International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:5, 2014

Fuzzy C-Means Clustering for Biomedical Documents Using Ontology Based Indexing and Semantic Annotation

S. Logeswari, K. Premalatha

***Abstract*—**Search is the most obvious application of information retrieval. The variety of widely obtainable biomedical data is enormous and is expanding fast. This expansion makes the existing techniques are not enough to extract the most interesting patterns from the collection as per the user requirement. Recent researches are concentrating more on semantic based searching than the traditional term based searches. Algorithms for semantic searches are implemented based on the relations exist between the words of the documents. Ontologies are used as domain knowledge for identifying the semantic relations as well as to structure the data for effective information retrieval. Annotation of data with concepts of ontology is one of the wide-ranging practices for clustering the documents. In this paper, indexing based on concept and annotation are proposed for clustering the biomedical documents. Fuzzy c-means (FCM) clustering algorithm is used to cluster the documents. The performances of the proposed methods are analyzed with traditional term based clustering for PubMed articles in five different diseases communities. The experimental results show that the proposed methods outperform the term based fuzzy clustering.

***Keywords*—**MeSH Ontology, Concept Indexing, Annotation, semantic relations, Fuzzy c-means.

I. Introduction

ANNOTATION is an elementary bustle in biomedical research. It acquaintances a commentary or approved judgment (textual comment, citation, classification, or other related object) to a target of annotation, such as a text or image. It can be produced for individual use, as in note-taking and personal classification of documents. Biocuration, or biomedical database resource annotation, is extremely constructive and ubiquitous in biomedical research. Nevertheless, most widely accessible biomedical data are unstructured and seldom described with ontology concepts available in the domains. This affluence of openly accessible biomedical data is beginning to facilitate cross-cutting integrative translational bioinformatics studies. The existing translational discovery methods for biomedical resources are hampered because they lack in standard terminologies and ontologies to describe their elements. The dispute is to generate consistent terminology labels for each component in the public resources that would permit the recognition of all elements that relate to the same type at a given level of granularity. One of the mechanisms for achieving this translation is to map the text meta data to the ontology concepts to devise refined or coarse search criteria.

Habitually annotations can be generated for the Meta data in three ways, manually, automatically and semi-automatically [1], [2]. In manual annotation, links between data and concepts are specified by human realm experts. In automated annotation, specialized programs are parsing data for providing such links. In semi-automated annotation, specialized programs are signifying links between data and concepts, which are consequently validated by domain experts. Annotation becomes an appealing strategy to reduce the complexity in retrieving important documents from the Web [3]. Annotations also facilitate an optimized indexation of documents along with the performance improvement.

Existing search engines, such as *Google* and *Yahoo*, offer a keyword-based search, which is based on mainly the surface string similarity between query and document terms [4]. In addition to that, sometimes the search is extended as a simple synonym expansion, excluding other types of information about terms, such as polysemy and homonymy. In order to deal with this problem and to get better search results, the usage of ontologies is recommended for document annotation. The practice of ontologies provides a content-based access to the data, which makes it possible to process information at the semantic level. This content-based processing extensively improves the search efficiency of relevant documents. However, due to the huge amount of existing documents, the idea of generating document annotations is not a trivial task. In this paper, an effective method called, concept based indexing and a method for enriching documents with semantic annotations are proposed for clustering where document terms are indexed and annotated according to domain ontology.

The rest of the paper is organised as follows. Section II involves with discussion of related works. Section III provides the insight of the proposed semantic annotation using MeSH ontology. Section IV presents the fuzzy c-means clustering process of biomedical text documents using ontology based semantic annotation. The experimental setup and results drawn from the proposed work and comparison with other semantic methods are discussed in Section V. Finally, Section VI gives the conclusion of the paper.

II. Review of Related Works

A MeSH ontology based annotation method is proposed for biomedical literatures by [4]. Annotation is implemented using Maximum Entropy (MaxEnt) classifiers. A MaxEnt model is trained for each of the terms and it can be applied to any

S. Logeswari is with the Bannari Amman Institute of Technology, Erode, Tamil Nadu, India (e-mail: slogesh76@gmail.com).

document for deciding whether it should be annotated with respect to the term or not. Experiments were conducted on PubMed documents with the feature types title, abstract, year and journal. The performance is validated with the measures such as average precision, average recall and average F-Measure. The results showed that the proposed method was able to produce high accuracy annotations.

