Performance Comparison between Keyword-based and WQCA-based Information Retrieval System

Naw Thiri Wai Khin[1], **Nyo Nyo Yee**[2]

[1, 2]Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar

Abstract: ***Today, semantic logics are very important in query understanding to create successful web search engines. A user might not formalize the query when he seeks information although he knows what he wants. As a result, understanding the nature of the information that is needed behind the queries are important research problem. So, this system proposes the Web Query Classification Algorithm (WQCA) for efficient Information Retrieval (IR) system. In the WQCA process, this system firstly classifies the web queries into each characteristic (taxonomies). Then, this system extracts the domain terms from the query. By using NoSQL graph database, this system classifies each domain term into their relevant categories according to the WQCA algorithm. In the WQCA-based IR process, this system uses the classified query to find the relevant document form the document collection. Finally, this system compares the performance between keyword-based IR and WQCA-based IR to show the effectiveness of the web query classification.***

Keywords: web query, domain term extraction, WQCA, NoSQL graph database, information retrieval

1. Introduction

Information Retrieval (IR) systems provide populations of users with access to a large collection of stored information. These systems are concerned with the structure, analysis, organization, storage, and searching of such information. A good IR system is able to accept a user query, understand from the user query what the user requires, search a database for relevant documents, retrieve the documents to the user, and rank the documents according to their relevance.

The keyword-based Information Retrieval (IR) system uses the vector space model to retrieve user query relevant information. The use of the vector inner product is the measure of similarity between the query and a document. The semantic indexing of the document changes from the keyword-based approach to the semantic based approach for effective retrieval. IR system will improve its performance if the documents which retrieve are represented by relevant category rather than words and it can potentially benefit from the correct meanings of words provided by query classification method.

The semantic-based information retrieval system eliminates the possibility of retrieving information that obtained the presence of the irrelevant information because of provision of the correct relevant category of the word in the searching process.

Web query classification based IR system is one type of semantic based IR system. So, this system proposes the Web Query Classification Algorithm (WQCA). Based on the concept term strategy and NoSQL graph database, this system uses the WQCA to classify query characteristics and the ambiguous domain terms. Using classified user query, this system performs the information retrieval process. In the web query classification based IR system, the WQCA and vector space model are used to retrieve user query relevant information. According to the concept terms analysis results, this system becomes a good IR system by extracting documents which are more relevant to user's requirements. To show the better performance of the WQCA based IR system, this system compares the keyword based IR system.

The rest of the paper is organized as follows: related work is described in section 2. Web query characteristics and classification are shown in section 3 and 4. NoSQL graph database is described in section 5. In section 6, vector space model (VSM) is presented. Then, the proposed system design and explanation of the system are described in section 7 and 8. Finally, experimental result of the system and conclusion are given in section 9 and 10.

2. Related Work

In 2012, S. M. Fathalla and Y. F. Hassan [1] presented hybrid method for user query reformation and classification depending on fuzzy semantic-based approach and K-Nearest Neighbour (KNN) classifier. The overall processes of the system are query pre-processing, fuzzy membership calculation, query classification and reformation. Classification is performed using KNN classifier not just by keyword-based semantic but using a sentence-level semantics. After classification, user's query is reformulated to be submitted to a search engine which gives better results than submitting the original query to the search engine. Experiments show significant enhancement on search results over traditional keyword-based search engines' results.

In 2015, A. Katariya [2] presented ontology based web query classification. Query classification is one technique in which query should classify to the number of predefined categories. Query classification use ontology as a model to classify the input search queries. Ontology stores a set of concepts and semantic rules to classify user queries.

