A Fuzzy-Ontology Based Information Retrieval System for Relevant Feedback

Comfort T. Akinribido[1], Babajide S. Afolabi[2], Bernard I. Akhigbe[3] and Ifiok J. Udo[4]

Information Storage and Retrieval Group (ISRG),
Department of Computer Science and Engineering,
Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria.

Abstract

Obtaining correct and relevant information at the right time to user's query is quite a difficult task. This becomes even complex, if the query terms have many meanings and occur in different varieties of domain. This paper presents a fuzzy-ontology based information retrieval system that determine the semantic equivalence between terms in a query and terms in a document by relating the synonyms of query terms with those of document terms. Hence, documents could be retrieved based on the meaning of query terms. The challenge has been that surface form does not sufficiently retrieve relevant document to user's query. However, the results presented showed that the Fuzzy-Ontology Information Retrieval system successfully retrieve relevant documents to user's query. This is irrespective of different meaning and varieties of domain. The System was tested on words with different meanings and some set of user's query from varied domains.

Keywords: *Information Retrieval, Synset, Probability Corpus Relevance, Term Frequency, Fuzzy techniques.*

1. Introduction

Fuzzy-Ontology Information retrieval system (FOIRS) is a system that typically measures the relevance of documents to users' query based on meaning of dominant words in each document. The weight of this dominant word in terms of both surface form (which is the matching of query terms with document terms) and domain concept should be measured according to their frequency and threshold values. Each document has a domain concept and target word that is elaborated and emphasized in the document. This word is mostly not repeated with the same spelling to prevent repetition or tautology, but it is written in different forms but with the same meaning. Hence, surface form can not only be used to determine the relevance of a document to user's query.

[12] states that context distance model compares the similarity of the contexts where a word appears, using the local document information and the global lexical co-occurrence information derived from the entire set of documents to be retrieved. The system must be able to adapt its behaviour autonomously to the changing context, expand the query to get synonyms of the query terms and originate the search task. Then, it can filter the results from irrelevant documents, organize them, and present them as useful information to the users in their current activities. [7] reported that specifying the context of a search can significantly improve search results. Thus, assess the context of any search is an important task not to be ignored. For this, they developed a probability dominant meaning space and context vector.

According to [5], FOIRS greatly improves retrieval effectiveness by expanding the query which can be a single word, keywords or longer phrase. The query terms can be expanded through a database that contains keywords and their synonyms. In the work term frequency and corpus relevance of words, which make up user's query were determined. This they did to tackle the challenge of word polysemy in document retrieval so that words can correlate better based on their meanings rather than on their surface forms.

Also, surface form representation of query terms does not sufficiently retrieve document that is relevant to the user's query. For instance, words like bank can occur in the context of river bank and financial bank; close (of door) and close meaning near; bat (the name of a bird) and bat used in sports (e.g. in baseball); wood (for firewood) and wood for the name of a person; caterpillar can mean a heavy equipment or a worm that would later develop into a butterfly. Words in these forms normally cause the retrieval of irrelevant documents as feedback to user's query, if surface form is applied.

In literature Boolean Information retrieval system (BIRS) allows the relevance of a document to be determined as relevant or non relevant; that is as (0 and 1). This in real sense does not specify any Grade of

relevance. Also the Vector space model (VSM) for Information retrieval (IR) computes the weight of query terms for each document. However, it does not relate the meaning of words in a document with meaning of words in user's query, as well as determine the corpus relevance of the query terms. This limitation is often the case with the use of VSM for IR [15]. [12] only proposed an approach that uses the local context vector analysis. This contains occurrences of query terms based on surface form. This limitation would have been well addressed through the introduction of fuzzy techniques as suggested in this study.

