Journal of Theoretical and Applied Information Technology
31st January 2019. Vol.97. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

NEWS RETRIEVAL BASED ON SHORT QUERIES EXPANSION AND BEST MATCHING

[1] ZAINAB A. KHALAF, [2] INTISAR A. SHTAET

[1,2] Basrah University, College of Science, Department of Computer Science, Basrah, Iraq
E-mail: [1]zainab_ali2004@yahoo.com, [2] intisar.redha14@gmail.com

ABSTRACT

An information retrieval (IR) is a system that locates information in collections of unstructured or semi-structured document which are relevant to the user query. Usually, the query represented by short query that have a few features or keywords that used to describe the user needed of information. These short queries containing only a few words can be vague and ambiguous. As a result, IR system returns documents that are generally not relevant to the user's information needs. In this paper, an automatic approach of query expansion is proposed based on latent semantic indexing (LSI) method in order to enhance the performance of the information retrieval system. Development database of queries are used to find the best match between the user query and the development queries by using LSI algorithm. The best result that obtained from LSI will combine with the user query to create a new query that used later as new input in information retrieval system to retrieve the documents. To evaluation, the information retrieval system is compared before and after query expansion by using latent semantic indexing (LSI) method. The proposed approach improved the retrieval system performance from 70% to 75% F-measure with an average relative improvement of 7.2%, and it is better than the conventional LSI approach.
Keywords: *Information Retrieval; Latent Semantic Indexing; Query Expansion; News Retrieval.*

1. INTRODUCTION

Accuracy of the information retrieval (IR) system is a domain dependent. There are many factors that effect on the IR system performance such as short queries, out-of-domain problem, and out of vocabulary (OOV) [1,2,3].

The query is statements of the information that needed by the user. Sometimes, the user query is very short which have little number of features that not reflect the user information need, so this called short query [4]. Thus, when search in the IR system by using the short query to find the best relevant documents that satisfied the user need, the IR system may be failed or give mismatched documents. These retrieved documents may be far away from the user needed. Therefore, the short query problem is one of the important factors that impact on the performance of IR system [5].

Recently, query expansion (QE) has received a great deal of attention in the IR literature and the research on resolving the problem of the short query. Query expansion (QE) is an approach that have been used in the IR system which reformulates a user query in order to enhance the effectiveness of information retrieval, where by the original query of the user is enhanced by adding new features that have a similar meaning [6].

Almost all the popular techniques of the query expansion used to expand the initial user query via combining the new related features with the features of the original user query. This can also involve selective retention of features from the user query. The expanded or the reformulated query is then used to retrieve more relevant results. Automatic query expansion is actually considered as promising technique to improve the retrieval effectiveness of documents ranking [7].

The key to an effective expansion of query is how to choose features which are appropriate for the context, or the topic of the query [6].

In this work, we proposed approach to expand short queries that is based on in-domain develop dataset. It searching the related terms from this dataset to expand the original query. The main idea behind query expansion is to increase the number of features in order to get better matching and that

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

will affect the performance of the system. The proposed approach is applied automatically based on combining the user query and the new query expansion through the use of the semantic resources.

The rest of the paper is organized as follows: after showing the introduction in section 1, the related work is shown in section 2. Section 3 shows the information retrieval system. Section 4 explains the proposed system. Next, section 5 provides the experiments to evaluate the proposed approach. Section 6 gives the conclusions. While section 7 provides the limitations of this study. Finally, the suggestions for future works is given in the section 8.

2. RELATED WORK

There are some related works that have been proposed for information retrieval (IR) applications by using LSI and many other techniques. Also, information retrieval can be improved by using two main approaches. The first one is by develop the IR system algorithms or by improve the query modeling such as query expansion [6].

Indu et al. 2014 [8] proposed such an automatic technique for bug labeling using semantic information from the Latent Semantic Indexing (LSI). Based on the TF-IDF and LSI, the identification of the appropriate developer such as accurate assessment of priority and severity of bugs issue, the proposed system produced such automatic bug labeling system. In this approach, the accuracy results have been reached to 62.8% of the security bug reports in comparison with 53.8% accuracy for polish. However, the system has achieved 61% using TF-IDF alone.

Xiong and Callan 2015 [6] presents a method to using Freebase as a base of knowledge in order to improving the expansion of user query. A proposed method investigates the two methods for identifying the entities that associated within the user query. Also, these entities of the two methods using in order to perform the expansion of user query. Experiments done on the dataset of ClueWeb09 with the queries of TREC Web Track illustrate that those methods are 30% further effective than other query expansion algorithms. In contrast, the disadvantage of this work that it focused only on the single dataset. Furthermore, methods used in this research are independent from each other and each method gives different results.