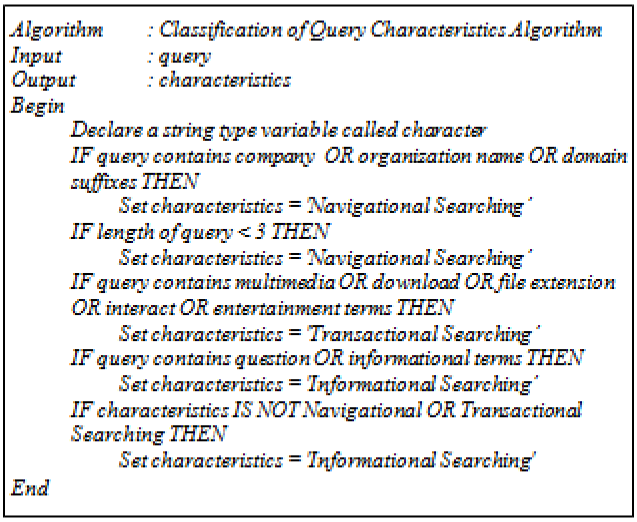
In 2006, W. Yue, Z.Chen and X. Lu [3] proposed a novel information retrieval algorithm based on query expansion and classification. The algorithm is induced by the observation that very short queries with the traditional information retrieval methods often have low precision, although they can get high recall. Their approach attempted to catch more relevant documents by query expansion and text classification. The results of the experiments showed that the proposed algorithm is more precise and efficient than the traditional query expansion methods.

3. Web Query Characteristics

In the web context, the "need behind the query" is often not informational in nature. Web queries are classified according to their intent into three classes:

- Navigational: The immediate intent is to reach a particular site. Navigational searching queries contain organization, business, company name, universities name, domain suffixes (eg. ".com", ".org") and domain prefixes (eg. "www", "http", "web"). Some Navigational queries contain URLs or parts of URLs.
- Informational: The intent is to acquire some information assumed to be present on one or more web pages. Queries for such search may consist of informational terms like "list" and "playlist" etc., question words like "who", "what", "when" etc.
- Transactional: The intent is to perform some web-mediated activity. Queries containing "audio", "video" and "images" are considered to be transactional [4], [5].

Algorithm for classification of query characteristics is presented in Figure 1. The input for algorithm is a user query and the output is classified characteristics. The input query is classified into the specific characteristics: Navigational, Transactional or Informational Searching, according to the If… Then… rules.



```
Algorithm    : Classification of Query Characteristics Algorithm
Input        : query
Output       : characteristics
Begin
        Declare a string type variable called character
        IF query contains company OR organization name OR domain
        suffixes THEN
              Set characteristics = Navigational Searching'
        IF length of query < 3 THEN
              Set characteristics = Navigational Searching'
        IF query contains multimedia OR download OR file extension
        OR interact OR entertainment terms THEN
              Set characteristics = Transactional Searching'
        IF query contains question OR informational terms THEN
              Set characteristics = Informational Searching'
        IF characteristics IS NOT Navigational OR Transactional
        Searching THEN
              Set characteristics = Informational Searching'
End
```

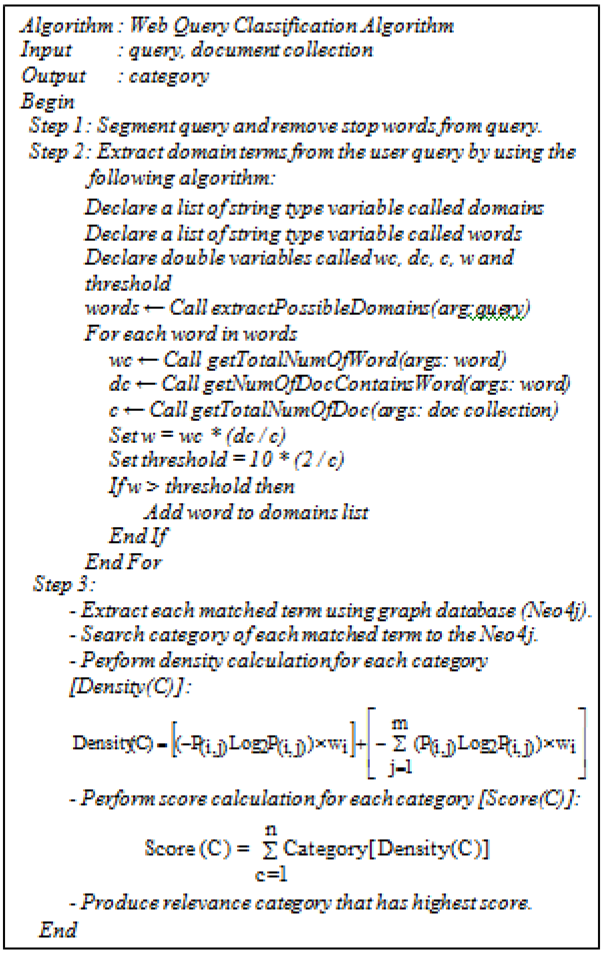
Figure 1: Classification of Query Characteristics Algorithm

4. Web Query Classification

Web query classification is significant to search engines for the purpose of efficient retrieval of appropriate results in response to user queries. User queries are short in nature, contain noise and are ambiguous in terms of user intent. Web query classification is to classify a user query Q_i into a list of n categories c_{i1}, ci_2, … , c_{in} [6].