An ontology based approach for automatic annotation is presented for document segments by [5]. Annotation is performed using their headings in conjunction with ontology and an annotation algorithm. This method is implemented on only specific domain rather than the general purpose ontology. The context information of the entities is used in the annotation process along with their relationships to the other neighboring entities. The approach proposed in this work significantly improved the search efficiency without any additional complexity. It is used to build an agriculture search engine.

A meta-data model for semantic annotation is developed for an effective information extraction system by [6]. The annotation is implemented by the mapping of instances to the classes in the ontology. Semantic based indexing and retrieval is performed based on this annotation by combining traditional information retrieval queries and ontology-based ones. In this proposed model, various issues related to the representation and usage of annotation is addressed. Based on MeSH ontology, an annotation method is presented for determining set of papers related to the topic, such as a gene, an author or a disease [7]. In this proposed method, MeSH Over-representation Profiles (MeSHOPs) are generated for the quantitative summarization of annotations in a convenient form for further computational analysis and visualization. The degree of association between any entity and the annotated medical concepts is measured based directly on relevant primary literature. The performance is analyzed by the comparisons with the hierarchical clustering techniques. The trustworthiness of MeSHOP annotations is evaluated based on the capacity to re-derive the subset of the Gene Ontology annotations with equivalent MeSH terms.

A Multiple-Ontology Knowledge Representation (SMOKR) system is presented to alleviate the problems associated with the query representation methods due to the lack of cross-ontology integration and semantic relations by [8]. The phrases are annotated and the semantic relations between them are identified using different domain ontologies before the instantiation of ontologies with the annotated phrases. Annotation is integrated with the ontology by mapping their instances using simple natural language processing techniques and also by matching their concepts using the state-of-the-art Biomedical Ontology Alignment Tool (BOAT). The performance of this SMOKR is assessed by testing it with a set of semantic queries and the results are compared with the keyword-based search engine Lucene.

A keyword based annotation model is proposed for biomedical documents based on the MeSH ontology by [9]. In this proposed annotation, semantic relations are extracted among the terms in the documents and are used to analyze the documents. Semantic relations are identified based on the graph of local co-occurrence relations among the terms and by a new global similarity metric which is computed using the connectivity of a graph.

A comparative study on fuzzy c-means and entropy based fuzzy clustering algorithms are analyzed based on the quality of clusters and their computational time by [10]. SOM algorithm is used to map the higher dimensional clustered data in 2-D for visualization after preserving the topological information intact. The quality of clusters is assessed using the performance measures such as discrepancy factor, compactness and distinctness.

A hybrid approach is presented for document clustering which combines the fuzzy c-means (FCM) algorithm with the harmony search (HS) algorithm by [11]. In this proposed method, harmony search is first applied on the documents to find near global optimal clusters. In order to get high quality clusters, FCM is applied with the best vector obtained from the harmony search as initial point. The performance of the hybrid approach is assessed using the measures such as F-measure and purity and it is compared with tradition FCM and HS algorithms. The experimental results exhibit better performance than the traditional algorithms.

III. Proposed Method

The main objective of this work is to improve the quality of the clustering based on the process of concept based indexing and semantic annotation using MeSH concept hierarchy as the domain reference for biomedical documents.

A. Ontology

MeSH published by the National Library of Medicine mainly consists of the controlled vocabulary and a MeSH Tree [12]. The controlled vocabulary contains several different types of terms, such as descriptor, qualifiers, publication types, geographic, and entry terms. Descriptors and Entry terms are used in the proposed indexing method. Descriptor terms are the main concepts or main headings in the ontology. Entry terms are the synonyms or the related terms to descriptors. MeSH descriptors are organized in a MeSH Tree, which can be seen as a MeSH Concept Hierarchy.

B. Concept Based Indexing

Concept based indexing considers both keywords as well as phrases in the documents. The phrases are identified using trigram technique. The concepts in the documents are identified using MeSH ontology concept hierarchy. The weight for the terms and phrases can be calculated based on the semantic relationships exists between them. Concept based weighting scheme computes the significance of the underlying text by converting the documents into a bag of concepts.

