**British Journal of Mathematics & Computer Science**
**18(4): 1-10, 2016, Article no.BJMCS.26935**
***ISSN: 2231-0851***

**SCIENCEDOMAIN** *international*
www.sciencedomain.org

# Conceptual Query Expansion Model for Web Information Retrieval

**Olufade F. W. Onifade[1] and Ayodeji O. J. Ibitoye[2*]**

*[1]Department of Computer Science, University of Ibadan, Nigeria.*
*[2]Department of Computer Science and Information Technology, Bowen University, Nigeria.*

***Authors' contributions***

*This work was carried out in collaboration between both authors. Author OFWO designed the study, wrote the protocol and supervised the work. Authors OFWO and AOJI carried out all laboratories work and performed the statistical analysis. Author AOJI managed the analyses of the study. Author AOJI wrote the first draft of the manuscript. Authors OFWO and AOJI managed the literature searches and edited the manuscript. Both authors read and approved the final manuscript.*

***Article Information***

DOI: 10.9734/BJMCS/2016/26935
*Editor(s):*
(1) Qiang Duan, Information Sciences and Technology Department, The Pennsylvania State University, USA.
(2) Kai-Long Hsiao, Taiwan Shoufu University, Taiwan.
*Reviewers:*
(1) Afeni Babajide Olakunle, Joseph Ayo Babalola University, Ikeji, Arakeji, Nigeria.
(2) Wei Gao, Hohai University, China.
Complete Peer review History: http://www.sciencedomain.org/review-history/16027

***Original Research Article***

***Received: 11th May 2016***
***Accepted: 23rd July 2016***
***Published: 2nd September 2016***

## Abstract

The process of retrieving relevant documents from user query is to begin with the clustering of documents with high semantic similarities between terms, and lower inner noise values. Here, the research extends normal keywords document clustering techniques in automatic thesaurus construction to building a Concept Based Thesaurus Network. The applied concept matching algorithm uses the Multi-Fuzzy Concept Network to generate sub clustered documents with relative degree of relationship across the clustered document. The proposed system achieved a higher cohesion rate between concepts and lower entropy rate in document. Also, a concise and relevant potential retrieved document were better ranked when compared with other existing document clustering techniques.

*Keywords: Concept based thesaurus network; concepts; document clustering; fuzzy concept network; document clustering; concept based document clustering.*

**Corresponding author: E-mail: ibitoye_ayodeji@yahoo.com;*

# 1 Introduction

A retrieved document is no doubt an extract from the clustered documents gathered from the user query in search time. Whenever users present an intention to the information retrieval system, the system identifies the terms that are present in this query to cluster potential retrievable documents [1]. Several existing Information Retrieval System (IRS) like the traditional keyword search, relevance feedback uses parameters which include but not limited to co-occurrence frequency of terms, degree of relationship, and word weight to determine the value of potential retrievable documents [2]. However, from the clustered documents used by these existing algorithms, many useful or relevant pages are not returned and many returned pages are useless or irrelevant. Since document clustering is a major stake that determines the relevance of retrieved documents from different user queries, to obtain an optimal result for the new trend of search that evolves round Concept Based Information Retrieval (CBIR) is to also cluster document based on concept from the user query [3,4]. There is no essence gained in trying to retrieved documents that is concept oriented on the surface of just any document clustering technique without the initial rudiment of concept cognitive approaches. The research established that only a Concept Based Document Clustering (CBDC) technique should be used for Concept Based Information Retrieval. This is because it engenders documents that are more relevant, precise and similar to users' search goal at different search time.

