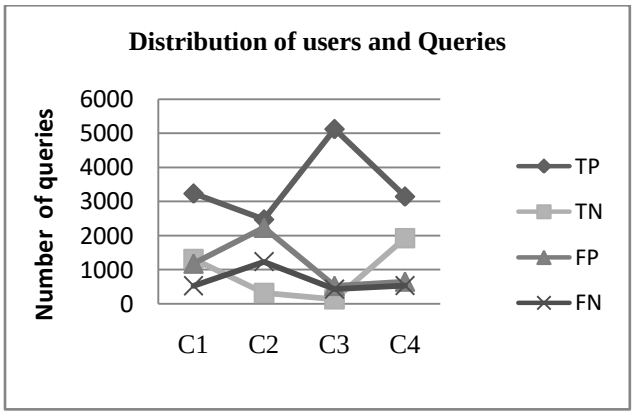

Figure 2. Distribution of Users and Queries

The graph in figure 2 represents the distribution of users and queries. It depicts the fraction of correct answer obtained for each query and the query count.The graphs shows the mapping of number of queries used to test the performance measure of the information retrieval system and the number of truly positive and truly negative queries. It was found that the number of true positives for Experts with Query Enhancement was the highest compared to the other categories.

V. CONCLUSION

Searching medical information is a crucial task of the search engine. Fetching the information has a great effect on the well being on the user .The medical terms used makes the article inaccessible by the laypeople. An effort has been taken to make the content accessible by both experts and laypersons. The enhanced querying solves major issues faced by the laypersons in accessing the medical information. It is evident from the results that the proposed method bridges the gap between medical experts and laypersons.The results show that the proposed method outperforms the existing state of the art query expansion methods for the medical domain and can be extended to different sub-domains in future.

REFERENCES

1. Abdou, S & Savoy, J (2008),” Searching in medline: Query expansion and manual indexing evaluation”, Information Processing & Management, vol.44 ,no.2,pp. 781–789.
2. Can, A. B., & Baykal, N. (2007). Medicoport: A medical search engine for all. Computer methods and programs in biomedicine, vol.86 ,no.1, pp.73–86.
3. Dwork, C, Kumar, R, Naor, M, & Sivakumar, D, (2001). “Rank aggregation methods for the web”, In Proceedings of the 10th international conference on world wide web, WWW ’01, New York, NY: ACM, pp. 613–622.
4. Griffon, N, Chebil, W, Rollin, L, Kerdelhue, G, Thirion, B, Gehanno, J. F, & Darmoni, S. J (2012), “Performance evaluation of unified medical language system’s synonyms expansion to query”, PubMed. BMC medical informatics and decision making, vol.12 no.1, pp 2.
5. Jalali, V, & Borujerdi, M, R, M, (2008), “The effect of using domain specific ontologies in query expansion in medical field’, In Innovations in Information Technology, pp. 277–281. IEEE.
6. Liu, Z. & Chu, W. W. (2007),”Knowledge-based query expansion to support scenario-specific retrieval of medical free text”, Information Retrieval, vol .10 no.2, pp.173–202.
7. Milne, D. N, Witten, I. H, & Nichols, D. M. (2007), “A knowledge-based search engine powered by Wikipedia”, In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 445–454, ACM.
8. Lau, T, & Horvitz, E. (1999), “Patterns of search: Analyzing and modeling web query refinement”, In UM’99: Proceedings of the seventh international conference on user modeling pp. 119–128.
9. A.F.Smeaton,C.J Van Rusbergen,”The Retrieval Effects of Query Expansion on a Feedback Document Retrieval System”The Computer Journal,vol .26, no.3,pp 239-246.
10. Salton.G. and McGill,M.G(1983),Introduction to modern Information Retrieval McGraw Hill Book Co,NewYork

11. Uma, R, Latha, B (2016),“ Sub-topic modeling-a hierarchy model for topic correlations”, International Journal of Control Theory Application 9(28), pp.175–179 (2016) 22.
12. Uma, R, Latha, B(2015),“ Enhanced clustering of correlated probabilistic graphs”. International Journal of Science and Engineering Research.
13. Uma, R., Latha, B (2018),“ Noise elimination from web pages for efficacious information retrieval”,Cluster Computing The Journal of Networks, Software Tools and Applications ISSN 1386-7857, Cluster Computing, DOI 10.1007/s10586-018-2366-x.
14. Uma, R., Latha, B (2018), “An efficient voice based information retrieval using bag of words based indexing”, International Journal of Engineering & Technology, vol 7, pp. 622-627.

AUTHORS PROFILE

R.Uma graduated from Madras University B.E(EEE), India in 1996 and received her Master’s in Computer Science and Engineering from College of Engineering Guindy, Anna University, India in 2002.Currently a Part Time Research Scholar of Anna University and working as Associate Professor in the Department of Computer Science and Engineering, Sri Sairam Engineering College, Chennai, India.

Dr.B.Latha graduated from Annamalai University (BE/CSE), India in 1998 and received her Master’s in Computer Science and Engineering from Sathyabama University, India in 2005 and completed her Doctorate Degree in the Faculty of Information and Communication Engineering, Anna University Chennai, India in 2010. Currently, working as Professor and Head of the Department, in the Department of Computer Science and Engineering,, Sri Sairam Engineering College, Chennai, India. She has published her paper over 41 International Journals and presented papers in 26 International conferences and 15 National conferences and also she has published 2 books. She is a member of IEEE, CSI, IACSIT and life member of ISTE. Her current research interest includes Artificial Intelligence, Network Security, Machining of composite materials, Computer aided modeling and optimization. She is guiding 10 Ph.D research scholars.

R.Valarmathi graduated from , Madras University B.E(CSE), India in 1999 and received her Master’s in Computer Science and Engineering from Sathyabama University, India in 2010.Currently a Part Time Research Scholar of Sathyabama University and working as Assistant Professor in the Department of Computer Science and Engineering, Sri Sairam Engineering College, Chennai, India