Proc. of Int. Conf. on Computing Electrical and Electronics Engineering, CEEE

An Efficient Semantic Relationship based Concept Indexing for PubMed Document Clustering using MeSH Ontology

Logeswari S[1] and Gomathi R[2]
[1-2]Bannari Amman Institute of Technology
Email: logeswaris@bitsathy.ac.in, gomathir@bitsathy.ac.in

***Abstract*—Text mining is critical for information retrieval, especially in the health-care field. The results of text mining are based on noisy data and false positives from natural language text. This is owing to inconsistencies in semantics, syntax, the scarcity of class-specific core terms, and the enormous complexity of the data. As a result, traditional term/based techniques are ineffective for clustering medical articles containing complicated semantics. As a result, a semantic-relationship based indexing technique for categorising research publications in the healthcare sector is proposed with MeSH ontology as a domain reference to extract the relevant relationships between the document's terms.**

***Index Terms*— Text mining, Document clustering, Ontology, Indexing, MeSH, Euclidean distance.**

I. INTRODUCTION

Medical document clustering is one of the issues faced by researchers because of its ambiguity and incompatibility. Words or phrases are the natural elements of text documents, and a document collection can contain millions of individual features. Medical documents have a greater number of generic words than core words that are unrelated to the search query. Due to the lower frequency of class dependent core terms in medical documents, clustering becomes more difficult.

As a logical evolution of accessible technology to deal with the information assault, ontology-based computing has promise. The background knowledge offered by ontology can be used to improve the document representation. In the conventional Bag-of-Words approach, it is one of the solutions for the synonym and polysemy difficulties. Medline and PubMed repositories are rich in medical literature [7]. The exponential growth of these online repositories and availability of large volume of documents has motivated the search for the hidden knowledge from these resources. An automated system is essential to extract information from PubMed. A semantic relationship based indexing approach is proposed to group the articles available in PubMed repository based on the diseases which are known as concepts. The proposed method converts words/phrases in the documents into concepts and characterizes diseases using ontology mapping.

II. LITERATURE REVIEW

The ability of traditional query-based information search to discover abstracts based on diseases is restricted. The query - based search method may be appropriate for content -focused querying. It necessitates the user's

Grenze ID: 02.CEEE.2022.7.505_1

understanding of the subject as well as the ability to select keywords for the search items.

Elsayed et al. (2015) proposed a distributed implementation of the bisecting k-means utilising the MapReduce programming architecture to handle the difficulty of clustering dense data documents. They've also presented an improved document clustering method that combines the WordNet ontology with bisecting k-means to take advantage of semantic relationships between words. The results of the experiments demonstrate that utilising lexical categories for nouns only improves internal evaluation measures of document clustering and reduces the number of features in the texts from thousands to tens. They used Amazon Elastic MapReduce to install the Bisecting k-means algorithm and examined the performance of their proposed solutions.

Jasila et al. (2019) proposed a hybrid strategy that focused on both the pre-processing and clustering algorithms. It is proposed with the goal of lowering the number of features and execution time, removing synonymous difficulties, and improving clustering accuracy. The measure of similarity is cosine similarity. A feature extraction method based on WordNet Ontology is used in the pre-processing stage. The Red Black Tree based sorting approach is used to get the initial centroids in clustering. By maintaining the line of similarity between data points and nearest cluster centroids, the data points are assigned to the appropriate clusters using a unique approach. The findings of certain current cosine similarity clustering techniques are compared to those of new clustering technique. On the Newsgroup dataset, the suggested hybrid technique performs better, with significant gains in dimensionality reduction, running time, and accuracy.

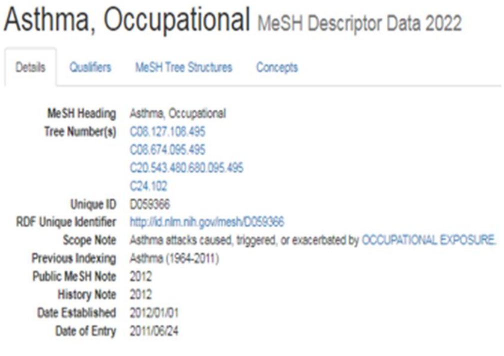
Shanavas et al. (2020) presented an ontology-based strategy for classifying medical documents. With the use of a domain-specific similarity matrix developed using Unified Medical Language System (UMLS) ideas in the training documents, a concept graph is automatically constructed and enriched for each medical text document in this suggested method. A graph kernel is used to compare medical text texts based on their enriched idea graphs. The comparison result is then used to classify the data. This method has the advantage of allowing domain information to be incorporated into the classification framework. Experiments on biomedical abstracts and clinical report classification reveal that the proposed method is effective.

Using the weighted concepts derived from the domain ontology [1], the semantic based indexing method could increase the accuracy and performance of text document clustering, according to this research review. This literature review lays the way for the development of semantic-based indexing techniques for medical document clustering, using the MeSH ontology as the domain reference.

III. PROPOSED METHOD

The controlled vocabulary and a MeSH Tree are the main components of MeSH, which is published by the National Library of Medicine. In the proposed indexing method, descriptors and entry words are used. Entry phrases are synonyms or terms that are similar to descriptors. Hidden concepts and semantic relationships are extracted from documents via ontology mapping. The proposed concept-based model evaluates terms in the document using the tri-gram technique.

The system is given an abstract of the PubMed article and the disease name that describes the concept. The documents are tokenized and stop-words are deleted during pre-processing. The subject hierarchy for each search word is extracted from the MeSH ontology. For the search query "Asthma," Figure 1 shows the MeSH descriptor table, and Figure 2 shows the MeSH tree hierarchy. In general, hypernyms indicate a term's forebears, whereas meronyms reflect a term's descendants in the concept hierarchy [2,3].