Web query classification includes three step processes. The first process is domain term extraction that is a categorization or classification task in which terms are categorized into a set of predefined domains [8]. The second process is learning step where a classification model is constructed. The third process is classification step where the model is used to predict class label for given data. If a certain category in an intermediate taxonomy is given, web query classification is directly mapped to a target category if and only if the following condition is satisfied: one or more terms in each node along the path in the target category appear along the path corresponding to matched intermediate category [7].

In the web query classification process, the input is the user query and the output is the relevance category that has the highest score. The web query classification algorithm (WQCA) is shown in Figure 2.



```
Algorithm : Web Query Classification Algorithm
Input     : query, document collection
Output    : category
Begin
 Step 1: Segment query and remove stop words from query.
 Step 2: Extract domain terms from the user query by using the
         following algorithm:
         Declare a list of string type variable called domains
         Declare a list of string type variable called words
         Declare double variables called wc, dc, c, w and
         threshold
         words ← Call extractPossibleDomains(arg: query)
         For each word in words
            wc ← Call getTotalNumOfWord(args: word)
            dc ← Call getNumOfDocContainsWord(args: word)
            c ← Call getTotalNumOfDoc(args: doc collection)
            Set w = wc * (dc / c)
            Set threshold = 10 * (2 / c)
            If w > threshold then
                Add word to domains list
            End If
         End For
 Step 3:
     - Extract each matched term using graph database (Neo4j).
     - Search category of each matched term to the Neo4j.
     - Perform density calculation for each category
     [Density(C)]:
       $Density(C) = \left[(-P_{(i,j)}Log_2P_{(i,j)}) \times w_i\right] + \left[-\sum_{j=1}^{m}(P_{(i,j)}Log_2P_{(i,j)}) \times w_i\right]$
     - Perform score calculation for each category [Score(C)]:
           $Score\,(C) = \sum_{c=1}^{n} Category[Density(C)]$
     - Produce relevance category that has highest score.
 End
```

Figure 2: Web Query Classification Algorithm (WQCA)

5. NoSQL Graph Database

Neo4j is a high performance NoSQL graph database which provides object oriented, flexible network structure. It is based on a Property graph data model which comprises of nodes and relationship along with their properties. It is reliable, ACID compliant, highly available and scalable. It offers REST interface and Java API quiet convenient to use. It can also be embedded into jar files. It uses CYPHER as its query language. Neo4j must be used in software involving complex relationships like social networking, recommendation engines etc. Neo4j must be avoided if relationships do not exist among the data. Some of the fortune 500 companies that use Neo4j are Adobe, Accenture, Cisco, Lufthansa, Telenor and Mozilla [9].

6. Vector Space Model (VSM)

The vector space model (VSM) is one of the classical and widely applied information retrieval models to rank the web page based on similarity values. The vector space model represents documents and queries as vectors in multidimensional space, whose terms are used as dimensions to build an index to represent the documents. Each dimension corresponds to separate term. If a term occurs in the document, its value in the vector is non-zero. It is used in information retrieval, indexing and relevant ranking and can be successfully used in evaluation of web search engines. The vector space model procedure can be divided into three stages. These are as follows:

- The first stage is the document indexing where content bearing terms are extracted from the document text.
- The second stage is the weighting of the indexed terms to enhance retrieval of documents relevant to the user.
- In the last stage, rank the documents with respect to the query according to similarity values.