Also, while [9] used the meaning of queries (that is; ontology) to disambiguate queries, and did not use the context distance to determine the grade of relevance; [2] shared how useful the introduction of the use of ontology can be in IR, but did not develop any IR application using ontology concepts. Likewise, [6] only discussed the concept of contextual retrieval (CR). The claim made was that it combines both search technologies as well as the knowledge about query and user context. With CR is able to provide the most appropriate answer for a user's information need. The provided query if expanded using ontology without any further interaction from the user [1], would provide relevant feedback. Thus, the system would be a useful IR system that can be applicable in different areas like Intelligent Distance Learning Environment.

[8] In his work only discussed issues concerning satisfaction and frustration metrics. He reported that while the satisfaction metric takes into account only relevant documents, the frustration metrics concerns non-relevant documents. The main weakness of the IR models discussed so far is in the way they represent a document. That is as a 'bag of words'. However, search engines can only find words, which have been indexed. Therefore, developers must have it at the back of their mind that the author of a document may have used other words in the same context. These words must be synonyms.

[12] Reported that an increasing number of approaches to Information Retrieval have been proposed using models that are based on concepts rather than on keywords. But in this work, the concept of Ontology was used and thus defined as objects with two fields (Keywords and Synset). For each Keyword the corresponding synset was obtained and related to words in a document. The purpose was to search for new documents that semantically correlate to user's query. Thus, with a tolerance for imprecision and a positive use of fuzzy logic, the ranking of retrieved documents in order of relevance was enhanced.

Finally, relating the meaning of terms in user's query to document concept was sufficiently taken care of using the proposed methodology for this work. Consequently, the Fuzzy-Ontology concept used is discussed in section 2. While section 3 contains the proposed architecture and the algorithm for the system (FOIRS); both the implementation of the system and results are discussed in section 4 and 5 respectively.

2. Relating the Meaning of Terms in User's Query to Document Concept

Fuzzy-Ontology allows the easy determination of the precise meaning of a word as it relates to a document collection. [3] stated that Fuzzy-Ontology could be used in IR to locate precise information, which may be contained in a document content collection. Also, concepts represent a single sense, which is a set of synonyms called synset. Since a word is assumed to have a fixed number of senses as defined in the lexicon, such as WordNet (Thesaurus), the semantic similarity between the query terms is determined by incorporating a database that contains a dictionary of synonyms into the IR system.

2. 1 Representing Ontology Properties

The concept of Ontology was used to describe the meaning of query terms by getting the synonyms of all the keywords that make up the user's query. The set of synonyms of keywords in the query is called synset. Ontology was represented by objects stored in a database with two fields (keywords and synset, which is a set of synonyms for the keywords). For each keyword the corresponding synset was obtained and then related with words in a document. The rationale for this is that most writers prefer to change words in documents without omitting the main content. Instead they use words that have the same meaning as the main content. The synset and the query terms were therefore matched with the document terms to calculate the term frequency and corpus relevance.

2. 2 Term Frequency and Probability Corpus Relevance

Term frequency (TF) is the number of occurrences of the query terms in each document. It was improved by first getting the target word from query. Then the frequency of each word in a query, which appeared in the context of the target word, was divided by the frequency of the target word in each document. This was important since for instance, the document for Financial Bank would have a number of financial terms/issues than river terms (which could be assumed to be river bank). Thus, the expected document to be retrieved will not be for river bank. Current search engines do not have this technique, hence they retrieve both relevant and irrelevant documents provided they have same spelling. The emphasis therefore has been on the number of occurrences of query terms, which a

search engine matches the queries it receives against the index they create. The index consists of words in each documents, plus pointers to their locations with the documents [4], [5].

2. 2. 1 Probability of Corpus Relevance

A Corpus Relevance is how far a word is closely associated in the context of other word. For instance words like wheat, grains, cereal, and corn are mostly found in the same document. The Probability of Corpus Relevance of the target words and each word that make up the user's query was pre-computed as:

for i = 0 to p

$$R(W_q, W_{i+1}) = FID(W_q, W_{i+1}) / (FID(W_q) + FID(W_{i+1}) + FID(W_q, W_{i+1}))$$

next i.

