International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 4
ISSN: 2454-4248
357 – 362

IRQX: A Framework for Information Retrieval Algorithms Using Query Expansion Techniques

R. Thamarai Selvi
Associate Professor and Head
Department of Computer Applications, Bishop Heber College
Tiruchirappalli-620017, INDIA
e-mail: thams.shakthi@gmail.com

Dr. E. George Dharma Prakash Raj
Assistant Professor of Computer Science
Bharathidasan University
Tiruchirappalli-620023, INDIA
e-mail: georgeprakashraj@yahoo.com

***Abstract*—**The number of information retrieval users and their operations are continuously increasing with the rapid growth of internet technologies. Information Retrieval is one of the most prevalent operations that is frequently used by the Internet users. The process of Information Retrieval may cause two problems. First, the search engine may retrieve irrelevant documents and second it may fail to retrieve the relevant documents. Many approaches have been proposed to improve the query representation by reformulating the queries. Among them, Query Expansion (QE) is one of the most effective approaches. In Information Retrieval, Query Expansion is referred to as the techniques or algorithms that reformulate the original query by adding or modifying new terms into the query, in order to achieve better retrieval results. This paper contributed to the process of information retrieval algorithms using query expansion techniques to improve the precision and recall. The proposed framework Information Retrieval algorithms using Query Expansion (IRQX) facilitates the users to select their choice of algorithms based on their need.

Keywords *- Information Retrieval, Semantic, Query Expansion, Fuzzy Set, Ranking,Synonyms Antonyms, Precision, Recall, Ontology,Question and Answering*

*****

I. INTRODUCTION

In information retrieval, ad-hoc retrieval is a simulation of how a collection of textual documents might be used, to search for a static set of documents using a new set of topics. In ad-hoc retrieval the number of possible queries is also large. The main aim of an IR system is to find relevant information for a given topic of request.

Many approaches have been proposed to improve the query representation by reformulating the queries. Among them, Query Expansion (QE) is one of the most effective approaches. In IR, Query Expansion is referred to as the techniques or algorithms that reformulate the original query by adding or modifying new terms into the query, in order to achieve better retrieval results.

One of the techniques to improve the retrieval process using Query expansion involves semantic. Semantic means the meaning and interpretation of words, signs and sentence structure. Sometimes the relevant documents may not contain the specified keyword. The lack of the given term in a document does not necessarily mean that the document is not a relevant. Because more than one terms can be semantically similar although they are lexicographically different.

Sometimes the documents may contain the opposite words(antonyms) of the given words preceded by the words like ‘not’, ‘im’,‘un’, ‘ir’ etc. For example, the word ‘good’ has the meaning ‘not bad’. So, the document contains the word ‘not bad’ is also a relevant document for the user. Retrieval, by classical information retrieval models which are based on lexicographic term matching. Therefore, these methods do not retrieve documents with semantically similar terms.

Search engines have become a crucial tool upon which millions of users are dependent for finding desired information. One of the core problems that search engines face in order to satisfy users' information needs is judging whether a piece of information is relevant to a given information need as specified by a text query. So, IR system needs some kind of feedback from the user to choose their choice.

A good IR system has to find all the relevant documents and rank them by using some ranking function. The quality of this ranking function is an important factor that determines the quality of the IR system. Question answering can be viewed as a sophisticated Information Retrieval task where a system automatically generates a search query from a natural language question and finds a concise answer from a set of documents.

The rest of this paper is organized as follows. Section II describes the related work in Information Retrieval. Section III describes the IRQX framework, Section IV describes the algorithms in the IRQX framework and Section V analyses the performance of SBIR, UFSBIR, ESBIR and FSBIR algorithms. And finally, the conclusion is given in Section VI.

II. RELATED WORK

The significance of Boolean Information Retrieval (BIR) has been revealed in many retrieval systems because of its simplicity [1]. Most of the commercial IR systems use thisBoolean model to predict that each document is either relevant or non relevant [2].

