ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 2, No.14 , March 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Dynamic Multi-Fuzzy Semantic Concept Based Document Clustering Technique

Ayodeji, O. J. Ibitoye
Department of Computer Science and Information Technology, Bowen University, Iwo, Nigeria
ibitoye_ayodeji@yahoo.com

Abstract

The fast pace of change in web content and the consistency in the exponential growth of web data, calls for a matching approach that can proportionately increase the relevance ratio of retrieved documents from potential clustered documents gathered by the user query. The research introduces the Dynamic Multi-Fuzzy Semantic (DMFS) technique as an enhanced methodology that seeks for the retrieval of concept oriented document by establishing a conceptual structure on the document clustered with the goal of obtaining a well partitioned sub cluster of documents with associated degree of relationship that is structured to retrieving a concise, specific and pertinent document that is based on concept matching and more preferable to the users' goal when compared with existing document clustering approach

Keywords: *Fuzzy, Automatic Thesaurus Construction, Fuzzy Concept Network, Concept, Concept Based Thesaurus Network, Information Retrieval.*

1. Introduction

Since the Web contains material of quite different varieties, users cannot anticipate what is available, sometimes users also finds it difficult to express their interest and later, they are surprised at the obtained result due to word ambiguities [1]. The reason may be because most times, query inputs by users contain terms that do not match the keywords used to index the majority of the relevant documents and sometime the un-retrieved relevant documents are indexed by a different set of keywords other than those in the query [2]. In the course of trying to retrieve relevant document irrespective of the indexing problems, several approaches such as word clustering, relevance feedback, query expansion and more with relative tools have been implemented. However, many results from Information Retrieval (IR) still curtail the possibilities of retrieving irrelevant document [3] either through expanded query as performed by thesaurus or other embraced method known to it developers if the document clustered used in the retrieval excise is not conceptually driven. Therefore, this research supports the fact that in order to retrieve a more relevant result with respect to user's intention, the mode of document clustering needs to be streamlined to the user query semantics. Hence, information retrieval systems must operate beyond the level of just near synonyms, and thesaurus construction towards outlining the semantic similarities between concepts in a clustered document. While we do not underestimate the importance of some existing techniques, we also understand that for IR to retrieve a better and relevant result, a better tool that is concept driven from the basis of document clustering to result display is necessary to achieve the goal of information retrieval system.

2. Overview of Concept Base Information Retrieval

The meaning of a text (word) depends on conceptual relationships to objects in the world rather than to linguistic or contextual relations found in texts or dictionaries. This assertion is a new trend of information retrieval model that is drawn from the intellectual survey; based on inference [4]. Informally, concept-based information retrieval can be defined as a search for information objects based on their meaning rather than on the presence of the keywords in the object [4]. However, when sets of words, titles, noun-phrases, and keywords are mapped to encoded concepts, the process is called concept based information retrieval. Here, the terms in the user query have representing domain concepts, and not as literal strings of letters. It can fetch documents even if they don't contain the specific words in the user query. This model pays much attention to representing conceptual information without acquisition of that knowledge [5]. Definitely, users would prefer to retrieve documents of interest without having to define the rules for their queries.

3. Basis for Concept Oriented Document Clustering Technique

In order to retrieve document based on concepts, it becomes a necessity to first identify the terms in the

user query [6] to cluster the require documents. Then, concepts inside the document can be classified according to the given conceptual structure. Concepts here are abstract or generic idea that is generalized from particular instances. As these concepts are expressed by natural language, it is possible to identify the concept in a document by analyzing it phrases. Hence, Natural Language Processing (NLP) [7] is an effective tool that can be used to analyze syntax and semantics of the text for categorization. Though there are other several ways of identifying concepts which are contained in a document, fuzzy reasoning about the terms found in such document are also used for calculating the likelihood of a concept present in the document [8]. After the successful identification of concepts, a need to identify the conceptual structure wherein the concepts belong is of necessity. One of the conceptual structure type is the conceptual taxonomy. Conceptual taxonomy is a hierarchical organization of concept descriptions according to generalization relationship [9]. Here, each concept in taxonomy has link to its most specific list ("Animals" or super concepts) and links to its most general subsumes ("rabbit", "goat", "cat" or subconcepts) in taxonomy. While, formal ontology is another good conceptual representation for entities, events, and other relationships that is composed of a specific domain, two primary relationships are abstraction (subsumption) and composition ("part-of" relationship). Various conceptual structures like dictionary, thesaurus, ontology or automatically generated one can be used. It is reported in many papers that pre-existing dictionaries often do not meet the user's needs for interesting concepts, or ontology like WordNet does not include proper nouns. Usually, conceptual taxonomies are constructed manually by deciding where in the taxonomy each concept should be located or automatically using special conceptual indexing technique. The proposed technique in section (4) executes the above principle with an extension by introducing the multi-fuzzy concept network [10, 11, 12] on the clustered document. The multi-fuzzy concept network gives the clustered document ability to reveal the conceptual knowledge of a document collection [13, 14]. The semantics of the possible fuzzy concept network between concepts reviewed here are the fuzzy positive, negative, and generalization and specialization association This helps to relate concepts with a fuzzy similar meaning in some contexts.