The term frequency – inverse document frequency also called as TF-IDF, is a well-known method to evaluate how important is a word in a document. TF-IDF is also a very interesting way to convert the textual representation of information into a vector space model (VSM). TF-IDF term weight is given as follows:

$$w_{ij} = tf_{ij} \times idf \qquad (1)$$

$$idf_i = \log \frac{N}{df_i} \qquad (2)$$

$$f_{ij} = \frac{f_{ij}}{\max\{ f_{1j}, f_{2j}, ..., f_{|v|j} \}} \qquad (3)$$

In this scheme, N is total number of documents in the system. The df_i is number of documents in which term t_i appears at least once. The f_{ij} is the raw frequency count of term t_i in document d_j. Then, the tf_{ij} is the normalized term frequency of t_i in d_j.

The term weight w_{iq} of each term t_i in q can also be computed in the same way as in a normal document. Weighting scheme for query is as follows:

$$w_{iq} = \left[0.5 + \frac{0.5 f_{iq}}{\max\{ f_{1q}, f_{2q}, ..., f_{|v|q} \}} \right] \times \log \frac{N}{df_i} \qquad (4)$$

Dice similarity method measures the similarity between the document vector d_j and the query vector q. Dice similarity method is as follows:

$$sim(d_j, q) = \frac{2 \left| \sum_{i=1}^{|v|} w_{ij} \times w_{iq} \right|}{\sum_{i=1}^{|v|} (w_{ij})^2 + \sum_{i=1}^{|v|} (w_{iq})^2} \qquad (5)$$

7. Proposed System Design

The proposed system compares the performance between keyword-based IR and WQCA-based IR process. In this system, the user can choose the desired IR process to retrieve the relevant information. In the keyword-based IR process, the relevant information is retrieved by using the vector space IR model that is based on the TF-IDF weighting scheme and similarity method. But, the WQCA-based IR process retrieves the relevant information by using both web query classification algorithm and vector space IR model.

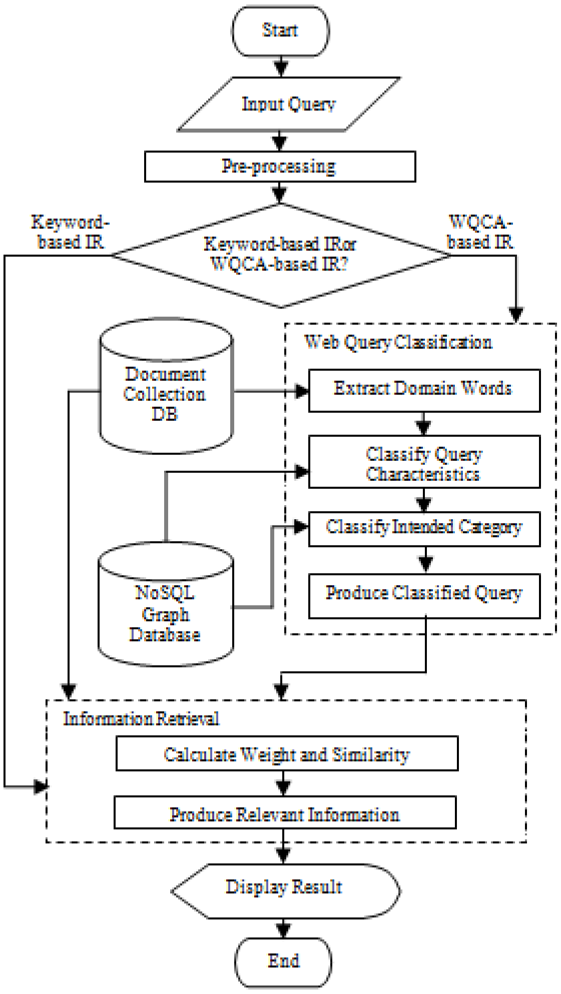


Figure 3: Proposed System Design

In this WQCA-based IR process, this system classifies each extracted domain terms into each categories by using WQCA algorithm. Then, this system retrieves the user query relevant information by using classified query. Proposed system design is shown in Figure 3.

8. Explanation of the System

In the performance comparison, this system uses the classified query and the original user query for each WQCA-based IR and keyword-based IR process. By using WQCA algorithm, this system classifies the original user query into the classified query. In the WQCA process, this system first defines the query characteristics category. Then, each domain terms are extracted from the user query. By using NoSQL graph database that stores matched terms and related category, this system classifies the most relevant category for the domain term. Among the relevant categories, the most relevant category has the highest score result. After performing the classification process, this system produces the classified query that includes each keyword, domain term and its relevant category. Table 1 shows the original user query and classified query.