The semantic retrieval [3] approach is used to discover semantically similar terms using WordNet. In many works, WordNet is used to identify similar concepts that correspond to document words. The Porter stemmer [4] is a context sensitive suffix removal algorithm. WordNet expansion technique was used [5] over a collection with minimal textual information. The Porter Stemmer Algorithm implemented in java performs this process for each word from the synsets. Porter’s stemmer is more compact and easy to use than Lovins [6].

Philip resnik et al. [7] annotated the Biblical text to create the aligned corpus such as Bible for linguistic research which also includes the automatic creation and evaluation of translation lexicons and similar tagged text. It has the feature of parallel translations over huge number of languages. It also represents the comparison with dictionary and corpus resources for modern English. Thus, it makes the Bible a multilingual

 ISSN: 2454-4248
357 – 362

corpus which considered to be a unique resource for linguistic research.

Wei Song et al. [8] proposed a fuzzy control genetic algorithm (GA) in conjunction with a novel hybrid semantic similarity measure for document clustering. In order to evaluate the performance of the algorithm, two standard data sets such as Reuter (21578 version) corpus and 20-newsgroup (18828 version) corpus, are used for test. Thesaurus-based and corpus-based semantic methods are used to solve the complicated term indexing method. WordNet is used as the thesaurus-based ontology. It is concluded that Fuzzy control GA performed better than conventional GA with the same similarity measures.

S. Niveditha et al. [9] proposed an algorithm that works with various user's search goals for a query and reflecting each idea with some keywords. They have done the above process by clustering the proposed feedbacks. They have determined number of user search goals for a query and clustered the Pseudo Documents using Fuzzy Self Constructing Algorithm to get the final restructured search result. The final results on user click from a commercial search engine demonstrated the effectiveness of their proposed methods. It has been finalized that in future when a user searches for a same topic for many time that topic will be sent to the user.

Narina Thakur et al. have implemented of an efficient Information Retrieval (IR) System to compute the similarity between a dataset and a query using Fuzzy Logic. TREC dataset has been used for evaluate the proposed approach. The dataset was parsed to generate keywords index which was used for the similarity comparison with the user query. Each query was assigned a score value based on its fuzzy similarity with the index keywords. The relevant documents were retrieved based on the score value. Results indicated that proposed similarity measure technique based on fuzzy logic, was better than cosine based similarity measure technique for handling vague, uncertain and imprecise queries.

Sharmela Shaik et al. [Sha, 11] explained the procedure and importance of reforming the natural language (NL) query into SPARQL query to apply to the database to retrieve the accurate semantic results. SPARQL is an RDF query language which is a semantic query language used to retrieve data and give precise results.

Xiaoqiang Liu et.al [12] implemented the Question Answering system based on the Jena API and supports natural language querying OWL ontology with respect to the ontology knowledge representation of refrigerator, a question answering system was designed based on ontology. Ontology knowledge representation for refrigerator is built with instances of the refrigerators', attributes and values with the subject-predicate-object triples format. The natural language question is analyzed, extract key words, analyze the dependencies between keywords, and build the query for ontology triples.

Sharvari et al. [13] proposed a Question Answering system for Marathi natural language by using concept of ontology as a formal representation of knowledge base for extracting answers. Ontology is used to express domain specific knowledge about semantic relations and restrictions in the given domains. The ontology is developed with the help of domain experts and the query is analyzed both syntactically and semantically. The results obtained here are accurate enough to satisfy the query raised by the user. The level of accuracy is enhanced since the query is analyzed semantically. The system is tested with Marathi documents of various domains like History, sports, festival, politics, etc and shows an overall precision of 93.95%, recall of 94.55% and accuracy of 89.28%.

Chandra et al. presented a survey of Question Answering and introduced the architecture of Question Answering system. This survey paper describes different types of Question Answering systems and general architecture of Question Answering system. Closed domain question answering system is restricted to a specific domain and the quality of answers is high. There is Open domain question answering system not restricted to any domain and the quality of answers is low. The closed domain question answering system gives more exact and correct answers than open domain question answering system.