Safi et al. 2015 [9] suggested a linguistic query expansion based on a semantic user's profile that is suitable for the Arabic users. The proposed approach used Arabic WordNet and ontologies of the Arabic WordNet Amine to observe users' profile by means of the hierarchal concepts and expand the user query with the suitable concepts. The proposed system has been implemented and evaluated based on using 150 queries over 50 users for testing, 130 queries was improved the performance; while other queries (20 queries) were degraded, and 7500 quires over for evaluation.

Siriguleng et al. 2015 [10] proposed a Mongolian Semantic methodology of information retrieval based on Latent Dirichlet Allocation (LDA). The main contribution of this approach is producing a new methodology that has been on using the Mongolian documents by considering the keywords and retrieval documents relationships. The experimental results show that the proposed approach gets the higher performance in the topic semantic.

Nik et al. 2016 [11] proposed a Malay hadith translated document retrieval model based on parallel Latent Semantic Indexing (LSI) methodology. In this research, such a new methodology of LSI is proposed based on using standard multi-core computers (PC). Around 2028 text documents are extracted from different four volumes of the Malay-translated book of Hadith. The experimental results show that the time consuming to create the LSI space by suing the parallel system is significantly faster than the sequential system.

Dalton et al. 2014 [12] proposed a technique of the entity query features expansion (EQFE) that enriches a user query by using the features from the entities. They find that the results of entity depend on the feature expansion in the significant improvements in the effectiveness of retrieval more than the approaches of state-of-the text expansion.

Singh et al. 2017 [13] presented an approach of fuzzy logic depended on automatic query expansion (AQE) for the document retrieval and the techniques of pseudo-relevance feedback (PRF) via mining the supplementary terms for query expansion (QE). The method computes importance degrees of the relevant terms in order to find the additional query terms. They select different and several numbers of the high (top) candidate terms depended on the similarity value within the query terms. The experimental results achieve the significant improvements through every method of the individual term selection, aggregated method and the regarding method of state-of-the-art.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Singh and Sharan 2017 [14] explored the power of combining various methods of query expansion term selection in order to improve the possibility of the information retrieval system performance by using the automatic expansion of user query. They studied the rank combination of four query expansion term selection methods on two real datasets with or without semantic filtering and semantic genetic filtering approaches. The experimental results show that the proposed approaches achieved a significant improvement over each method of the individual query expansion term selection and related state-of-the-art approaches.

Lin et al. 2018 [15] proposed a method of the Bayesian Query Expansion (BQE) in order to generate a new query with information from an initial ranking list. For each query, the top ranked candidates in the initial ranking list are selected, and the features are pooled with the query using Bayesian model. The new query is then can be used in order to produce the new ranking list, with which the process of retrieval is performed again to obtain the final results. They evaluate BQE with multiple methods of feature extraction and methods of distance metric learning on four large-scale re-ID datasets. Experiments show that the proposed approach improves the performance of baselines and is robust to feature representation and metric learning methods. They observe a consistent improvement over all baselines and report competitive performances compared with the state-of-the- art results.

Lu et al. 2015 [16] proposed an approach in order to expand the user query within the synonyms that are generated from the WordNet. The proposed approach extracts the phrases of natural language from the identifiers of the source code. After that, the expanded queries will match these phrases. Finally, it sorts out the results of the search. The empirical study shows that synonyms are used in order to expand queries of the user which help to recommend the better alternative queries.

3. INFORMATION RETRIEVAL SYSTEM

Information Retrieval (IR) is the process that used to deals with the representation, storage, and access of information for the purpose of finding the most relevant documents as a response to the request of user [17]. Mainly, the system of information retrieval consists of major steps which have been used in order to retrieve relevant documents. These steps are the database collection, preprocessing, indexing and representation, dimension reduction, matching by using one of the IR algorithms, and IR evaluation [18]. Figure (1) shows the typical information system.

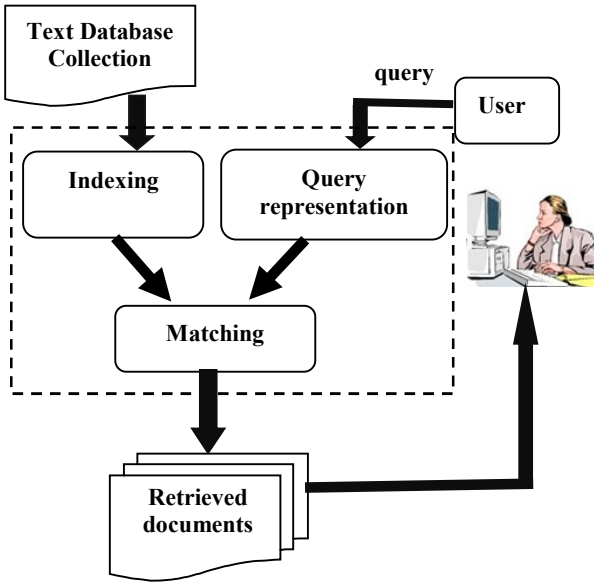


Figure (1): Information retrieval system

3.1 Text Database Collection

Text database collection is the first and an essential step of information retrieval system. Different types or different formats of documents are collecting like web pages, and reports that collected from multiple resources.