As a result, the Probability Corpus Relevance =

$$\frac{1}{p}\left[\sum_{v=1}^{p} \frac{R\left(W_q, W_{i+1}\right)}{R_c}\right]$$

where;

FID = frequency in document
i = position of each word in the query
R_c = Maximum corpus relevance
W = number of corpus relevance
q = number where the target word belong in the query
P = total number of word in a query

It is interesting to note that some sample word pairs well with high corpus relevance scores, while others with low corpus relevance scores. Also, important is the need to get the Corpus Relevance of words that make up user's query. This was necessary, since if the query terms have high Corpus Relevance to a document, the document will be adjudged relevant to the query and vice versa. Similarly, the Corpus Relevance was also used to obtain the weight of query terms as well as the synset in the sample document.

2. 3 Ranking Using Fuzzy Concepts

Fuzzy techniques can be used to avoid rigid definitions and to manage uncertainty in hierarchical representations of concepts and in matching processes [11]. Therefore, fuzzy techniques were applied using term frequency and Corpus Relevance result to rank relevant document in order of relevance with specified threshold value. The technique was also used to rank every sample retrieved document in order of relevance. This was necessary, since if only relevant documents that satisfy the user's query are retrieved, users will be prevented from the burden of reading through many pages to get what they really needed. The easy applicability of the Fuzzy rule was possible, since the Probability Corpus Relevance had been achieved. Thus, if it is high then the relevance of the document to the query will be high and vice versa.

Consequently, User's preference or choice of (and access to) relevant document would be easy and precise through the use of fuzzy techniques for efficient ranking. For instance, a document that is 90% relevant will be retrieved before a document that is 80% relevant and so on. The overall implication of this that the time spent in trying to locate relevant document as mentioned earlier will be reduced.

The first step in applying the concept of Fuzzy was to determine the fuzzy set (Probability Corpus Relevance and Relevant). While the Probability Corpus Relevance was used as the fuzzy input variable, Relevant was used as the fuzzy output variable. See Table 1 and 2 below.

Table1: Membership Function of the Fuzzy-Ontology Information Retrieval System (Fuzzy Input Variable)

Fuzzy Input Variable	Membership Function
Probability Corpus Relevance	High
	Medium
	Low

Table 2: Membership Function of the Fuzzy-Ontology Information Retrieval System

Fuzzy Output Variable	Membership Function
Relevant	High
	Medium
	Low

In order to get the Grade of relevance of retrieved document, which is the strength (advantage) of this system (FOIRS), the following was adopted:

(i) First, if the degree of membership of one of the retrieved document is 0.7, then the document is highly relevant;
(ii) secondly, if it is 0.5, then the document is moderately relevant; and
(ii) thirdly, if the membership function is 0.1, then the document is not relevant.

Unlike the FOIRS, others like the BIRS will only categorize the document to be retrieved as (0.7, 0.5 and 0.1), which means (Relevant, Relevant, and not

Relevant) respectively. A second weakness with the system (BIRS), like others is that it scatters the result (retrieved documents) all over the result page. Thus making it very cumbersome for users to read through and fish out the most relevant feedback (document, which satisfy their information need).

3. The System Architecture

The diagram in figure 3.0 below is a pictorial representation of the FOIRS System Architecture.