III. IRQX FRAMEWORK

This framework will identify whether the query is a topic or a question. If the given query is a particular keyword or topic, the user will choose the algorithm to retrieve the documents based on their requirements. For example, if the user wants to retrieve more number of relevant documents which have the same meaning of the given topic or keyword, the SBIR algorithm will be selected. The index will be created for this topic and it will be stored. Then, the retrieved documents will be displayed. The IRQX framework is given in Figure 1.

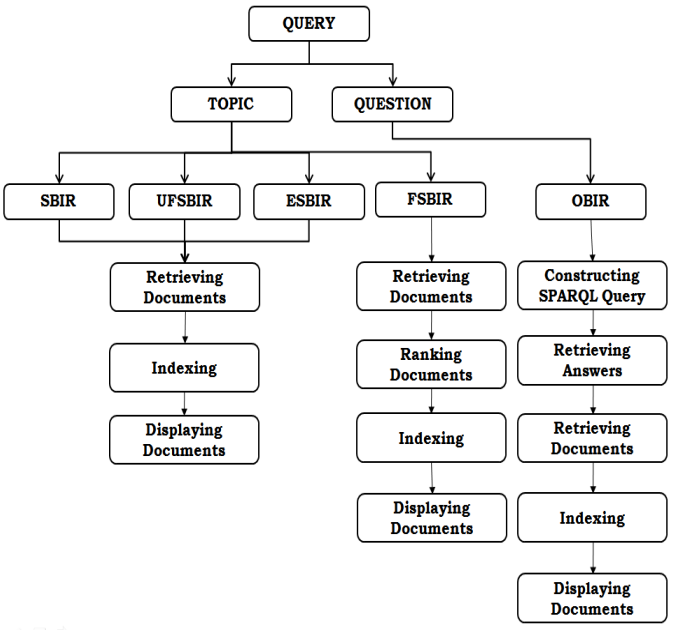

Figure 1. IRQX Framework

Sometimes, the user may want to feedback the IR System with the interested synonyms to retrieve the documents. In this case, the UFSBIR algorithm will be selected. The words "bad" and "not good" are having same meaning. So, the document with the words "not good" is also a relevant document for the topic "bad". Here, the query is expanded with the negation of the antonyms. If the user is interested in retrieving documents using antonyms, the ESBIR algorithm will be selected. Generally the corpus is a collection of textual documents. The task of ad-hoc Information Retrieval is finding the relevant document and displaying the order in the same way they appeared in the collection. Since the query is expanded with synonyms and antonyms, the documents that are retrieved may be ranked. In this case, the algorithm FSBIR will be selected. If the query is a question then the type of the question is identified and SPARQL query is formulated and the answers and documents are retrieved from the ontology.

 ISSN: 2454-4248
357 – 362

IV. PROPOSED ALGORITHMS

A. SBIR: Semantic based Boolean Information Retrieval

SBIR (Semantic based Boolean Information Retrieval) was proposed [15] to retrieve the documents with semantically similar terms for a keyword search. SBIR algorithm provides a novel perspective to the task of ad-hoc retrieval. Here, the given query is expanded with the synsets extracted from WordNet dictionary. Then, each synset is stemmed to root word and the documents are retrieved.

B. UFSBIR: User Feedback Semantic based Boolean Information Retrieval Algorithm

UFSBIR algorithm is proposed to retrieve the documents with semantically similar terms to enhance the performance of Boolean Information Model to reduce the time to retrieve the documents by allowing the user to feedback the query [16]. This work targets to develop a system to address the Information Retrieval for a static data set and aims to provide documents from within the collection that are relevant to an arbitrary user information need. UFSBIR retrieves documents from the document set by finding their synonyms using WordNet data base to find more similar documents are retrieved. The synonyms are then stemmed to find the root words using Porter Stemmer algorithm. And then the user is allowed to select the list of words for the IR process. The user interaction with IR system improves the performance by selecting query terms based on the document collection. The precision and recall values are calculated and the results reveal that the efficiency of UFSBIR is due to the feedback given by the user after collecting the synsets.