3.2 Preprocessing

Preprocessing is an essential and preparatory step to test the quality of the collected data before implementing the algorithms for information retrieval. A main goal of preprocessing is to convert the different data forms into more consistent form [19]. Preprocessing involves several sub-operations which are: Tokenization, normalization, stop words removal, and the stemming [20].

3.3 Indexing

Indexing is the process of representing the content of the documents [21] through sets of weighted terms which reflects the best summarize to the content of the given documents [22, 23, 24]. There are several popular techniques that have been used with indexing process such as Term

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

frequency- inverse document frequency (TF-IDF) technique [25].

TF-IDF scheme is designed to consider about the feature discriminative power within and over the documents [19]. Term frequency (TF) can be define as a factor which is measuring the terms frequency that occurrence in the document, and the inverse document frequency IDF which measuring an inverse of the documents number that contains in the document term [26, 27]. TF-IDF can be calculated via using a following equation:

$$TF - IDF = f_{ij} * log\left(1 + \frac{N}{N_i}\right) \quad (1)$$

where f_{ij} is the occurrence of the word i in the document j, while Ni is the frequency of document that contains the word i in the collection and N is the entire number of the documents in a collection [28].

3.4 Dimension Reduction

To produce the very compact representation of datasets, dimension reduction is used to reduce the components number of a dataset by representing the original data as narrowly as possible with a fewer number of features. So, the goal of the dimension reduction is to produce the very compact representation of data sets with the restricted loss about the information for reducing the storage and the runtime requirements [29].

3.5 Information Retrieval Algorithms

The similarity scores are computed between the given user query and the feature weights by using an algorithm. K-Nearest Neighbors (KNN), Latent Dirichlet Allocation (LDA), Support Vector Machines (SVM), and LSI represents as examples of the machine learning algorithms that can be used to build IR system [30]. In the following subsection will describe LSI algorithm that used in this paper.

- **Latent Semantic Indexing (LSI)**

LSI is the statistical method for the information retrieval which has been designed in order to help with a problem of the synonyms and polysemy words [10].

LSI is depending on the assumption which there is the structure of underlying semantic in the textual document data. This relationship among the features and the documents can be re-described in the form of semantic structure. Also, it tries to solve problems of the lexical matching via using statistically that derived the conceptual indicators instead of the individual words for the retrieval. LSI suppose that there is some of underlying or a latent structure in the word usage which is partly obscured via the variability in a word choice [11].

The LSI method can be described within five main steps:

1. Vector space model creation: the textual documents will represent as the vectors in the vector space (A). A matrix A is generated which is a vector space model (VSM) $(m \times n)$ matrix that has m features and n documents. Each cell in matrix A[i, j] will represents the frequency of the word occurrence i in a document j. Besides, the user query is also converted to vector (q). This vector contains frequency of occurrence of words from matrix A in the query. Here, q will be a word occurrence matrix of size $(m \times 1)$ [31].

2. Singular Value Decomposition (SVD): The rectangular matrix A within the order m×n generated from step (1) is decomposed into the three matrices (U, S, and V) as in equation (2). The matrix U which consists of the left eigenvectors of matrix A and it will describe the relationships among the features and the documents. This matrix is generated by using an equation U = A^TA. The matrix S is the m×n diagonal matrix together with the entries that sorted in the decreased order. The S matrix entries comprise the singular values (eigenvalues) which describe related strengths for every feature. V represent the matrix which is defined via using the equation V = A AT, which is contains the left eigenvectors of matrix A.

$$A_{m \times n} = U * S * V^T \quad (2)$$

3. Ranking: SVD is the method to reducing the feature space dimensionality without significant loss of the specificity. It is used in order for reducing the rank and a size of a text file. The reduced rank SVD implemented on a matrix that a

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

largest singular values of k are retaining while remainder values are set to 0 [32].