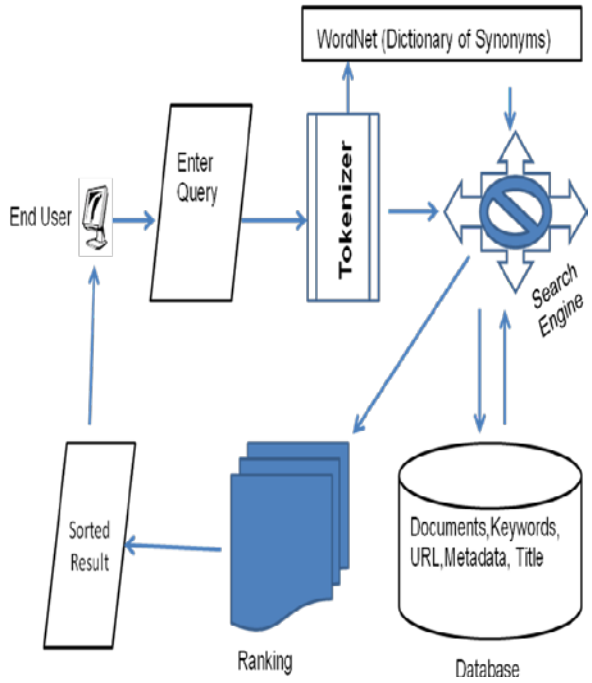

Fig. 3.0: The FOIRS System Architecture

From figure 3.0, when an end user types his/ her query through the text field (provided for user's query entry), the tokenizer divides the user's query into words. The crawler or search engine then locates the document's directories (URL), keywords, metadata, and the title of document. The query of an end user that is divided into tokens is matched with the corresponding document terms, keywords, metadata, and title in the database. The real documents whose keywords, metadata and title match the tokens of the end user's query are then retrieved to the document's directories.

A second responsibility of the search engine is to count the total number of terms in each document. Then it determines the frequency of occurrence of the tokens of the query in each document, which is the term frequency. The Synset of the tokens that make up the query was obtained using the Java WordNet library. Two parameters: The term frequency and the synset were used to calculate the Probability of corpus relevance. Also, using the fuzzy technique, the search engine pick the retrieved documents that are related to the query and ranked them based on a specified thresholds value. After the ranking the results are sorted into a list of relevant document that are arranged in a hierarchical order, and displayed in the user interface of the end-user.

3.1 Algorithm

Step 1: Get target word from textbox or Input box.
Step 2: Get the query from textbox or Input Box
Step 3: Break it into tokens or words
Step 4: Match each query term with metadata, title, keywords of document and each document term to determine their frequency.
Step 5: Find query terms' contextual meaning in Word Net Library by getting corresponding synset for each query terms in MySQL with database for Dictionary of Synonyms. Thus the number of synset of words in the query that appear in each document will be determined
Step 6: Determine Term Frequency for each query terms and their synset in the collection
Step 7: Find Probability Corpus Relevance.
Step 8: The URL of the relevant document is obtained and stored in database
Step 9: The probability of corpus relevance in each metadata, title, keywords and in the words that compose the document is obtained and used for ranking based on a threshold value.
Step10: Display relevant document in the list box according to their level of relevance in a hierarchical order.
Step 11: The title of the document is linked using hyperlink to the URL of the relevant document
Step 12: Finally, click the title and see the documents

4. Implementation

The database for the system (FOIRS) was implemented using MySQL. The database shown in figure 4.0 below contains dictionary of synonyms.

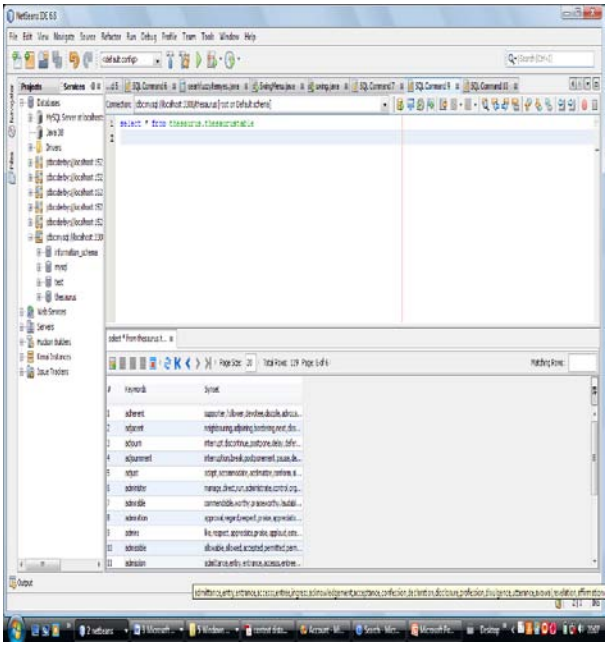

Fig. 4.0: The Database of FOIRS

Also, while Java programming language was used to develop and design the system as shown in figure 4.1 below; the Graphical User Interface was achieved using NetBeans 6.8.