C. ESBIR: Extended Semantic Based Information Retrieval Algorithm using Synonyms and Antonyms

An Extended Boolean model using synonyms and antonyms is used to predict whether each document is relevant or not. It refers an online lexical reference called WordNet to find the semantically similar terms. ESBIR Algorithm retrieves the semantically relevant documents with the use of Word Net database, antonyms and stemming algorithm [17]. IR provides the required information according to the user query within a collection of data. The lack of the given term in two documents does not necessarily mean that the documents are not related. Sometimes the documents may contain the opposite words(antonyms) of the given words preceded by the words like ‘not’, ‘im’, ‘un’, ‘ir’ etc. For example, the word ‘good’ has the meaning ‘not bad’. So, the document contains the word ‘not bad’ is also a relevant document for the user. For example, the synsets for the word “impossible” are “not capable”, “unacceptable”, “unimaginable”, “not possible”, “not attainable”, “not acceptable”. These words are stemmed to their root word. After stemming process, the documents are retrieved from the database for all the root words.

D. FSBIR: Fuzzy Implemented Semantic Based Information Retrieval Algorithm using Query Expansion

FSBIR algorithm exemplifies the implementation of fuzzy ranking for semantic based information retrieval using synonyms and antonyms. Query expansion is often effective in increasing recall [18]. Generally, the given query is expanded with the synonyms, but in this proposed algorithm the antonyms of the given query is also considered to expand to retrieve more number of relevant documents. A good IR system has to find all the relevant documents and rank them by using some ranking function. The quality of this ranking function is an important factor that determines the quality of the IR system. “Fuzzy Implemented Semantic Based Information Retrieval using Query Expansion” (FSBIR) is proposed to retrieve the documents using synonyms and antonyms with fuzzy set ranking function. The idea behind this FSBIR is to first add terms with close meaning to the original query to expand it, and then reformulate the ranked documents to improve the performance of the overall retrieval.

E. OBIR: Ontology Based Semantic Information Retrieval using Query Expansion for Question and Answering

OBIR addresses the task of providing answers for the given question using semantic based ontology construction in the specified domain. Question answering is the task of finding a concise answer to a natural language question. The answers will be supported by the document context. Question answering can be viewed as a sophisticated Information Retrieval task where a system automatically generates a search query from a natural language question and finds a concise answer from a set of documents. In this work, new novel algorithm “Ontology Based Semantic Information Retrieval using Query Expansion for Question and Answering” is proposed to find the correct answer for the given question. Here, the given query is expanded semantically and SPARQL query is framed and then the answer is retrieved from the ontology along the documents for the reference. This algorithm always gives the correct answer for the question with document reference.

V. PERFORMANCE ANALYSIS OF PROPOSED ALGORITHMS

A. Number of Documents – Analysis

Queries are given to retrieve the relevant documents from the Bible data set. The Bible is one of the religious books which is referred by many people. It contains many verses. People may want to refer the verses which have the same meaning for the given word. In this work, The Bible (Kings James Version) database is used. This database contains 66 Books, 1189 Chapters, 31,102 Verses and 7,882,80 words. The comparison of number of documents retrieved by IR algorithms is shown in Table I and it is graphically represented in Figure 2.

TABLE I. COMPARISONS ON NUMBER OF DOCUMENTS RETRIEVED BY PRPOSED IR ALGORITHMS

Query	BIR	SBIR	UFSBIR	ESBIR	FSBIR
impossible	**9**	**184**	155	193	193
wrong	34	167	143	150	150
close	27	116	97	110	110
accept	86	161	153	180	180
reject	30	123	110	165	165
weak	57	278	229	239	239
forget	61	74	64	82	82
happy	25	350	206	218	218

 ISSN: 2454-4248
357 – 362

Precision - Analysis

The precision of the algorithms are calculated and given in the following Table II and it is graphically represented in Figure 3. The ESBIR and FSBIR algorithm give a better performance than the other algorithms.