# 2 Related Works

Everyday documents are scaling at a very high level, and as huge amount of documents are generated it becomes difficult to search a particular or a group of document(s) due to data veracity. To this effect, document clustering as become a significant technique that aimed toward grouping similar documents in clusters. The essence is to make a system that clusters similar sort of documents efficiently. Different approach like the k means, Expectation Maximization and Hierarchical Clustering has been proposed by various researchers. For instance, the global K-means [5] clustering technique creates initial centers by recursively dividing data space into disjointed subspaces using the K-dimensional tree approach. The objective function of K means is to minimize the average squared distance of objects from their cluster centers, where a cluster center is defined as the mean or centroid μ of the objects in a cluster. However, the main limitation of K-means approach is that it generates empty clusters based on initial center vectors and there are no conceptual relationship between clusters. [6] proposed a modified version of the K means algorithm that effectively eradicates this empty cluster problem while a novel clustering algorithm which utilizes the swarm intelligence of ants in a decentralized environment that blended partitioned and hierarchical clustering was designed by [7]. This algorithm proved to be very effective as it performed clustering in a hierarchical manner without conceptual relationship between clusters. Moreover, [8] posited an approach for clustering heterogeneous data streams with uncertainty as [9] also proposed a novel Multi Representation Indexing Tree (MRIT) algorithm for constructing a hierarchy that satisfies arbitrary shape clusters with a good performance. Automatic and manually thesaurus construction system has also been identified as a good medium for retrieving relevant document in information retrieval. Several previous techniques for automatic thesaurus construction includes concept lattice-based information retrieval (IR) random indexing, and contextual document ranking modeled as basis vectors [10]. However, [11] addressed some problems of automatic thesaurus construction; this include the quality of automatically extracted semantic relations as compared with the semantic relations of a manually crafted thesaurus and a simple algorithm for representing both single word and multiword terms in the distributional space of syntactic contexts alongside a method for evaluation quality of the extracted relations was proposed. The experiments show significant difference between the automatically and manually constructed relations: while many of the automatically generated relations are relevant, just a small part of them could be found in the original thesaurus. Looking beyond the scope of automatic and manually thesaurus construction to gather relevant document, there are higher possibilities for different users to enter almost the same query with different aim toward achieving same or different goals. No doubt users present their request to search engine based on how they feel and what best term they consider can be combined together to retrieve their desired relevant document [12]. Most of the times, query input by users contain terms that do not match the terms used to index the majority of the relevant documents and sometime the un-retrieved relevant documents are indexed by a different set of terms than those in the query or in most of the other relevant documents [13]. One of the most difficult things is to guess accurately (100%) the intentions of a user by human not to talk of a system since most of the terms are ill-defined by searchers. Therefore, to retrieve documents based on concepts calls for a necessity to first of all identify concepts that are presented in the document before these concepts can be classified according to a conceptual structure. This system of concept based document clustering technique has the capacity to enhance the concept based information retrieval system with the ability to fetch documents even if they don't contain the specific terms in the user query while users can also retrieve more relevant document without having to define the rules for their queries.

# 3 Document Clustering

Document clustering (DC) is an important process because the retrieved documents are subsets of the set of the clustered document wherein the clustered document is also a subset of the entire search engine corpus. It is observed to be the first significant approach for all Information Retrieval (IR) techniques. Wherein, after a user had defined the problem on the text field and had clicked on the search button, terms from the query are extracted in order to gather relative documents that have these terms from the document corpus to form a document cluster for that particular user query. These documents are clustered based on the term frequency, word weight to form the initial clustering, intermediate clustering and final clustering as illustrated in Fig. 1.