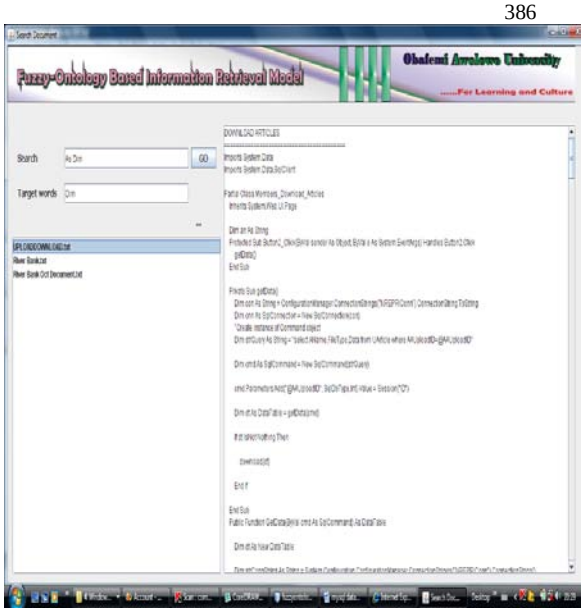

Fig. 4.1: The Graphical User Interface of FOIRS

4.1 System Evaluation

The system was evaluated using selected set of homonyms (words with the same spelling but different meanings). The FOIRS demonstrated the ability to strictly retrieve documents that are relevant to the specific meaning of the user's query and rank relevant document in order of relevance. Unlike FOIRS, current search engines retrieve both relevant and irrelevant document to the query of the user. This happens provided they have the same spelling with the query of the user. Two parameters: Satisfaction and frustration metrics were used for the evaluation of the system. The resultant feedback was compared with that of Google search engine. While the satisfaction metrics take into consideration only relevant documents, the frustration metrics considered non-relevant documents.

5. Result

The Table 3, presented below is used to indicate the results of some homonyms and samples of user's query tested on both FOIRS and GOOGLE.

Table 3: Results of Homonyms and User's query on both FOIRS and GOOGLE

Homo-nyms	Sample of User's Query	Frustration (Metrics)		Satisfaction (Metrics)	
F(FOIRS) and G(GOOGLE)		F	G	F	G
Bank	Financial Bank	1	5	11	7
	River Bank	0	4	7	3

Bat	Bat for baseball	0	6	10	4
	Vampire bat	0	3	8	5
Cater-pillar	Caterpillar sandal	0	10	15	5
	Caterpillar butterfly	0	9	17	8
Punch	Punch a paper	4	7	9	6
	Punch Newspaper	4	4	8	8
	Weed Management	5	6	7	6
	Organic crop production	1	4	7	4
Percentage in total document		**13.2 %**	**50.8 %**	**86.8 %**	**49.1 %**

As shown in table 3; some sample of user's query, such as weed management and organic crop production, which contain hynonyms but not stated under the homonyms column were entered for both FOIRS and GOOGLE. The purpose was to avoid preempting the system's ability to measure up with other systems, in terms of retrieval quality and strength. Thus the irrelevant documents retrieved are as indicated under frustration metrics, while the relevant documents are stated under the satisfaction metric. The percentages of both irrelevant and relevant documents are shown in the last row of table 3. Under the frustration and satisfaction columns; F and G is used to represent FOIRS and GOOGLE respectively. Thus, the result from the table and under satisfaction metrics indicate that the percentage of relevant documents retrieved with FOIRS is 86.8%, while that of GOOGLE is 49.1%.

This result is further buttressed using the graphs in figure 5.0 and 5.1 respectively below. Therefore, the graph in figure 5.0 below shows the relationship between the irrelevant documents and the relevant ones retrieved in GOOGLE.