4. The Proposed Methodology

A required intention of this research is to extend the normal keyword information retrieval document clustering technique with the view to maximize the volume of document clustered, apply a conceptual taxonomy on the clustered document to engender sub clusters before establishing a conceptual structure between document in a cluster and terms in the document using the multi fuzzy concept network to describe all possible context-independent relationships and context dependent relationships between keywords concepts. Hence, the dynamic multi fuzzy semantic concept based document clustering technique is packaged to perform these three distinct operations on clustered documents gathered from user's query has illustrated in fig 1 below. The effect of this is to enable pages which would not be included in the result set, but with potential to be more relevant to the user's desired intention to be included.

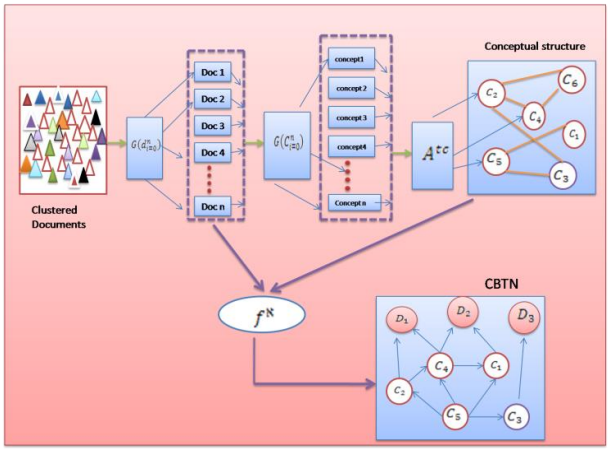

Fig 1: Diagram illustrating the dynamic multi-fuzzy model

From the fig 1, the initial clustered documents contain respective documents that have the presence of at least a user query terms. This is further illustrated using the Eq (1) below

$$u_q \overset{\beta_c}{\rightarrow} \gamma d \qquad (1)$$

where $\boldsymbol{u_q}$ is the user's query, β_c is the keyword extractor from query, and γd is the clustered document that contains the term, synonyms and other related concepts from the user query.

The documents in the cluster are randomly selected using the function $G(d_{i=0}^{n})$with the view to eliminate non informative words and extract concepts (text) that are necessitated in the building of conceptual relationships between terms in a document. Thereafter, the concepts are later ranked based on weight, frequency and co-occurrence degree of relationship that may exist using the function $G(c_{i=0}^{n})$. Then a conceptual structure is established through an enhanced automatic thesaurus construction approach before the application of multi-fuzzy concept network is used on this conceptual structure to identify the degree of relationship between concepts and respective documents. Hence, we further illustrate this technique by using Eq(2) and Eq(3) below.

$$\boldsymbol{\nabla_{\rho}} = \mu_{\rho} * \alpha^{tc} \qquad (2)$$

Where μ_{ρ} is the text finder, α^{tc} is automatic thesaurus constructor and $\boldsymbol{\nabla_{\rho}}$ is the conceptual structure.

$$\boldsymbol{\varphi}_{\cap} = \frac{\Delta f_{\tau}}{\boldsymbol{\nabla_{\rho}}} \qquad (3)$$

where Δf_{τ} is the fuzzy concept network $\boldsymbol{\varphi}_{\cap}$ is the redefine document in selected groups of different concepts. The essence is to identify terms that can be linked with one another with some level of association. Each concept cluster contains documents that have some degree of relationship while individual document contains concepts with some degree of association also. Therefore, instead of reducing the entire clustered document or using the entire clustered document without retrieving relevant document to user query, the proposed technique helps to develop an exclusive association between the user query and the potential retrievable documents with the view to generate a conceptual sub clusters of documents wherein the obtainable result can be channeled directed on the sub cluster with a highest degree of fuzzy relationship value to the user query. We therefore use fig 2 and 3 below to further juxtapose our intentions.

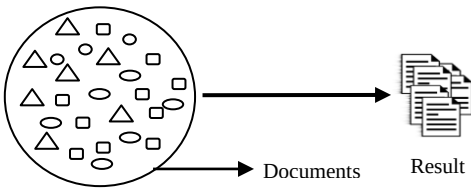

Fig 2: Ordinary keywords clustered document to retrieved result

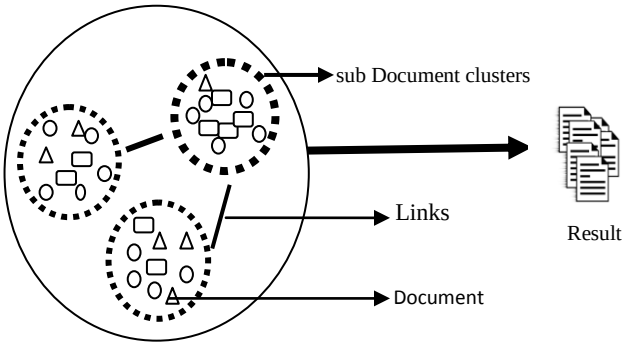

Fig 3: dynamic multi-fuzzy semantic concept based document clustering to retrieved result