TABLE II. PRECISION ANALYSIS OF PROPOSED IR ALGORITHMS

Query	BIR	SBIR	UFSBIR	ESBIR	FSBIR
impossible	1	0.64	0.71	0.83	0.83
wrong	1	0.69	0.78	0.95	0.95
close	1	0.72	0.86	0.77	0.77
accept	1	0.87	0.85	0.99	0.99
reject	1	0.89	0.87	0.98	0.98
weak	1	0.73	0.83	0.95	0.95
forget	1	0.88	0.94	1	1
happy	1	0.59	0.87	0.96	0.96

Recall – Analysis

The recall of the algorithms are calculated and given in the following Table III and it is graphically represented in Figure 4.

The ESBIR and FSBIR algorithm gives a better performance than the other algorithms since it displays the relevant documents based on the rank. This algorithm gives the correct answer and the document reference for the query using ontology. So, all the relevant documents are retrieved. It produces the recall value 1 for the queries.

TABLE III. RECALL ANALYSIS OF IR PROPOSED ALGORITHMS

Query	BIR	SBIR	UFSBIR	ESBIR	FSBIR
impossible	0.06	0.74	0.69	1	1
wrong	0.24	0.81	0.78	0.96	0.96
close	0.32	0.98	0.98	0.98	0.98
accept	0.48	0.79	0.73	0.95	0.95
reject	0.19	0.68	0.6	0.93	0.93
weak	0.25	0.89	0.83	0.92	0.92
forget	0.74	0.79	0.73	0.82	0.82
happy	0.12	0.97	0.86	0.91	0.91