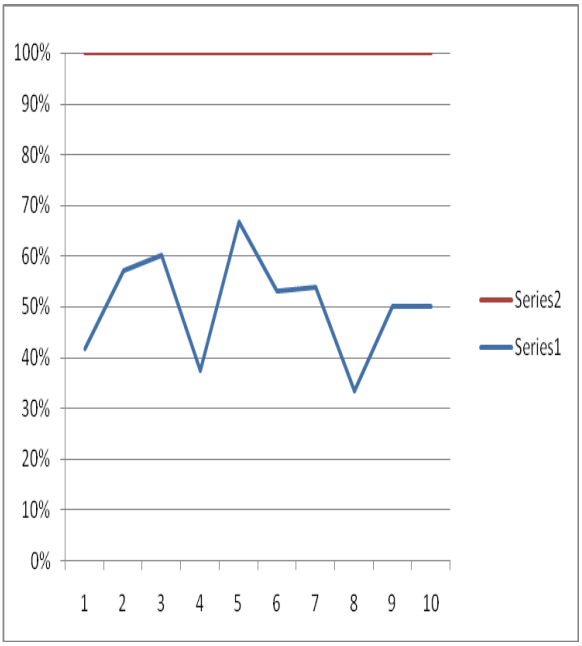

Fig. 5.0: The relationship of irrelevant and relevant documents retrieved in GOOGLE.

Also, the graph in figure 5.1 below shows the relationship between the irrelevant documents and the relevant ones retrieved in FOIRS.

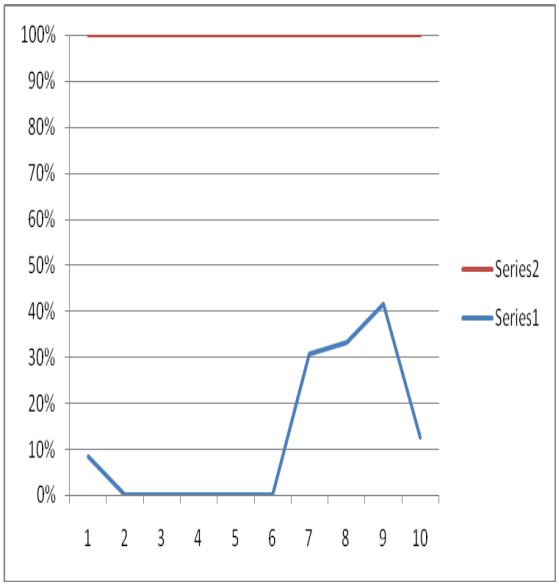

Fig. 5.1: The relationship of irrelevant and relevant documents retrieved in FOIRS.

In summary, both the relevant and irrelevant documents retrieved and as shown by the graph in figure 5.0 above is quite significant. That is, the ratio of 100% to almost 70% cannot be ignored. But from the graph in figure 5.1, the ratio of 100% and a little above 40% can be easily ignored. Thus, the ratio of irrelevant document retrieved is not significant as compared to the ratio of relevant document retrieved. This shows that there is significant improvement in the retrieval of relevant document to user's query, when FOIRS is used.

6. Conclusion

For a retrieved document to be really relevant most of the words added to the query must be related to the search context. Also, the retrieval quality of final results is likely to be high when context-based approach (CA) is applied to the design and implementation of IR systems. This could also bring some improvements in the retrieval of document with relevant feedback. CA was achieved through the introduction of fuzzy logic as proposed in this work. Thus, the use of fuzzy techniques in IR system according to the results reported, confirms that very good a result, more reasonable and satisfactory results in response to user's query will be ensured.

7. References

[1] Ciorascu, C., Ciorascu, I., & Stoffel, K.(2003) Knowler-Ontological Support for Information Retrieval Systems Proceedings of SIGIR 2003 Conference, Work-shop on Semantic Web, Toronto, Canada.