Table 1: Original User Query and Classified Query

ID	*Original User Query*	*Query Characteristics*	*Classified Query*
1	comparison between RDBMS and graph database	Informational	comparison between RDBMS and graph database - Database Management System
2	survey about neural network	Informational	survey about neural network - Artificial Intelligence
3	cluster analysis classification techniques using matlab	Informational	cluster analysis classification techniques using matlab - Data Mining
4	Network Security Principles and Practices from Wikipedia	Navigational	Network Security Principles and Practices from Wikipedia - Computer Networking
5	Learning PHP programming form w3school.com	Navigational	Learning PHP programming form w3school.com – Programming Language
6	AES encryption algorithm source code	Transactional	AES encryption algorithm source code - Cryptography
7	Download speech recognition software	Transactional	Download speech recognition software - Digital Signal Processing

By using the classified query, the WQCA-based IR process can more retrieve query relevant information than the keyword-based IR process. Table 2 and 3 show WQCA and Keyword based information retrieval results for user query "cluster analysis classification techniques using matlab", which intends to data mining category and informational searching. According to the retrieval results, the WQCA-based IR process can retrieve the most query relevant document. So, it is more precise than the keyword based IR process.

Table 2: WQCA-based Information Retrieval Results

ID	*Document*	*Category*	*Characteristics*	*Similarity Value*
1	Data Mining Cluster Analysis.htm	Data Mining	Informational	0.27609
2	What is classification in data mining - Quora.htm	Data Mining	Informational	0.27726
3	Cluster analysis - Wikipedia.htm	Data Mining	Navigational	0.24548
4	MATLAB for Image Processing.pdf	Image Processing	Transactional	0.20782
5	Data Mining - Concepts and Techniques.pdf	Data Mining	Transactional	0.14138
6	Web query classification - Wikipedia.htm	Data Mining	Navigational	0.16104
7	Intro to Digital Image Processing - MATLAB and Simulink.html	Image Processing	Informational	0.13974

Table 3: Keyword-based Information Retrieval Results

ID	*Document*	*Category*	*Characteristics*	*Similarity Value*
1	MATLAB for Image Processing.pdf	Image Processing	Transactional	0.39999
2	Cluster analysis - Wikipedia.htm	Data Mining	Navigational	0.39000
3	Data Mining Cluster Analysis.htm	Data Mining	Informational	0.34223
4	What is classification in data mining - Quora.htm	Data Mining	Informational	0.23897
5	Intro to Digital Image Processing - MATLAB and Simulink.html	Image Processing	Informational	0.2165
6	Web query classification - Wikipedia.htm	Data Mining	Navigational	0.19954
7	Digital Signal Processing Using MATLAB and Wavelets.htm	Signal Processing	Informational	0.18833

9. Experimental Result of the System

To evaluate the performance of IR system, the precision, recall and f-measure methods are used as shown in equation (6), (7) and (8). Where, TP denotes the number of relevant documents in retrieved documents. FP is the number of non-relevant documents in retrieved documents. FN denotes the number of relevant documents in non-retrieved documents.

$$\text{Precision (P)} = TP / (TP + FP) \tag{6}$$

$$\text{Recall (R)} = TP / (TP + FN) \tag{7}$$

$$\text{F-measure (F)} = 2 \times [(P \times R) / (P + R)] \tag{8}$$

For the experimental results, this system is tested by using different ambiguous query. For training data, this system used 550 documents that are relevant 22 categories. These training documents are different file types that include ".doc", ".pdf" and ".html". The training categories and documents are collected from the Wikipedia category source and Google search engine. The precision, recall and F-measure results for WQCA-based IR process are shown in Table 4. Experimental results about WQCA-based IR process are shown in Figure 4.