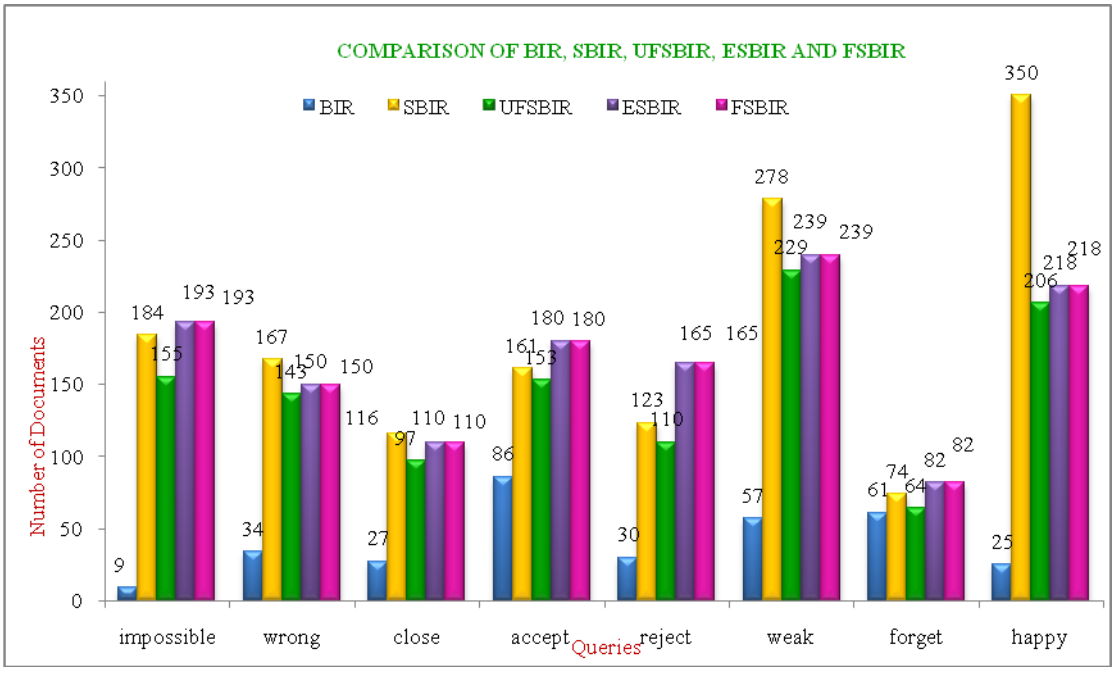

Figure 2. Comparisons on Number of Documents retrieved by Proposed IR algorithms

ISSN: 2454-4248
357 – 362

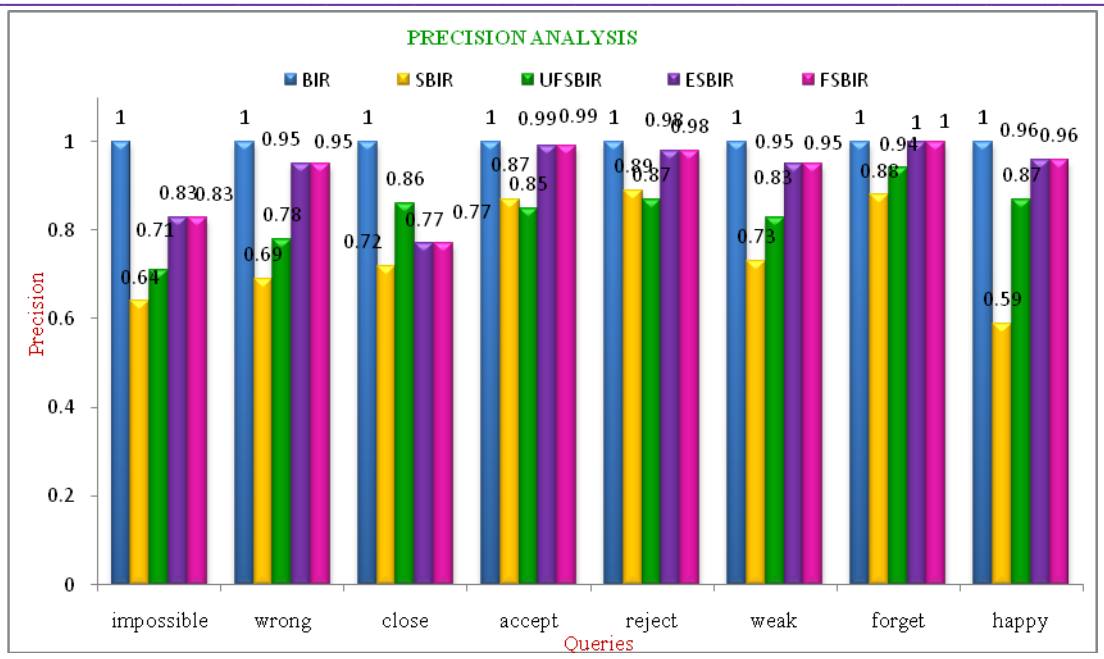

Figure 3. Precision Analysis of Proposed IR Algorithms

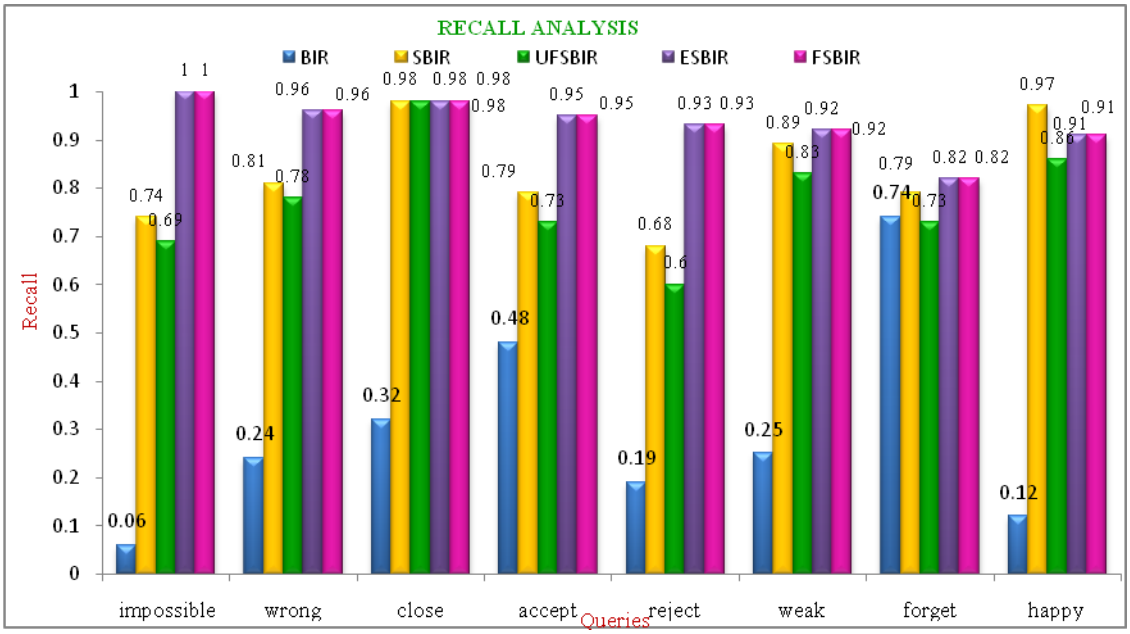

Figure 4. Recall Analysis of Proposed IR Algorithms

The precision, recall are set-based measures. They are computed for the unordered sets of documents. Further these measures to evaluate the ranked retrieval results needs to be extended. In a ranked retrieval context, appropriate sets of retrieved documents are naturally given by the top k retrieved documents. For each query the precision and recall, values aer plotted to give a precision-recall curve. For each information need, the interpolated precision is measured at the 11 recall levels of 0.0, 0.1, 0.2 …1.0. The arithmetic mean of the interpolated precision for each recall level and for each information need is calculated. Table IV shows the average 11-point interpolated precision of Extended Semantic based Boolean Information Retrieval and FSBIR. And, a composite precision recall curve is drawn to evaluate overall system performance on the corpus.

TABLE IV 11-POINT INTERPOLATED AVERAGE PRECISION

RECALL	INTER-PRECISION	
	ESBIR	FSBIR
0	0.81	1.00
0.1	0.67	0.91
0.2	0.63	0.84
0.3	0.55	0.75
0.4	0.45	0.72