Let's assume that A be an $m \times n$ matrix, and let the rank of A be k. Then matrices of U, S, and V^T as shown in Figure (2) with the following properties [26, 33]:

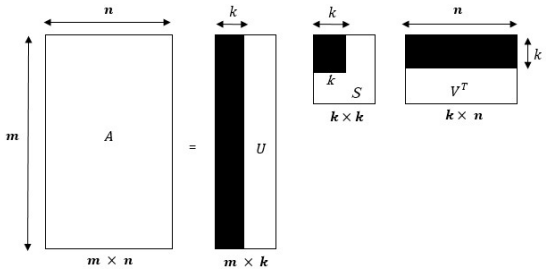


Figure (2): Singular-value decomposition form

The benefit of this reduction to the optimal rank k approximation is that the calculation of term similarity is only based on the most significant features of the document collection. The noise which blurs the clear view on the hidden relations is repressed.

4. Query vector transformation: To find the similarity degree between the vector of user query q and the documents matrix A after ranking to k, query vector should be transformed into a same space as the documents, hence it is mapped into the chosen k-dimensional space as obtained in the following equation (3):

$$q = q^T U_k S_k^{-1} \qquad (3)$$

5. Similarity score computation: The measure of cosine similarity is used in order to calculate a similarity degree among the vector of the user query and each document vector (which represented by the columns in V_k^T) as it is obtained in the following equation (4) [31]:

$$sim(q,d) = \frac{q.d}{|q||d|} \qquad (4)$$

Finally, the result is ranking based on the similarities.

3.6 Performance Evaluation

The major objective of the evaluation is to estimate the IR system performance. The evaluation measure allows estimating a system ability to retrieve the relevant documents and discard irrelevant ones [34].

There are two essential measures for estimating the text retrieval quality and evaluating search strategies [35, 36]. Precision is the fraction/ percentage of relevant documents among the retrieved documents. Precision can be calculated from the equation:

$$\text{Precision} = \text{relevant retrieved} / \text{total retrieved} \qquad (5)$$

Recall is the fraction/ percentage of the relevant documents which have been retrieved over the total amount of the relevant documents [36, 22]:

$$\text{Recall} = \text{relevant retrieved} / \text{total relevant} \qquad (6)$$

The F-measure which is compromise from both precision and recall in order to measuring the performance of text information retrieval [34, 37]:

$$\text{F-Measure} = 2 \times R \times P / (R + P) \qquad (7)$$

4. PROPOSED SYSTEM

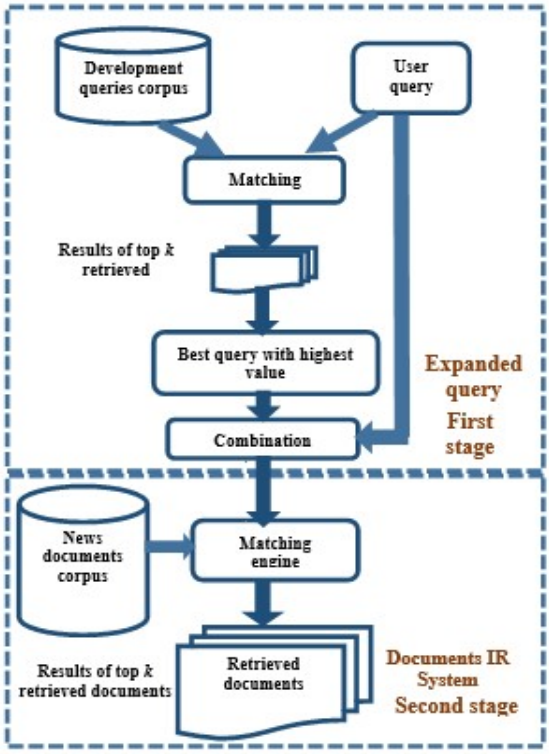
Information retrieval with a query formed by just a few features has become a big challenge. To overcome this problem, query expansion is a good suggested idea to increase the accuracy and performance of the information retrieval. So, the query expansion is the main idea that used in the proposed system. Our idea is that the query expansion will produce a more accurate hypothesis (query) by using a development query collection that collects from different news websites. The

Figu

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

framework of the proposed study consists of two main steps: query combination and query expansion to expand the user query, while the second step is documents IR system to identify the most relevant documents.

Figure (3) illustrate the main diagram of the proposed system for documents retrieval based on query expansion.



The first step is query combination which deals with the collection of queries. Different queries are collected from different website in order to use it as specific development dataset. The proposed system takes advantage of the strength of specific domain of the related dataset of the queries. The main idea behind query combination is to increase the number of features in order to get better matching between the given user query and the specific domain dataset of queries. A good query becomes very important to find the relevant documents. The LSI is used to find the best matching to the user query which will be combined as expanded query. LSI is chosen because it is a statistical approach based on SVD algorithm to find semantic concepts between the terms. LSI method is used in this work as IR system because it is a powerful method of document indexing which organizes the information from documents into the semantic structure.