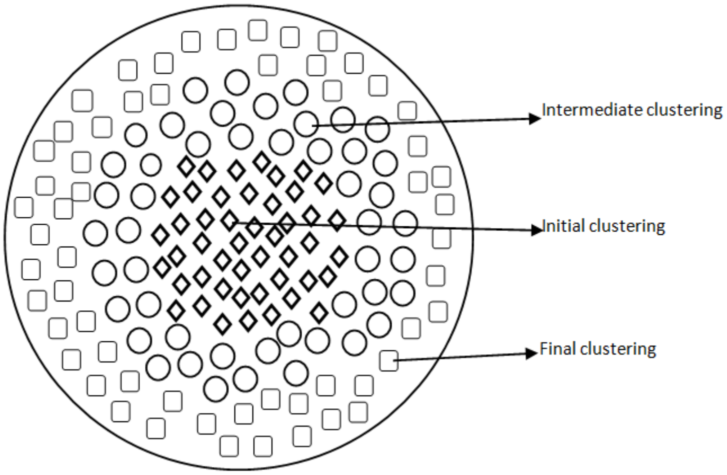

**Fig. 1. Typical view of a clustered document from user query**

In one of the approaches involved in information retrieval; such as automatic thesaurus query expansion, the final document clusters and/or intermediate document clusters are largely used to determine the type of retrieved document to a user query [14]. The outcome of neglecting the initial clustered document as part of potential retrieved documents from user query is a proof that, not the entire document in the cluster was considered in constructing a thesaurus. This also affects the degree of effect that existed between terms since the larger the document and terms in a document cluster, the lower the degree of effect between the terms. And once the degree of effect between terms is low, the degree of relationship between terms will also be low [15]. Hence, output from this clusters are also document with higher degree of inner noise, low precision rate and low recall rate.

Also, document clustering techniques involved in traditional keyword based search do not consider other terms that are significant in the document but the terms that was used in clustering such [16]. Hence, authors with this knowledge use the advantage to increase the ranking of their document by increasing the frequency of the terms. This approach will not in any way retrieve relevant document from the cluster because no term in a document is independent. This often time result in query drift since it has no clue of the field of search by the user. It important to emphasis that documents consist of several terms with associated links. This links defines the relationship that exists between terms that are presented in a document [17]. The link definition approach is to extract words from documents and then to compare documents with each other, documents with higher degree of matching will be clustered together. This approach can give us accuracy alongside the degree to which documents are similar to each other. Thus, user query retrieved document from document clusters that uses the keywords, will only retrieve document with the highest frequency of terms which may not be relevant or similar to user's intention. In order to achieve a more relevant retrieval with respect to users' query, Information Retrieval (IR) techniques must operate beyond the level of term based document clustering to a concept based document clustering approach wherein all nouns phrase, words are treated as concept to define a relationship in the document and between documents.

## 4 Proposed Concept Based Thesaurus Network

Concept Based Thesaurus Networks (CBTN) is an extension of the automatic thesaurus document clustering techniques; but here, documents are partitioned into sub clusters which showcased concepts that contains document with some level of degree of relationship. Once a user typed in the query ($\delta u_q$ ), the document classifier ($\partial_c$) identifies the terms in such query with the view to gather documents that contains it terms and respective synonyms. Here, equation 1 as indicated below is obtained

$$\delta u_q \overset{\partial_c}{\rightarrow} \nabla d \qquad (1)$$

Where $\boldsymbol{\delta u_q}$ is the user's query, $\partial_c$ is the document classifier based on term in the query, and $\nabla d$ is the clustered document that contains the term, synonyms and other related concepts.

After a successful document clustering, there is need to eliminate irrelevant document from the cluster in order to focus on relevant once. Hence, the fuzzy document classifier and thesaurus constructor tool, (FDCTC ($\boldsymbol{\varphi_f}$) ) is applied on the clustered documents to build the CBTN. It is from the application of FDCTC tool that we obtain equation 2 as illustrated

$$\nabla d \overset{\varphi_f}{\rightarrow} C_{\aleph} \qquad (2)$$

where, $\nabla d$ is the clustered document base on terms embedded in the user query, $\boldsymbol{\varphi_f}$ is the fuzzy document classifier and thesaurus constructor tool, $C_{\aleph}$ is the generated concept based thesaurus network.

The FDCTC ($\boldsymbol{\varphi_f}$) performs three leveled operations on clustered documents from users query. The essence is to cluster documents that are more relevant to user's intention and to establish conceptual relationship between the documents. Hence, FDCTC manipulates the user query by performing the following actions.