[2] Darijus Strasunskas and Stein L. Tomasscu, (2006); Department of Computer and Information Science, Norwegian University of science and Technol-ogy,NO_7491 Trondheim,Norway

[3] Dwi H. Widyantoro,(2001),'' A Fuzzy-Ontology Based Abstract Search Engine and its User Studies, Department of Computer science Texas A &M University College Station,TX 77843-3112,USA

[4] Liddy Elizabeth, (2005)," How a Search Engine Works", Director of Center for Natural Language Processing Professor, School of Information Studies, Syracuse University

[5] Salton G., Wong A. and Yang C.S (1975),'' A Vector Space Model for Automatic Indexing "Commun-ications of the ACM, vol.18, nr.11, pages 613-6230

[6] Allan James (editor) et al (2002),"Challenges in information Retrieval and Language Modeling" Report of a workshop held at the Center for Int-elligent Information Retrieval, University of Massachusetts Amherst

[7] Mohammed A. Razek, Claude Frasson, Marc Kaltenbach, (2003) " A Context-Based Information Agent for Supporting Intelligent Distance Learning Environments, Budapest, Hungary.

performance can be obtained if there is thorough understanding of: Term frequency and Inverse document frequency; Corpus Relevance; Use of WordNet for query expansion and the basic principles of information retrieval theory. However, the Unified Modeling Language (UML), if used in future for the design of the system could also contribute as a major factor to enhance the system. Other suitable Object oriented (OO) programming languages would also serve as a good alternative to the programming language used for the system design. As

[8] Korfhage, R.(1993). Information Storage and Retrieval Morgan kayfmann Publishers.

[9] Nagypal, G (2005), "Improving Information Retrieval Effectiveness by Using Domain Knowledge Stored in Ontologies," OTM Workshops 2005,LNCS 3762, Springer-Verlag, 780-789.

[10] Stefania Gallora(2007)" Fuzzy Ontology and Information Access on the web" Technical University of Kosice

[11] Tzoukermann Evelyne, Hongyan Jing, (2003)."Content Distance and Morphology Approach in Information Retrieval", Columbia University.

[12] Xu, J. and Croft W. B.(2008)" Improving the effectiveness of information retrieval with local context analysis" ACM Transactions on Information Systems (TOIS), Vol. 18, No.1.

[13] Yi- Chun Liao, (2007) "A weight –Based approach to information retrieval and relevance feedback" Hsuan Chuang University.

[14] Rubens N.O.,(2006) " The Application of Fuzzy Logic to the Construction of the ranking function of Information retrieval Systems; University of Massachusetts, Department of Computer Science. Computer Modeling and New Technologies,
Vol 10, No.1, 20-27

[15] Manning, C.D., Raghavan, P., Schütze, H. (2009). An Introduction to Information Retrieval. Cambridge University Press Cambridge, England. Retrieved from *http:/www*.nlp.stanford.edu/IR-book/pdf/00front.pdf

7. Biography of Authors

Akinbirido C.T. studied Computer Science at Adekunle Ajasin University, Akungba-Akoko in Ondo State. She obtained Second Class Upper Division. She is currently on her M.Sc degree programme in Computer Science and Engineering in Obafemi Awolowo University, Ile-Ife. Nigeria. Her areas of interest are Information Retrieval, Artificial Intelligence, Database Organization and Operation Research.

Afolabi, B.S. (Ph. D)

He is a Senior Lecturer in the Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife. Nigeria and head of the Information Storage and Retrieval Group research team.

Akhigbe, B.I

He is a member of Information Storage and Retrieval Group in the Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife. Nigeria. He has both B.Sc and M.Sc in Computer Science.

UDO Ifiok James

Information Storage and Retrieval Group (ISRG), Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife. Nigeria.

He obtained his B.Sc in Computer Science from University of Calabar, Calabar in 2005. Udo is currently on his M.Sc degree at Obafemi Awolowo University, Ile-Ife. His area of specialization is Information system design and Data reduction in Very Large Databases (VLDB).