The first step of the proposed system tries to locate the best candidates for the user query by searching inside the related queries. Searching for the more likely matching query in the related development queries with the high score of similarity are used to select as suggested candidate. For example, assume that the q and Q is the given user query and the development dataset respectively. Q= {q_1, q_2, q_3, …, q_n}. Also, suppose the best candidates for q is q_3 with high similarity score. Thus, (q) and (q_3) are combined together by using and logic to create new query ($Query_{new}$) with more features. Figure (4) show the first stage.

Later, the expanded query ($Query_{new}$) is used instead of the input user query (q) in order to retrieve the documents that needs by the user.

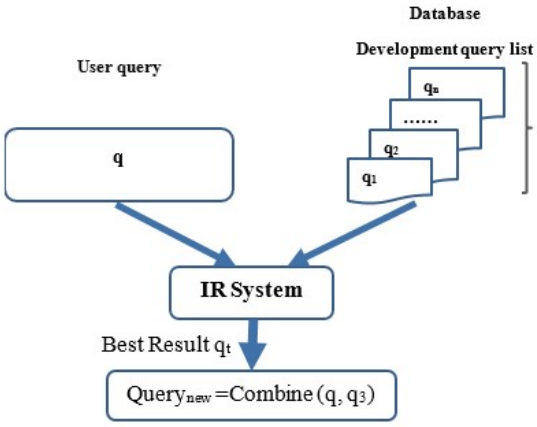


Figure (4): An example of the proposed system with combination

In the second stage, the matching between the queries expansion and the collection of the documents are accomplished by using LSI algorithm. Figure (5) shows the documents IR System.

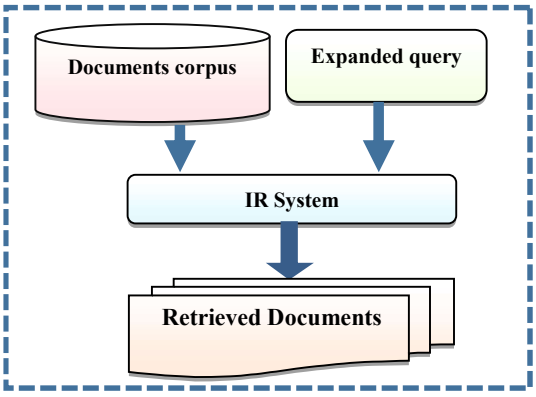


Figure (5): Documents IR System

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

To recap this stage, the query of user will compare against a collection of development queries to find the most relevance one. After that, the query expansion is done automatically by using the user query and the specific development query that results from the comparison of the queries with highest similarity value. While the second stage is using the expanded query instead of user query in order to retrieve the most relevant documents.

5. EXPERIMENTAL RESULTS

This section will discuss the results which get from the proposed system. These results have been tested and compared with results that obtained from the traditional IR system using LSI algorithm.

Databases that used in this work contains (500) English news article documents. These databases consist of four datasets. Three of this datasets are collected from British Broadcasting Corporation (BBC) news site which are: 1) Text news articles that contains (500) documents of English news articles, which has been manually collected to be used as a document corpus for information retrieval. 2) Queries list which represent a collection of queries (usually statements) that used by the user to find the information they need. This list contains (100) queries. 3) Gold standard which consists of the manual retrieval of the news documents. These documents represent the correct answer for the user query, while the latest dataset which called Development queries list represents the new queries that are collected to enhance the IR performance. This dataset is used in the propose study in order to find the most similar query to t user query and then it will use to improve the I result via query expansion. The developmen queries list is collected from other different websites which are Cable News Network (CNN), The Daily Telegraph news, al-jazeera English news, The Daily Mail news, and Independent news website. The queries list, and development queries datasets are collected manually. In additional, the gold standard dataset is evaluated manually based on the list of queries in order to use it later to evaluate the IR system performance.

LSI with information retrieval was applied. First LSI is applied without query expansion. LSI is divided for four steps which are: preprocessing, dimension reduction, SVD construction and computing the similarity by using the cosine metric. All the steps are explained in previous sections.

The preprocessing step involves several operations which are: tokenization, removed stop words, and stemming. In this study, a stop word list which contains (627) words was used. Also, the Stanford stemmer [37] is used which is a powerful and context sensitive suffix removal algorithm.

In this study, different *k* ranks were used to reduce the feature space. Ranks of (100, 150, 175, 200) were used and the best results was with the rank of (150). The following table (1) shows the results of *k* ranks with the threshold of (0.4).