In order to convert the unstructured text document into the vector space model the following preprocessing techniques are applied:

1. Tokenization
2. Stop word removal
3. Concept weight calculation

The main objective of the preprocessing is to enhance the quality of data as well as to reduce the complexity involved in clustering. For the given query the keywords of query are searched in the MeSH ontology for its existence. If it exists, the corresponding parent, children and its synonyms will be captured. Based on the query term, the complete path is captured from the root element to the leaf and their weights are assigned based on semantic relationships such as identity, synonymy, hypernymy and meronymy. The text document may encompass multiple concepts. The semantic weight of the each concept is the collective semantic amalgamation of all terms that are closer to the concept in the ontology. Initially, the relations are assigned with the values shown in Table I. The initial level value of meronym is set to 0.8. At each level it is decreased by 0.05 towards the leaf whereas, for hypernyms, the initial level value is assigned as 0.7. The value is decreased by 0.1 towards the root.

TABLE I
SEMANTIC RELATIONS AND THEIR WEIGHTS

Semantic Relation	Weight	Dynamic weight adjustment
Identity	1	-
Synonym	1	-
Hypernym	0.7	0.1
Meronym	0.8	0.05

Since the hypernyms are indicating more general form of the concept, all the upper level keywords are owed with the lesser values than the concept keyword. The meronyms are causative more in the concept identification than the hypernyms. Therefore the meronyms are assigned with the closure values of concept. The identity word and synonymy are given equal weight. Based on the words relations, the following equations are used to calculate the semantic weight of the concept and the abstract of a document.

$$W\left(Concept_i\right)=\frac{\sum_{j\in R} Freq_j \times weight_j}{N} \quad (1)$$

where $W(concept_i)$ is the weight of the $concept_i$, $Freq_j$ is the frequency of the $concept_i$ in the particular relation, $weight_j$ is the semantic weight assigned for the $concept_i$ in that relation and N is the number of unique concepts in the text document. In order to decide the core concept of the document, a concept based vector space model (VSM) is used for document representation. The concept based VSM records the weight of the individual concepts of every document. The concept with the maximum weight indicates the core concept of that particular document.

C. Semantic Annotation

Semantic annotation is the method of identifying pre-defined concepts and entities within a text based on the domain reference. Annotation is about tagging, attaching names, attributes, comments, descriptions, etc. to a document or to a selected part in a text. It provides additional information (metadata) about an existing piece of data, which requires knowing the similarity among documents. Considering an ontology G and a collection of documents D, the computation of annotation on the basis of word-by-word comparison is a straightforward method, but it is expensive in terms of number of comparisons. In order to alleviate this problem, a concept based model is used for annotation.

1. Part-of-Speech (POS) Tagging

POS tagging is the process of annotating each word in a sentence with a part-of-speech marker. POS is tagging is illustrated with the following example: Nephron sparing surgery (NSS) with a minimal tumor-free margin is considered the cornerstone in the contemporary management of renal cell carcinoma stage T1.

The result obtained from POS tagger is given below:

Nephron/***NNP,*** sparing/***VBG***, surgery/***NN***, tumor-free/***JJ***, margin/***NN***, renal/***JJ***, cell/***NN,*** carcinoma/***NN***, stage/***NN*** T1/***CD***

where NNP - Proper singular noun, VBG - Verb, gerund/present participle, NN - Singular noun, JJ – Adjective, CD - Cardinal number

2. Annotation

The secreted meaning of the patterns can be inferred from the domain ontology based on the patterns with similar meanings and the data objects that are co-occurring with it. The nouns are annotated with the underlying concepts of the MeSH ontology through concept mapping.

Nephron$^{\text{kidney}}$ renal$^{\text{kidney}}$ cell carcinoma$^{\text{cancer, kidney, virus, cardiovascular}}$

The importance of the individual concepts is computed in terms of concept weight using (1). The probability of occurrence of individual concepts is recorded in the VSM model. The greatest challenge in the clustering of biomedical documents is most of the diseases are having similar symptoms and are represented with similar terminologies. Therefore most of the nouns are annotated with more than one disease with reference to the MeSH ontology. Hence the traditional partitional and hierarchical clustering methods are not suitable for clustering biomedical documents. This ambiguous condition can be resolved by applying fuzzy c-means algorithm for clustering the biomedical documents.