The proposed technique in fig 3. has been able to transient form the representation in fig 2 by establishing conceptual structure between documents and a degree of relationship between concepts and sub concept clusters. Hence, in fig 3, the user query is as important as the clustered document where the clustered documents are also as important as the user query. Therefore, in comparing the precision rate, recall value and the degree of similarity (cohesion) that existed between documents using a set of 20 user query on a dataset of about 2500 on various fields of computer science subjects to compare the ordinary keyword clustering document technique with the proposed dynamic multi fuzzy concept based document clustering techniques, we obtained the following results has presented in table 1. below

Table 1: Summarized result on conducted experiments.

Methodology	Precision Rate (average %)	Recall Rate (average %)	Cohesion (average %)
keyword clustering	54	40	10
DMFS	73	64	68

From table 1, the dynamic multi-fuzzy semantic concept based document technique demonstrated an average of 19%, 24% and 58% performance rate in precision, recall and test for cohesion when compared with the ordinary keyword approach. This is because our research focuses more on the fact that the best form of information retrieval system is the one that can easily identify the concept in the user query,

establish an association between the query and potential retrievable documents, identify relationship between documents that will produce documents in the sub cluster that will portray to a large extend the intention of the searcher.

5. Conclusion

In this research, we started by creating a platform just like other information retrieval systems wherein the entire documents that contains the users query is clustered together. From this clustered document, we performed three distinct operations. These includes:

I. Identifying text, keywords from each document as distinct concepts
II. Building a conceptual structure through automatically thesaurus construction by using the generated concept in (I) above
III. Using the multi-fuzzy concept network to identify degree of relationship between concepts and respective documents

The research was able to establish the fact that a concept based document clustering with a degree of similarity between the user query and the potential retrieval documents is the basis for retrieving a more relevant document for the user's goal.

REFERENCES

[1] 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.

[2] Y.H Jung Park,. and D Du,. An Effective Term Weighting Scheme for Information Retrieval, Computer Science Technical Report TR008, Department of Computer Science, University of Minnesota, Minneapolis, Minnesota, pp. 1-15, 2000.

[3] M. Speretta, S. Gauch, Personalized search based on user search hierarchies, Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05), 2005

[4] J Choi,. M Choi,. V.V Kim,. and Raghavan , Conceptual Retrieval based on Feature Clustering of Documents, 2002.

[5] M. Dragoni, Celia da Costa Pereira, G. B Andrea. Tettamanzi, A Conceptual Representation of Documents and Queries for Information Retrieval System using Light Ontologies, Expert Systems with Applications 39 (2012) pp. 10376–10388, Elsevier, 2012.

[6] B.M Kim, J. Y Kim, and J Kim, Query term expansion and reweighting using term co-occurrence similarity and fuzzy inference, Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference, Vancouver, Canada, Vol. 2, pp. 715-720, 2001.

[7] Tanveer Siddiqui and U. S. Tiwari "Natural Language Processing and Information Retrieval" Oxford University press, 2008

[8] Y. J Horng,., S. M Chen,. and Lee, C. H., Fuzzy information retrieval using fuzzy hierarchical clustering and fuzzy inference techniques, Proceedings of the 13th International Conference on Information Management, Taipei, Taiwan, Republic of China, pp.215-222, 2002.

[9] S Loh, L K Wives and J P M de Oliveira, Concept-Based Knowledge Discovery in Texts Extracted from the Web, SIGKDD Explorations, ACM SIGKDD, July 2000, Vol 1, Issue 1, 29-39

[10] S.M Chen,. Y.J Horng, and C.H Lee. Fuzzy information retrieval based on multi-relationship fuzzy concept networks, Elsevier, 2002.

[11] S.J Chen and H.C Chu., A New Method for Fuzzy Query Processing of Document Retrieval based on Extended Fuzzy Concept Networks, International Conference on Electronics and Information Engineering, 2010.

[12] Y. C Chang, S. M Chen, and C. J Liau, A new query expansion method based on fuzzy rules, Proceedings of the 2003 Joint Conference on AI, Fuzzy System, and Grey System, Taipei, Taiwan, Republic of China, 2003

[13] Y. J Horng, S. M Chen. and C. H Lee., A new fuzzy information retrieval method based on document terms reweighting techniques, International Journal of Information and Management Sciences, Vol.14, No.4, pp.63-82, 2003.

[14] L. Y Chen and S. M Chen,, A new fuzzy hierarchical clustering method based on dynamic cluster centers, in Proceedings of the Ninth Conference on Information Management Research, Changhua, Taiwan, Republic of China, 2003.