Table (1) The results of k ranks with the threshold of (0.4)

***k* ranks**	**Precision**	**Recall**	**F-measure**
100	0.58	0.74	0.65
150	**0.70**	**0.71**	**0.70**
175	0.75	0.65	0.69
200	0.76	0.60	0.67

Also, the current study investigates the effect of different thresholds such as (0.3, 0.4, 0.5, 0.6, and 0.7) and the best result was with the threshold of (0.4). The following Table (2) shows the baseline system results with rank k of (150) and different thresholds.

Table (2) The baseline system results with ranks k of (150) and different threshold

Threshold	**Precision**	**Recall**	**F-measure**
0.3	0.46	0.84	0.59
0.4	**0.70**	**0.71**	**0.70**
0.5	0.65	0.49	0.55
0.6	0.44	0.31	0.36
0.7	0.33	0.18	0.23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

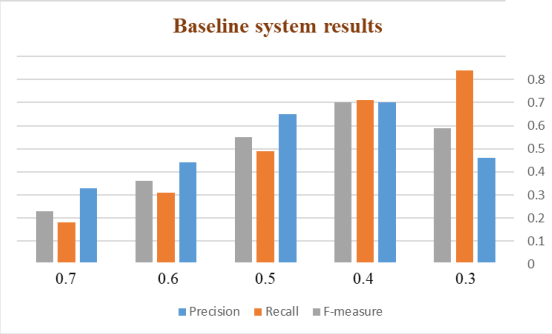


Figure (6): Baseline system results

Next, the performance of the proposed approach for enhancing the retrieval system output via the query expansion by using development queries list is applied. The query expansion comprises of two main steps:

A. IR system that based on LSI algorithm is build using the user queries database. Then, the development queries dataset is used to find the most similar matching between the user query and the development queries dataset. The query that have the highest score of similarity is selected for expand the user query. The selected query is combined automatically with user query by using the logic "AND" to generate the expanded query.

B. The expanded query is used to retrieve the documents instead of the user query. In this step, the extended queries are used as new input in the LSI retrieval system.

The proposed system used (100) queries for development to improve the IR system performance. Form the total (100) queries, only (40) queries retrieve more relevant documents with higher degree of similarity and F-measure scores compared to the LSI system without query expansion. While the remainder queries produced same results but with different degrees of similarity and the number of the relevant document was more specific.

The proposed system gets better results compared to the LSI system without expansion. The best results are obtained with the rank of (150) and threshold of (0.4). The following Table (3) shows the proposed system results with rank k of (150) and different thresholds.

Table (3) The proposed system results with ranks k of (150) and different threshold

Threshold	Precision	Recall	F-measure
0.3	0.64	0.79	0.71
0.4	**0.75**	**0.74**	**0.75**
0.5	0.74	0.63	0.68
0.6	0.66	0.56	0.61
0.7	0.57	0.50	0.53

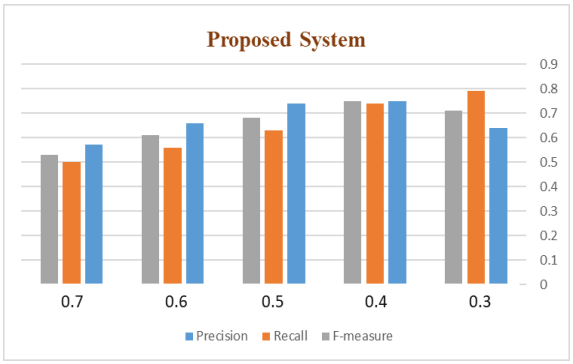


Figure (7): Proposed system results

Compared to LSI without query expansion, the proposed approach shows that LSI algorithm with expanded query improves the retrieval performance on English news. The results which are found by using the expanded query represent the quality of the search that could expect to be able to retrieve documents that are wanted to maintain performance better than that of the baseline retrieval system. In addition, LSI based on query expansion exhibits much better performance.

In general, the results prove that LSI and the expanded query significantly outperform than the LSI algorithm with user query (original query).

The results show that LSI method with expanded query (P= 0.75, R= 0.74, F-measure= 0.75) is more significant compared to LSI algorithm without expanded query (P= 0.70, R= 0.71, F-measure= 0.70).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

6. CONCLUSIONS

In this paper, an approach of a query expansion was proposed by making the use of LSI algorithm to retrieve the most relevant documents and to increase the performance. Thus, the essential goal of the current proposed system was investigating the effect of different queries from the development queries list by using a form of an expanded query to retrieve the documents of English news.

The results show that the query expansion displays the highest F-measure because it does take into consideration the short query problem. The results collected from IR system with query expansion provide the evidence that extending the short query gives better performance compared to the IR system without query expansion that used the short query as input. The IR system with a query expansion achieved an F-measure of 0.75; without the query expansion, it achieved an F-measure of 0.70. Generally, Experimental results show that the proposed framework improves the retrieval of English news articles.