D. Fuzzy C-Means Clustering Algorithm

FCM is one of the most popularly used algorithms for clustering which allows a piece of data to be in more than one cluster. The documents are assigned to each cluster using fuzzy membership values. These membership values are used to indicate the strength of the association between that data element and a particular cluster. The membership value is computed for each data point with respect to all cluster centers based on the euclidean distance between the data point and the cluster center. The range of membership values lies from 0 to 1. The closer values to 1 indicate that the data point is closely associated with that cluster. Clearly, the summation of membership of each data point should be equal to one. The

algorithm is based on minimization of the following objective function which is calculated based on (2):

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \| x_i - c_j \|^2, 1 \le m < \infty \quad (2)$$

where *m* (the Fuzziness Exponent) is any real number greater than 1, *N* is the number of data points, *C* is the number of clusters, u_{ij} is the degree of membership of x_i in the cluster *j*, x_i is the *i*th of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is the Euclidean distance expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by (3) and (4).

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\| x_i - c_j \|}{\| x_i - c_k \|} \right)^{\frac{2}{m-1}}} \quad (3)$$

where $\|x_i - c_j\|$ is the distance from point *i* to current cluster centre *j*, $\|x_i - c_k\|$ is the distance from point *i* to other cluster centers *k*.

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \quad (4)$$

The iteration will stop when $\max_{ij}\{| u_{ij}^{(k+1)} - u_{ij}^{(k)} |\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas *k* is the iteration steps. This procedure converges to a local minimum or a saddle point J_m.

The algorithm is composed of the following steps:

1. Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$
2. At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. Update $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\| x_i - c_j \|}{\| x_i - c_k \|} \right)^{\frac{2}{m-1}}}$$

4. If $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$,then STOP; otherwise return to step 2.

Clustering is performed on the biomedical documents using both the concept based indexing and the semantic annotation process. The results are compared with the tradition term based clustering using term frequency. The term Frequency-Inverse Document Frequency (TF-IDF) weight is a numerical statistic which shows the importance of a word in a document.

E. Performance Measures

The performance of the proposed methods for biomedical document clustering is assessed using the measures called Dave's validity index and Bezdek index.

1. Bezdek index

Bezdek proposed an index called Bezdek index for validating the performance of the fuzzy c-means algorithm. It is calculated as the sum of the internal products for all the membership values assigned to each point in the output matrix U. Its value ranges between [1/c, 1]. If the value of this index is higher, then the result is more accurate. The index is defined as:

$$V_{pc} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{2} \quad (5)$$

2. Dave's Validity index

Dave's defined the validity measure as:

$$V_{MPC} = 1 - \frac{c}{c-1}\left(1 - V_{pc}\right) \quad (6)$$

Dave's index usually ranges between 0 and 1. If the value of index is higher, then the result is more accurate.

IV. Experimental Results

The datasets used for this experimental analysis are collected from PubMed journals abstracts via MEDLINE. The abstracts contain disease details namely, neoplasm, asthma, conjunctivitis, dengue and cardiovascular system. In each category, 100 documents abstracts are considered and totally five hundred documents abstracts are processed.

Fig. 1 shows the comparison between proposed methods and the traditional term based method of clustering using fuzzy c-means algorithm for Bezdek index. The experimental results show that the proposed clustering algorithms based on MeSH ontology produces better quality clusters than the traditional keyword based method. In the proposed methods annotation based clustering produces high quality clusters than the clustering using concept based indexing method.

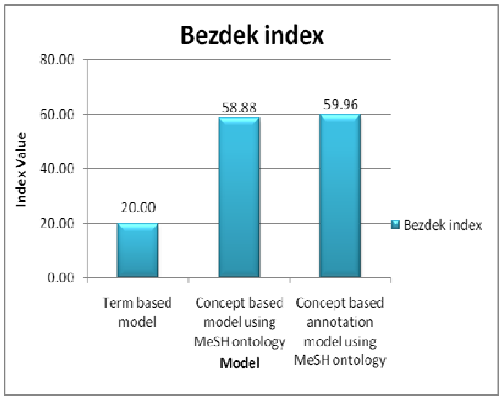

Fig. 1 Comparison of Term Based, Concept Based and Concept Based Annotation Model for Bezdek Index

Fig. 2 shows the comparison between the performance of the proposed methods and the traditional term based method of clustering using fuzzy c-means algorithm for Dave's index. The experimental results show that the proposed clustering algorithms based on MeSH ontology produces better quality clusters than the traditional keyword based method. In the proposed methods also, annotation based clustering produces high quality clusters than the clustering using concept based indexing method.