0.5	0.41	0.65

 ISSN: 2454-4248
357 – 362

0.6	0.37	0.60
0.7	0.30	0.52
0.8	0.22	0.45
0.9	0.14	0.35
1	0.08	0.12

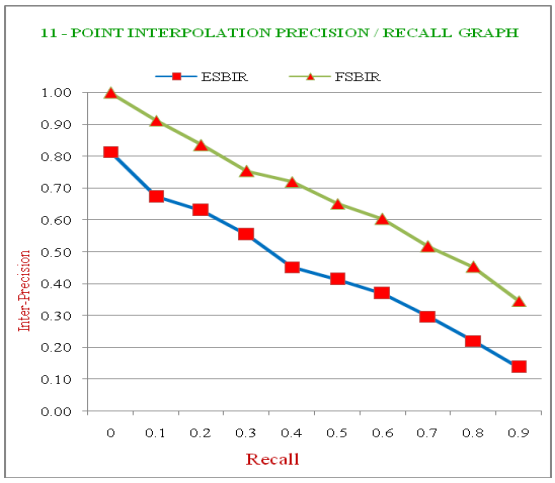

Figure 5. 11-Point Precision Recall Curve

The 11-point interpolation precision/recall values are also plotted in the Figure 5. The curve closer to the upper right-hand corner of the graph indicates the best performance. In this graph, the curve formed by the algorithm FSBIR is very closer to the upper right-hand corner compared to the curve formed by the ESBIR. This gives the better result. The result shows that the inter-precision values for the given recall is increased in the FSBIR algorithm compared to ESBIR algorithm since fuzzy logic is used to rank the documents.

VI. CONCLUSION

In this paper the overall framework for all the proposed IR algorithms is presented. The proposed IRQX framework facilitates the users to select their choice of algorithms based on their need. The given query is expanded with synonyms and antonyms. The performance analysis of the proposed SBIR, UFSBIR, ESBIR and FSBIR algorithms are compared in terms of number of retrieved documents, Precision and Recall.

REFERENCES

[1] R.B.-Yates and B.R.-Neto, "Modern Information Retrieval", Addison Wesley Longman, 1999.

[2] Salton, G., McGill,M., "Introduction to Modern Information Retrieval", McGraw-Hill, New-York, 1983.