Table 4: Precision, Recall and F-measure Results about WQCA-based IR

ID	*Category Name*	*P*	*R*	*F*
1	Analysis of Parallel Algorithm	0.937	0.96	0.948
2	Artificial Intelligence	0.884	0.996	0.936
3	Business Application	0.846	1	0.916
4	Computer Architecture	0.797	0.928	0.851
5	Computer Networking	0.833	0.93	0.877
6	Cryptography	0.853	0.964	0.902
7	Data Mining	0.884	1	0.937
8	Data Structure	0.701	1	0.82
9	DBMS	0.923	1	0.959
10	Digital Signal Processing	0.884	0.964	0.919
11	Distributed System	0.779	1	0.875
12	Electronic Circuits	0.867	1	0.928

13	Embedded System	0.9	0.968	0.932
14	Human Computer Interaction	0.895	1	0.943
15	Image Processing	0.862	1	0.925
16	Information System	0.682	1	0.809
17	Network Security	0.839	0.946	0.884
18	Operating System	0.904	1	0.948
19	Programming Language	0.864	0.972	0.913
20	Software Engineering	0.88	0.984	0.928
21	Web Application	0.879	0.976	0.919
22	Windows Application	0.826	0.948	0.882

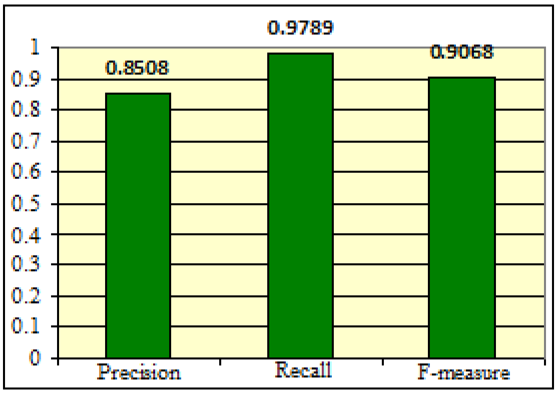


Figure 4: Experimental Results about WQCA-based IR Process

The precision, recall and F-measure results for keyword-based IR process are shown in Table 5. Experimental results for keyword-based IR process are shown in Figure 5. We can see the average F-measure values of all categories for all tested queries are under 0.7. It means that the combination of exactness and completeness is not over 70% in keyword based IR process.

Table 5: Precision, Recall and F-measure Results about Keyword-based IR Process

ID	*Category Name*	*P*	*R*	*F*
1	Analysis of Parallel Algorithm	0.796	0.86	0.823
2	Artificial Intelligence	0.576	0.62	0.577
3	Business Application	0.614	0.596	0.552
4	Computer Architecture	0.327	0.764	0.423
5	Computer Networking	0.636	0.758	0.669
6	Cryptography	0.662	0.576	0.54
7	Data Mining	0.539	0.58	0.499
8	Data Structure	0.424	0.948	0.564
9	DBMS	0.676	0.932	0.782
10	Digital Signal Processing	0.638	0.476	0.527
11	Distributed System	0.69	1	0.813
12	Electronic Circuits	0.706	0.928	0.798
13	Embedded System	0.594	0.564	0.525
14	Human Computer Interaction	0.508	0.748	0.562
15	Image Processing	0.718	0.88	0.773
16	Information System	0.436	0.932	0.592
17	Network Security	0.619	0.582	0.559
18	Operating System	0.655	0.576	0.537
19	Programming Language	0.548	0.88	0.666
20	Software Engineering	0.63	0.956	0.754
21	Web Application	0.652	0.732	0.622
22	Windows Application	0.694	0.788	0.727

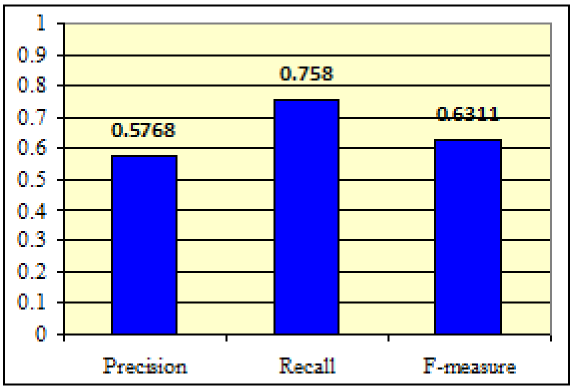


Figure 5: Experimental Results about Keyword-based IR Process

Then, performance comparison results between WQCA and keyword-based IR processes are shown in Figure 6.