I. Extract text or keywords from each document in a cluster as distinct concepts
II. Establish a conceptual network structure of concepts through automatically thesaurus construction.
III. Measure the degree of relationship between concepts and respective documents by using the multi-fuzzy concept network.

Here, the main approach used in generating conceptual query expansion model for web information retrieval is defined by using graph of concepts that contains related documents for the network analysis. From the initial clustering of documents that contains user query terms, the function $G(d_{i=0}^{n})$ is applied to select document at random, eliminate non informative words and rank the documents based on term weight has contained in each of the document clustered. Thereafter, individual concepts (text) that are required in the construction of a CBTN are extracted and ranked using the function $f(c_{i=0}^{n})$. The ranking is based on concept weight, alongside co-occurrence frequency and associated relevance degrees. Then, an automatic thesaurus construction ($A^{tc}$ ) is used to build conceptual structure between concepts and documents. The automatic thesaurus $A^{tc}$ is further illustrated using equation 3.

$$A^{tc} = \frac{\Delta c_{kp}}{\Delta c_k} \times \frac{Num_d}{Num_c} \times \frac{Min(v_x, v_y) \times E}{Max(v_{x,}v_y) \times T} \times \frac{co\_occ}{Max(occ_x, occ_y)} \tag{3}$$

Where $Num_d$ is the number of document in the document cluster that contains terms such as $t_x$ and $t_y, Num_c$ is the number of documents in the document cluster, $v_x$ and $v_y$ are the degrees of effect of terms $t_x$ and $t_y$ in the document cluster center, T denotes the number of terms in the document cluster and E is the entropy, $occ_x$ denotes the number of documents containing term $t_x$ in document cluster; $occ_y$ denotes the number of documents containing term $t_y$ in document cluster C; $co\_occ$ denotes the number of documents containing both terms $t_x$ and $t_y$ in document cluster. Within the concept-document network, the methodology reflects the degree to which the document has the concept weight by using edges between the concepts, and the documents by using a valued degree to indicate it relationship strength. Concepts and documents present in the network are represented with nodes. However, unlike existing document clustering techniques, the CBTN makes use of the entire documents that are clustered with the presence of the user query term, its respective synonyms and noun phrases. More so, instead of neglecting associative relationship between terms, synonyms and other concept has been perpetrated by the reviewed existing document clustering approaches, the research methodology choose to restructure query based on concepts matching using multi-fuzzy concept network in conjunction with other parameters such as term frequency, degree of relationship, word weight etc. as stated above to build the CBTN. This process is further illustrated using Fig. 2.

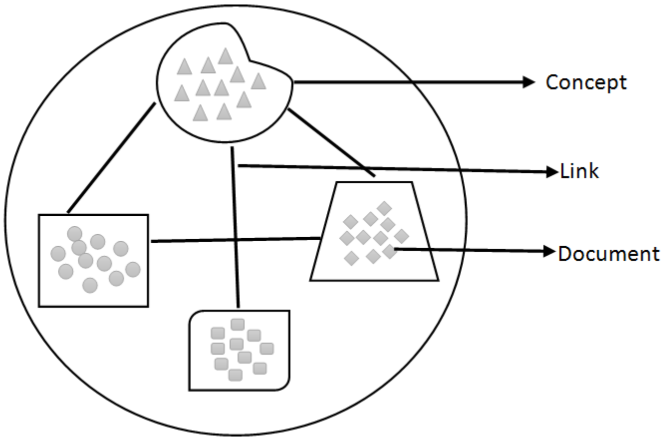

**Fig. 2. Proposed concept based thesaurus network**

The typical view of a concept based thesaurus network as shown in Fig. 2 indicates how concepts has, or contains respective documents and how these concepts are linked together with a degree of association. Fig. 3 is further used to illustrate Table 1 as an example to depict a possible level of association that could exist between documents and their respective concepts in CBTN.