[3] Fellbaum, C., "WordNet. Theory and Applications of Ontology: Computer Applications", Springer Science Business Media B.V,Vol.231,pp.231-243, 2010.

[4] Giridhar N S, Assistant Professor, Prema K.V, Professor, N .V Subba Reddy, "A Prospective Study of Stemming Algorithms for Web Text Mining", Ganpat University Journal of Engineering & Technology, Vol.1, No.1, Jan, 2011.

[5] Manuel, D., Maria, M., Alfonso, U. L., & Jose, P., "Using WordNet in Multimedia Information Retrieval", CLEF 2009 Workshop, Part II, LNCS, Springer-Verlag Berlin Heidelberg, 6242, pp. 185–188,, 2010.

[6] Deepika Sharma, "Stemming Algorithms: A Comparative Study and their Analysis", International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 4– No.3, 2012.

[7] Philip Resnik, Mari Broman Olsen and Mona Diab, "The Bible as a Parallel Corpus: Annotating the "Book of 2000 Tongues", Computers and the Humanities Vol.33, pp. 129–153, 1999.

[8] Wei Song a,b, Jiu Zhen Liang a, Soon Cheol Park b "Fuzzy control GA with a novel hybrid semantic similarity strategy for text clustering", Information Sciences Vol. 273, pp:156-170, 2014.

[9] S.Niveditha, T. Malathi, S.R.Sivaranjhani,"Efficient Information Retrieval using Fuzzy Self Construction Algorithm", International Journal of Computer Applications (0975 – 8887) Vol.104, No.1, October , 2014.

[10] Narina Thakur, Prabhjot Singh, SumitDhawan and Shubham Agarwal, "Implementation of an efficient Fuzzy Logic based Information Retrieval System", EAI Endorsed Transactions on Scalable Information Systems, Vol.2, No.5, pp.1-7, 2015.

[11] Sharmela Shaik, Prathyusha Kanakam, S Mahaboob Hussain, D. Suryanarayana, "Transforming Natural Language Query to SPARQL for Semantic Information Retrieval", International Journal of Engineering Trends and Technology (IJETT)", Vol.14, No.7, pp.347-350, 2016.

[12] Xiaoqiang Liu, Zhenbo Guo, Kaixi Wang, Wenxu Jiang, "Study and Development of Question Answering System based on Ontology Query", Advances in Computer Science Research, ISSN:2352-538X, pp.430-432, 2016.

[13] Sharvari, S. Govilkar, J. W. Bakal, "Question Answering System using Ontology in Marathi Language", International Journal of Artificial Intelligence and Applications (IJAIA)", Vol.8, No.4, pp. 43-64, 2017.

[14] Chandra Obula Reddy. A , Dr. Madhavi . K, "A Survey on Types of Question Answering System", IOSR Journal of Computer Engineering (IOSR-JCE), Vol.19, No.6, e-ISSN: 2278-0661, p-ISSN: 2278-8727, pp. 19-23 , 2017.

[15] Thamarai Selvi. R, E. George Dharma Prakash Raj, "An Approach to Improve Precision and Recall for Ad-hoc Information Retrieval using SBIR Algorithm", World Congress on Computing and Communication Technologies, IEEE Xplore Digital Library, pp. 137141, 2014.

[16] Thamarai Selvi. R, Dr.E.George Dharma Prakash Raj. "UFSBIR:A Semantic based Boolean Information Retrieval Algorithm with User Feedback", International Journal of Information Systems, Vol.1 , pp.36-40, 2014.

[17] Thamarai Selvi. R, Dr. E. George Dharma Prakash Raj. "ESBIR:Extended Semantic based Boolean Information Retrieval Algorithm using Synonyms and Antonyms", Journal of Emerging Trends in Computing and Information Sciences, Vol.6, No.4, pp. 198-202, April 2015.

[18] Thamarai Selvi. R, Dr. E. George Dharma Prakash Raj, "FSBIR: Fuzzy Implemented Semantic Based Information Retrieval Algorithm using Query Expansion", International Journal of Applied Engineering Research, Vol. 10, No.82, pp.498-502, 2015.