According to the graph the precision value of WQCA-based IR system is above 20% higher than the precision of keyword-based IR. The average precision or the exactness of proposed IR is about 85%. The average recall value of proposed IR is above 95%. It means that almost all relevant documents are retrieved by WQCA-based IR system. F-measure comparison results of WQCA based and keyword based IR are clearly described with bar chart. F-measure is the combination of precision and recall of IR. These results prove that WQCA based IR is above 20% more efficient than the keyword based IR.

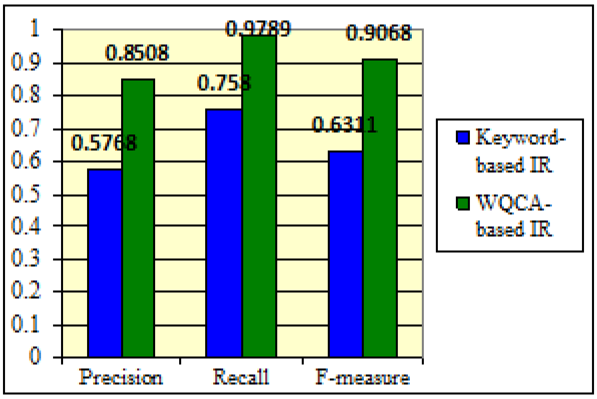


Figure 6: Performance Comparison Results

10. Conclusion

According to the experimental results, WQCA-based information retrieval system is more precise than the keyword-based IR system. So, WQCA-based IR system is a good information retrieval system by extracting more relevance user required documents. Moreover, this WQCA-based IR system can classify the web search taxonomies (query characteristics) and analyze the ambiguous domain terms. Web query classification method can solve the problems in lack of semantics correlativity in traditional retrieval system. The precision of the information retrieval system must increase by classifying the web query into the target categories. So, search engine can provide a set of relevant documents based on semantic retrieval.

References

[1] S. M. Fathalla and Y. F.14 Hassan, "A Hybrid Method for User Query Reformation and Classification", IEEE, pp. 132-138, 2012.

[2] A. Katariya, "Ontology-Based Web Query Classification", International Journal of Engineering Research and General Science, pp. 806-813, Volume 3, Issue 3, May-June, 2015.

[3] W. Yue, Z.Chen and X. Lu, "Using Query Expansion and Classification for Information Retrieval", Proceedings of the First International Conference on Semantics, Knowledge, and Grid (SKG), IEEE, 2006.

[4] B. Andrei, "A Taxonomy of Web Search", IBM Research, vol. 36, no. 2, 2002.

[5] A. Mohasseb, M. E. Sayed and K. Mahar, "Automatic Identification of Web Queries using Search Type Patterns", pp. 295-304, International Conference on Web Information Systems and Technologies, 2014.

[6] K. B. Shruti and D. D. Brian, “Leveraging Search Engine Results for Query Classification”, Department of Computer Science & Engineering, Lehigh University, 2013.

[7] A. Katariya, "Ontology-Based Web Query Classification", International Journal of Engineering Research and General Science, pp. 806-813, vol. 3, no. 3, May-June, 2015.

[8] S. M. Kim and T. Baldwin, “An Unsupervised Approach to Domain-Specific Term Extraction”, University of Melbourne, Australia, 2011.

[9] A. Nayak, A. Poriya and A. Poriya, “Type of NoSQL Databases and its Comparison with Relational Databases”, International Journal of Applied Information Systems (IJAIS), Foundation of Computer Science FCS, USA, Volume 5, March 2013.

[10] B. Liu, Web Data Mining, Department of Computer Science, University of Illinois at Chicago, USA, Springer-Verlag Berlin Heidelberg, 2007.

Author Profile

Naw Thiri Wai Khin received the B.C.Sc. and M.C.Sc. degrees from Computer University of Hinthada and Pathein in 2007 and 2010, respectively. During 2008-2012, she works as tutor in Application Department, Computer University of Pathein, Ministry of Science and Technology. During 2015-2018, she works as Assistant Lecturer in Application Department, Computer University of Hinthada. Now, she works as Lecturer in FITSM, University of Computer Studies, Yangon.