**Table 1. A sample of identified concepts in a document**

| Concept node | Concept |
|---|---|
| C1 | Network |
| C2 | Artificial intelligence |
| C3 | Computer science |
| C4 | Security and encryption |
| C5 | Intranet |
| C6 | Internet |

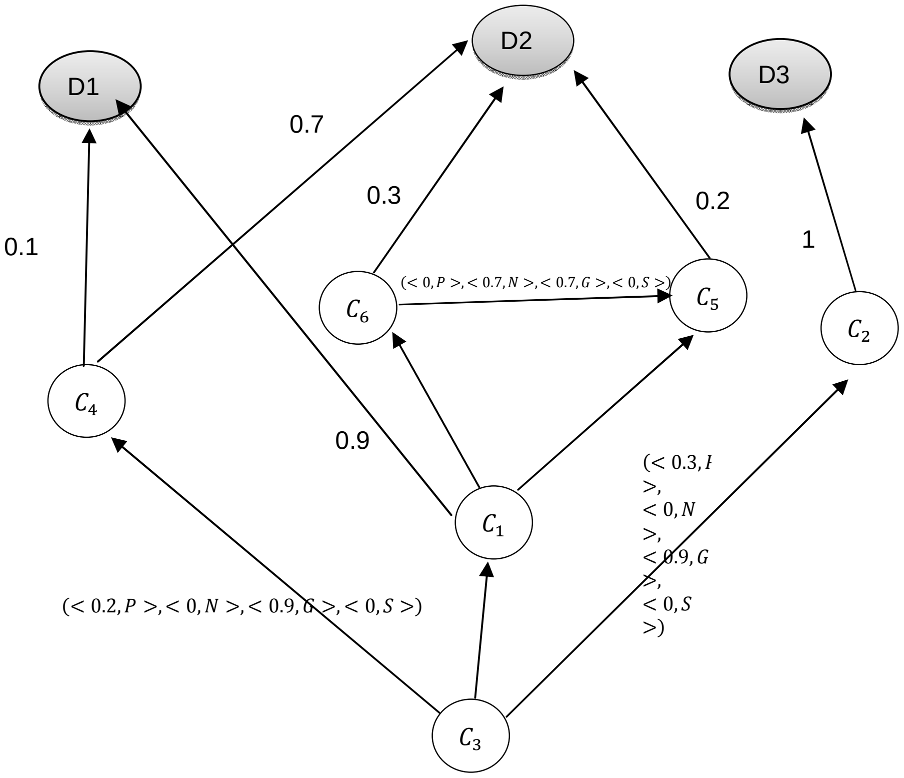
D1
D2
D3
0.7
0.3
0.2
0.1
1
$C_6$
$C_5$
$C_2$
$(< 0, P >, < 0.7, N >, < 0.7, G >, < 0, S >)$
$C_4$
0.9
$C_1$
(< 0.3, I >, < 0, N >, < 0.9, G >, < 0, S >)
$(< 0.2, P >, < 0, N >, < 0.9, G >, < 0, S >)$
$C_3$

**Fig. 3. Example of a concept based thesaurus network**

In Fig. 3, there is a fuzzy positive relationship of degree 0.7 and fuzzy general relation of degree 0.8 between concept three and four, fuzzy negative of degree 0.7, fuzzy general of degree 0.7, fuzzy special of degree 0.5 between concept six and five among others. The degree of relationship between the documents and concept for example $d_1$ has degree of 0.4 to concept $c_4$, 0.9 to concept $c_1$ and so on. This degrees indicate the strength at which documents has, or contains different concepts.