7. LIMITATIONS OF THE STUDY:

The study has many limitations which can be listed as below:

1- The proposed approach was tested only on the English news.

2- The approach tested only on a small dataset of English news which is broadcasted in 2017.

3- The manual collection of text news articles corpus.

8. FUTURE WORKS

For the future works, there are several suggestions based on the current study such as:

4- Employing a technique that preserves the top similar queries which results and use it as a database for the entered user query such as clustering techniques.

5- The proposed system also can be used with other languages such as Arabic via using the preprocessing which can be appropriate for it.

6- The proposed system can be used to retrieved documents by searching inside classified news to enhance the IR performance, reduces the time, and computational operations instead of searching inside unclassified news.

9. REFERENCES

[1] C. Carpineto and G. Romano, 2012, *"A survey of automatic query expansion in information retrieval."* in Proceedings of ACM Computing Surveys (CSUR), Vol. 44, issue 1, pp. 1-50.

[2] B. Haddow and P. Koehn, 2012, *"Analysing the Effect of out-of-domain Data on SMT Systems,"* in Proceedings of the Seventh Workshop on Statistical Machine Translation, Association for Computational Linguistics, pp. 422-432.

[3] Z. A. Khalaf, 2015, *"Broadcast News Segmentation Using Automatic Speech Recognition System Combination with Rescoring and Noun Unification,"* PhD Thesis, Universiti Sains Malaysia.

[4] Cai, Fei, and Maarten De Rijke. *"A survey of query auto completion in information retrieval."* Foundations and Trends® in Information Retrieval 10.4 (2016): 273-363.

[5] A. Addis, 2010, *"Study and Development of Novel Techniques for Hierarchical Text Categorization,"* PhD Thesis, Dept. of Electrical and Electronic Engineering, University of Cagliari.

[6] C. Xiong and J. Callan, 2015, *"Query Expansion with Freebase,"* in Proceedings of the International Conference on The Theory of Information Retrieval, ACM, pp. 111-120.

[7] B. He and I. Ounis, 2009, *"Studying Query Expansion Effectiveness,"* in Proceedings of the European Conference on Information Retrieval, Springer, pp. 611-619.

[8] Chawla. Indu, and Sandeep K. Singh. *"Automatic bug labeling using semantic information from LSI."* Contemporary Computing (IC3), 2014 Seventh International Conference on. IEEE, 2014.

[9] S. Houssem, J. Maher, and B. H. Lamia, 2015, *"Axon: a Personalized Retrieval Information System in Arabic Texts Based on Linguistic Features,"* in Proceedings of the 6th International Conference on Information Systems and Economic Intelligence (SIIE), IEEE, pp. 165-172.

[10] V. T. Turunen and M. Kurimo, 2006, *"Using Latent Semantic Indexing for Morph-based Spoken Document Retrieval,"* in Proceedings of the Ninth International Conference on Spoken Language Processing, pp. 341-344.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[11] N. N. Amirah, T. M. Rahim, Z. Mabni, H. M. Hanum, and N. A. Rahman, 2016, *"A Malay Hadith translated Document retrieval using parallel Latent Semantic Indexing (LSI),"* in Proceedings of the Third International Conference on Information Retrieval and Knowledge Management, IEEE, pp. 118-123.

[12] J. Dalton, L. Dietz, and J. Allan, 2014, *"Entity Query Feature Expansion Using Knowledge Base Links,"* in Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 365-374.

[13] J. Singh, M. Prasad, Y. A. Daraghmi, P. Tiwari, P. Yadav, N. Bharill, M. Pratama, and A. Saxena, 2017, *"Fuzzy Logic Hybrid Model with Semantic Filtering Approach for pseudo Relevance feedback-based Query Expansion,"* in Proceedings of the IEEE Symposium Series on in Computational Intelligence (SSCI), IEEE, pp. 1-7.

[14] Singh, Jagendra, and Aditi Sharan. *"Rank fusion and semantic genetic notion based automatic query expansion model."* Swarm and Evolutionary Computation 38 (2018): 295-308.

[15] Lin, Yutian, et al. "Bayesian Query Expansion for Multi-Camera Person Re-identification." Pattern Recognition Letters (2018).

[16] M. Lu, X. Sun, S. Wang, D. Lo, and Y. Duan, 2015, "Query Expansion Via Wordnet for Effective Code Search," in Proceedings of the IEEE 22nd International Conference on Software Analysis, Evolution and Reengineering (SANER), IEEE, pp. 545-549.