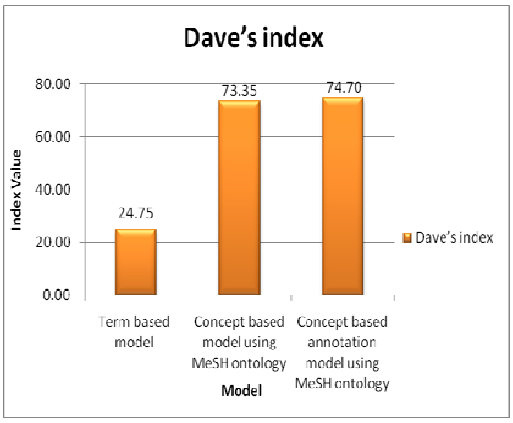

Fig. 2 Comparison of Term Based, Concept Based and Concept Based Annotation Model for Dave's Index

V. Conclusion

The traditional clustering algorithms, both k-means and hierarchical algorithms are suitable for hard clustering in which each document is assigned as a member of exactly one cluster. These algorithms are not much suitable for clustering the biomedical documents because of the existence of the more common terms in different domains of it. Semantic relations are also not considered in traditional hard clustering algorithms. Concept based indexing and annotation methods are proposed in this work for clustering the biomedical documents using the semantic relations that are derived from the MeSH ontology. The performances of the proposed methods are analyzed using the measures Bezdek and Dave's validity indices and compared with the traditional term based clustering. The results show that concept based methods outperform the traditional term based method. Concept based annotation produces better quality clusters than the concept based indexing method.

References

[1] Jonquet, Clement, Mark A. Musen, and Nigam Shah,"A system for ontology-based annotation of biomedical data," *Data Integration in the Life Sciences, Springer Berlin Heidelberg*, pp. 144-152, 2008.

[2] Adrien Coulet, Florent Domenach, Mehdi Kaytoue and Amedeo Napoli, "Using pattern structures for analyzing ontology-based annotations of biomedical data," *in Proc. Formal Concept Analysis,. Springer Berlin Heidelberg*, pp. 76-91, 2013.

[3] Fontes, Celso Araujo, Maria Claudia Cavalcanti, and Ana Maria de C. Moura. "An Ontology-Based Reasoning Approach for Document Annotation," *in Proc. IEEE Seventh International Conference on Semantic Computing (ICSC)*, pp. 160-167, 2013.

[4] Tsatsaronis, George, Natalia Macari, Sunna Torge, Heiko Dietze, and Michael Schroeder, "A maximum-entropy approach for accurate document annotation in the biomedical domain," *J. Biomedical semantics*, vol. 3, no. 1, pp.1-17, 2012.

[5] Hazman, Maryam, Samhaa R. El-Beltagy, and Ahmed Rafea, "An Ontology Based Approach for Automatically Annotating Document Segments," *Int. J. Computer Science Issues (IJCSI)*, vol. 9, no. 2, pp.221-230, 2012.

[6] Kiryakov, Atanas, Borislav Popov, Ivan Terziev, Dimitar Manov, and Damyan Ognyanoff, "Semantic annotation, indexing, and retrieval." *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 2, no. 1, pp. 49-79, 2004.

[7] Cheung, Warren A., BF F. Ouellette, and Wyeth W. Wasserman, "Quantitative biomedical annotation using medical subject heading over-representation profiles (MeSHOPs)," *BMC bioinformatics*, vol.13, no. 249, pp.1-11, 2012.

[8] Chua, Watson Wei Khong, and Jung-jae Kim, "Semantic querying over knowledge in biomedical text corpora annotated with multiple ontologies," *in Proc. of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pp. 400-407, 2012.

[9] W. Shuguang and H. Milos , 'Keyword annotation of biomedical documents with graph-based similarity methods', *in Proc. of* IEEE international conferences on bioinformatics and biomedicine, pp. 361-364, 2012.

[10] Chattopadhyay, Subhagata, Dilip Kumar Pratihar, and Sanjib Chandra De Sarkar, "A Comparative Study of Fuzzy C-Means Algorithm and Entropy-Based Fuzzy Clustering Algorithms," *Computing & Informatics*, vol. 30, no. 4, pp. 701-720, 2011.

[11]] Kang, Jiayin, and Wenjun Zhang, "Combination of Fuzzy C-means and Harmony Search Algorithms for Clustering of Text Document," *Journal of Computational Information Systems*, vol. 7, no. 16,pp. 5980-5986, 2011.

[12] Sridevi, U. K., and N. Nagaveni, "An ontology based model for document clustering," *Int. J. Intelligent Information Technologies (IJIIT), vol.* 7, no.3, pp. 54-69, 2011.