# 5 Experiments and Results

## 5.1 Entropy

Entropy was used as a measure of quality on the clusters with the caveat that the best entropy is obtained when each cluster contains exact relevant documents without or with minimal inner noise. For example, let *CS* be a clustering solution. For each cluster, the document distribution of the query is calculated first, i.e., for cluster *j* we compute $P_{ij}$, the “probability” that a member of cluster *j* belongs to document *i*. Then using this document distribution, the entropy of each cluster *j* is calculated using the standard formula

$$E_J = -\sum P_{ij} \log(P_{ij}) \tag{4}$$

where the sum is taken over all document clustered. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster:

$$E_{cs} = -\sum_{j=1}^{m} \frac{n_{j*E_j}}{n} \tag{5}$$

where *nj* is the size of cluster *j*, *m* is the number of clusters, and *n* is the total number of data points. Therefore, the Fig. 4 is a graph that demonstrates the entropy test conducted on two document clustering approach alongside the proposed concept based thesaurus network.

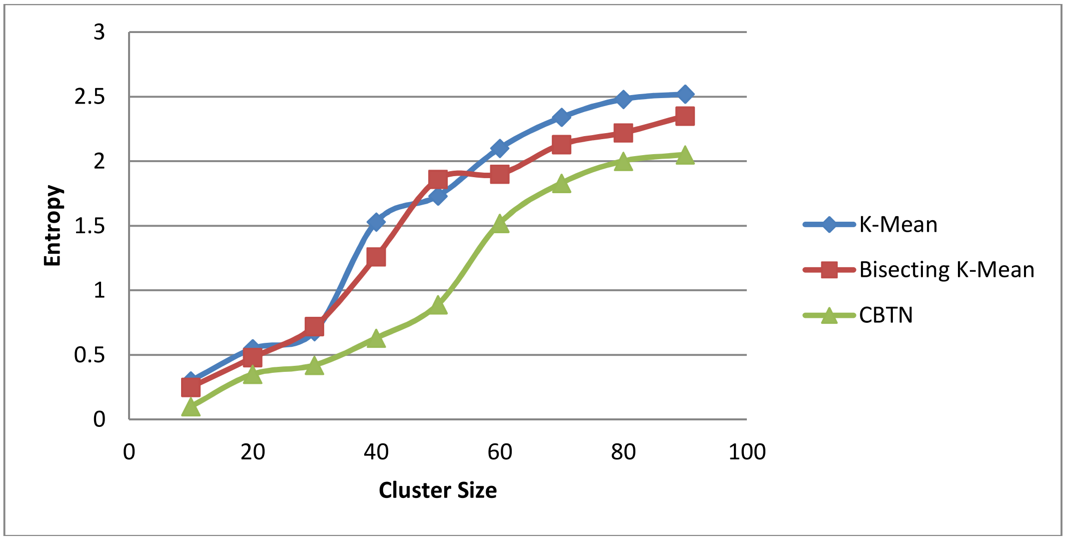

**Fig. 4. Entropy test on clustered document**

In Fig. 4, we can resolve that CBTN performs better in entropy test than the K-means and Bisecting K-means approach. This is because; there is minimal inner noise in respect to the numbers of documents present in a cluster as obtained from different user query. Although the three approach satisfies the Information Retrieval condition that the smaller the size of a document cluster, the minimal the entropy value. However, the values obtained for the rate of disorderliness from clustered document is more minimal in the proposed CBTN compared to the k-means and Bisecting K-means document clustering approach.

## 5.2 Degree of cohesion

This experiment is conducted to know the degree to which documents in a cluster can be related to each other. One major objective of the proposed CBTN is not just to cluster documents that have the presence of the user query term but to cluster documents that are similar in goal towards solving a problem. One common possible measure for computing the similarity between documents is the degree of cohesion test. This is defined as

$$c = \frac{1}{S}\sum_{d\in S} d \qquad (6)$$

Equation 3 is the vector obtained by averaging the weights of the various terms that are present in a document d from the set of documents *S*. In performing this experiment using the K-means, Bisecting K-means and the proposed CBTN, while using the cosine document similarity measure as indicated in equation 6, the Fig. 5 shows the obtained result.