[17] J. Wang, J. Liu, R. Higgs, L. Zhou, and C. Zhou, 2017, *"The Application of Data Mining Technology to Big Data,"* in Proceedings of the IEEE International Conference on Computational Science and Engineering (CSE) and Embedded and Ubiquitous Computing (EUC), IEEE, pp. 284-288.

[18] W. B. Croft, D. Metzler, and T. Strohman, 2010, *"Search Engines: Information Retrieval in Practice,"* Addison-Wesley Reading.

[19] V. Korde and C. N. Mahender, 2012, *"Text Classification and Classifiers: A Curvey,"* International Journal of Artificial Intelligence & Applications, Vol. 3, issue. 2, pp. 85-99.

[20] C. A. Gonçalves, C. T. Gonçalves, R. Camacho, and E. C. Oliveira, 2010, *"The Impact of Pre-processing on the Classification of MEDLINE Documents,"* in PRIS, pp. 53-61.

[21] I. Ruthven and M. Lalmas, 2003, *"A Survey on the Use of Relevance Feedback for Information Access Systems,"* The Knowledge Engineering Review, Vol. 18, issue. 2, pp. 95-145.

[22] M. Sharma and R. Patel, 2013, *"A Survey on Information Retrieval Models, Techniques and Applications,"* International Journal of Emerging Technology and Advanced Engineering, Vol. 3, issue. 11, pp. 542-545.

[23] P. D. Turney and P. Pantel, 2010, *"From Frequency to Meaning: Vector Space Models of Semantics,"* Journal of Artificial Intelligence Research, Vol. 37, pp. 141-188.

[24] N. Polettini, 2004, *"The Vector Space Model in Information Retrieval-Term Weighting Problem,"* Entropy, Citeseer, pp. 1-9.

[25] J. Geiss, 2011, *"Latent Semantic Sentence Clustering for multi-document Summarization,"* University of Cambridge, Computer Laboratory.

[26] Khalaf, Zainab Ali, and Tan Tien Ping. *"Novel Noun Pronunciation Unification Approach to Improve Story Boundary Identification in the Transcription of Malay News Broadcasts."* IJCSA 11.1 (2014): 37-55.

[27] A. Roshdi and A. Roohparvar, 2015, *"Information Retrieval Techniques and Applications,"* International Journal of Computer Networks and Communications Security, Vol. 3, issue. 9, pp 373-377.

[28] N. Alsaedi, P. Burnap, and O. Rana, 2016, *"Temporal TF-IDF: A High Performance Approach for Event Summarization in Twitter,"* in Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI), IEEE, pp. 515-521.

[29] T. Berka and M. Vajteršic, 2014, *"Dimensionality Reduction for Information Retrieval Using Vector Replacement of Rare Terms,"* in Data Mining for Service, Springer, pp. 41-60.

[30] W. H. Leung and T. Chen, 2003, *"Hierarchical Matching for Retrieval of hand-drawn Sketches,"* in Proceedings of the International Conference on Multimedia and Expo 2003 (ICME'03), IEEE, Vol. 2, pp. 1-29.

[31] D. Bhatia, 2016, *"The Power of Latent Semantic Indexing in Review Retrieval,"* Master Thesis, University of Texas Tech.

[32] C. Yu, 2013, *"Using Latent Semantic Indexing for an Online Research Interest Matching System,"* in Proceedings of the International Conference on Advanced Information Engineering and Education Science (ICAIEES), Atlantis Press, pp. 109-112.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[33] V. Snasel, P. Moravec, and J. Pokorny, 2005, *"WordNet Ontology Based Model for Web Retrieval,"* in Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration (WIRI'05), IEEE, pp. 220-225.

[34] L. Tamine-Lechani, M. Boughanem, and M. Daoud, 2010, *"Evaluation of Contextual Information Retrieval Effectiveness: Overview of Issues and Research,"* Knowledge and Information Systems, Springer, Vol. 24, issue. 1, pp. 1-34.

[35] R. S. Prerna and P. Bhadana, 2013, *"News Retrieval Based on Latent Semantic Index And Clustering"*, International Journal of Scientific and Technology Research, Vol. 2, issue. 4, pp. 215-217.

[36] R. T. Selvi and E. G. D. P. Raj, 2014, *"An Approach to Improve Precision and Recall for Ad-hoc Information Retrieval Using SBIR Algorithm,"* in Proceedings of the World Congress on Computing and Communication Technologies (WCCCT), IEEE, pp. 137-141.

[37] Khalaf, Zainab Ali, and Tan Tien Ping. 2015, *"MAHIR System: Unsupervised Segmentation for Malay Spoken Broadcast News Stories."* International Journal of Information and Electronics Engineering, Vol 5, issue. 3, pp. 211-215.

[38] Stanford stemmer ((https://github.com/stanfordnlp/CoreNLP))