In Fig. 5, a test for the level of similarities that existed between documents cluster from different user queries using the K-means, Bisecting K-means and the proposed CBTN document clustering technique was conducted. Although the three document clustering techniques result satisfies the information retrieval system condition that the smaller the number of documents in a cluster, the higher the cohesive force between the documents while the higher the number of documents in a cluster, the more prone to contain dissimilar documents. However, the values obtained from the proposed CBTN has indicated in Fig. 5

reflected that CBTN performs better than the existing document clustering approaches when considering similarities between documents in a cluster. After all, the CBTN is a re-cluster exercise on an already existing document, clustered from users query.

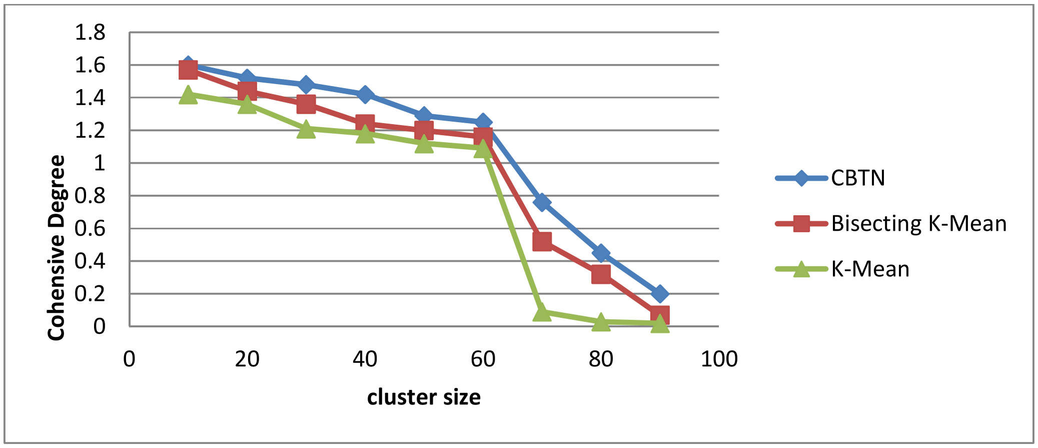

**Fig. 5. Degree of cohesion between documents**

Finally, in this section we present the precision, and recall values for the ten distinct queries. While precision rate is the ratio of the number of relevant documents to the total number of documents retrieved for a query, recall rate is the ratio of the number of relevant documents retrieved for a query to the total number of relevant documents in the entire collection. We illustrate the obtained result using Table 2.

**Table 2. Precision and recall values for ten user queries**

| User query | Precision value | | | Recall value | | |
|---|---|---|---|---|---|---|
| | K-means | Bisecting K means | CBTN | K means | Bisecting K means | CBTN |
| $Q_1$ | 45.43 | 47.54 | 50.24 | 60.24 | 62.46 | 66.17 |
| $Q_2$ | 43.05 | 52.17 | 58.62 | 60.78 | 64.12 | 69.56 |
| $Q_3$ | 46.68 | 50.45 | 58.72 | 63.98 | 65.45 | 70.12 |
| $Q_4$ | 43.42 | 49.33 | 52.14 | 60.49 | 63.1 | 68.21 |
| $Q_5$ | 45.42 | 50.34 | 54.68 | 61.62 | 63.82 | 69.66 |
| $Q_6$ | 42.83 | 52.45 | 55.64 | 60.24 | 62.42 | 67.15 |
| $Q_7$ | 44.29 | 50.81 | 56.25 | 62.47 | 69.43 | 74.82 |
| $Q_8$ | 44.01 | 52.45 | 59.23 | 61.44 | 67.25 | 71.42 |
| $Q_9$ | 45.56 | 52.67 | 56.46 | 62.74 | 66.33 | 71.63 |
| $Q_{10}$ | 46.44 | 53.6 | 55.76 | 60.23 | 64.78 | 68.83 |

From Table 2, we observe that the proposed CBTN outperforms the K-means and the Bisecting K-means approach for both precision and recall value analysis per user query. This approach in no doubt has helped to enhance the retrieval of relevant document from the data corpus based on the users query. The research was able to establish the fact that a concept based document clustering which uses the user query to establish a degree of conceptual similarity between the potential retrievable documents is the basis for retrieving a more relevant document for the user's goal.

## 6 Conclusion

Document clustering will forever remain an important factor that will always decide the document that will be retrieved as results to different user queries. Hence, to retrieve documents that is concept based; a concept based document clustering will always be required. CBTN has helped to showcase that the degree of effect between terms can be increased for the entire document cluster while the degree of relationship between terms can also increase since each concept in a multi-fuzzy network contains several documents with some degree of relationship as compared to other existing techniques. The relationship that existed or that may exist between concepts creates a bond that helps to identify the user query intentions. This also helped CBTN to eradicate unwanted or irrelevant documents that may be retrieved in connection to user query while document which do not contain all or any of the users query but relevant to users intention can be retrieved.

## Competing Interests

Authors have declared that no competing interests exist.

## References

[1] Custis T, Al-Kofahi K. A new approach for evaluating query expansion: query document term mismatch. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, Amsterdam, The Netherlands. 2007;575–582.

[2] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. Introduction to information retrieval. Cambridge University; 2008.

[3] Klink S. Query reformulation with collaborative concept-based expansion. In Proceedings of the First International Workshop on Web Document Analysis WDA; 2001.

[4] Choi J. Choi, M. Kim VV. Raghavan. Conceptual retrieval based on feature clustering of documents; 2002.

[5] Likas A, Vlassis N, Verbeek JJ. The global k-means clustering algorithm. Pattern Recognition. 2003;36(2):451-461.

[6] Malay K. Pakhira. A modified k-means algorithm to avoid empty. International Journal of Recent Trends in Engineering. 2009;1(1):220-226.

[7] Alam S, Dobbie G, Riddle P, Naeem MA. Particle swarm optimization based hierarchical agglomerative clustering. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WIIAT). 2010;2:64-68.

[8] Guo-Yan Huang, Da-Peng Liang, Chang-Zhen Hu, Jia-Dong Ren. An algorithm for clustering heterogeneous data streams with uncertainty. 2010 International Conference on Machine Learning and Cybernetics (ICMLC). 2010;4:2059-2064.

[9] Lifeng Wang, Hui Song, Xiaoqiang Liu, ―Incremental Document Clustering Using Multi representation Indexing Tree; 2010.

[10] Chen LY, Chen SM. A new method for automatic thesaurus construction and query expansion proceedings of the 2004 15th International Conference on Information Management, Taipei, Taiwan, Republic of China; 2004.

[11] Panchenko A. Technology of the automated thesaurus construction for Information Retrieval. Intelligence Systems and Technologies, Bauman Moscow State Technical University, Moscow. 2009;9:124–140.

[12] Tanveer Siddiqui, Tiwari US. Natural language processing and information retrieval‖. Oxford University Press; 2008.

[13] He B, Ounis I. Studying query expansion effectiveness. In Proceedings of the 31st European Conference on Information Retrieval (ECIR 2009). Springer, Toulouse, France. 2009;611–619.

[14] Hartmann RK, Brown K. Thesauruses. Encyclopedia of language & linguistics. 2nd ed. Elsevier. Oxford, pp. 668-676. Holscher C, Strube G (2000). Web search behavior of Internet experts and newbies. Computer Networks. 2005;33:337-346.

[15] Kalczynski PJ, Chou A. Temporal document retrieval model for business news archives. Information Processing and Management. 2005;41(3):635-650.

[16] John W. Kang, Hyun-Kyu. Kang A. Term cluster query expansion model based on classification information in natural language information retrieval. International Conference on Artificial Intelligence and Computational Intelligence; 2010.

[17] Wei Song, Cheng Hua Li, Soon Cheol Park. Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity. Journal of Expert Systems with Applications. 2009;36(5):9095-9104.

***Peer-review history:***
*The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)*
*http://sciencedomain.org/review-history/16027*