Asthma, Occupational MeSH Descriptor Data 2022

MeSH Heading	Asthma, Occupational
Tree Number(s)	C08.127.108.495 C08.674.095.495 C20.543.480.680.095.495 C24.102
Unique ID	D059366
RDF Unique Identifier	http://id.nlm.nih.gov/mesh/D059366
Scope Note	Asthma attacks caused, triggered, or exacerbated by OCCUPATIONAL EXPOSURE.
Previous Indexing	Asthma (1964-2011)
Public MeSH Note	2012
History Note	2012
Date Established	2012/01/01
Date of Entry	2011/06/24

Figure 1. MeSH Descriptor Table for Asthma

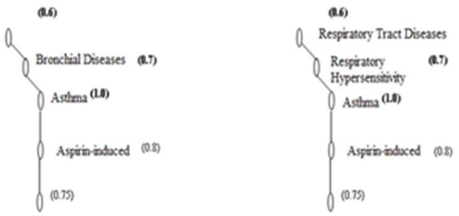


Figure 2. MeSH Tree Hierarchies for Asthma

A. Pseudo code for Concept Based Indexing

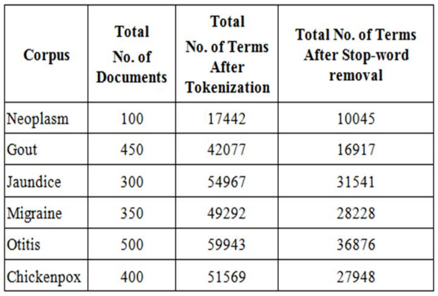
The pseudo code for the semantic relationship based indexing is given as follows [3]:

```
Input: Document Corpus from PubMed
Output: Vector Space Model with Concept – Document Weight Information
Method:
 Given the set of abstracts and search query
1.    Tokenize the content of documents
2.    Extract only nouns in the abstract after eliminating all the stop-words in it
3.    For the semantic based indexing, do:
        a.    Extract the possible tree hierarchies from MeSH ontology for the search query
        b.    Assign the initial weights for the immediate hypernym as 0.7 and meronym as 0.8 of the search item.
        c.    For the ancestors of hypernym, the value is decreased by 0.1 for each predecessor level of the hierarchy
        d.    For the subsequent meronyms, the value is decreased by 0.05 for each successive level of the hierarchy
              $W(C_i, d) = \sum_{j \epsilon R} \sum_{i=1}^{n} \frac{freq_i^{(j)} \times weight_i}{n}$
4. Compute the total concept weight for the possible diseases in the given abstract and store it in traditional
   VSM
```

B. Experimental Result Analysis

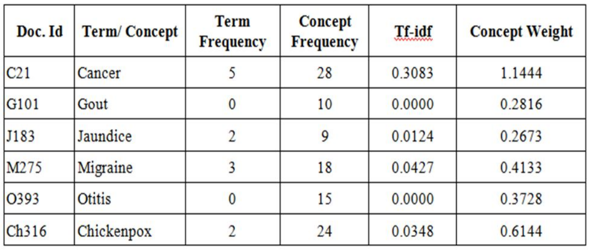
The performance of the proposed method is compared to traditional indexing techniques, term-based indexing, and the Latent Semantic Indexing (LSI) indexing technique. The number of clusters is set to 6 for this experiment, while the maximum number of algorithm iterations is set at 20. The results of the pre-processing on the document corpuses are shown in Table II and the Table III depicts the comparison between the conventional tf-idf weight and concept weight using the proposed approach for some sample abstracts that are considered as input from PubMed repository.

TABLE II. RESULTS OF THE PRE-PROCESSING



Corpus	Total No. of Documents	Total No. of Terms After Tokenization	Total No. of Terms After Stop-word removal
Neoplasm	100	17442	10045
Gout	450	42077	16917
Jaundice	300	54967	31541
Migraine	350	49292	28228
Otitis	500	59943	36876
Chickenpox	400	51569	27948

TABLE III. COMPARISON OF TF-IDF AND CONCEPT WEIGHT FOR SOME SAMPLE DOCUMENTS



Doc. Id	Term/ Concept	Term Frequency	Concept Frequency	Tf-idf	Concept Weight
C21	Cancer	5	28	0.3083	1.1444
G101	Gout	0	10	0.0000	0.2816
J183	Jaundice	2	9	0.0124	0.2673
M275	Migraine	3	18	0.0427	0.4133
O393	Otitis	0	15	0.0000	0.3728
Ch316	Chickenpox	2	24	0.0348	0.6144

For hard clustering methods, the internal validity index, silhouette index is used to determine the appropriate number of clusters and the quality of clusters based on Euclidean distance and Pearson correlation coefficient in Figure 2 and in Figure 3 respectively.

Starting from cluster size 3 move onwards, the proposed concept based indexing outperforms the traditional word based and LSI based indexing methods, according to the experimental results. It is also discovered that, regardless of cluster size, the K-Means algorithm outperforms the other hierarchical set of algorithms. In comparison to Euclidean distance, Pearson correlation coefficient is the more effective of the two proximity measures.

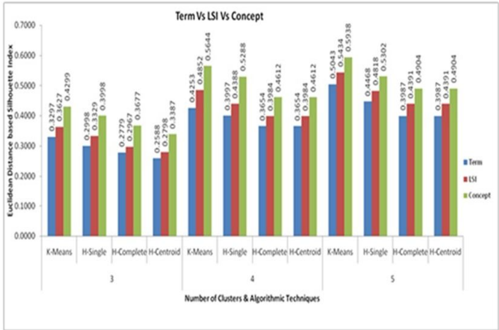


Figure 2. Performnace Analysis-Euclidean Distance based Silhouette Index

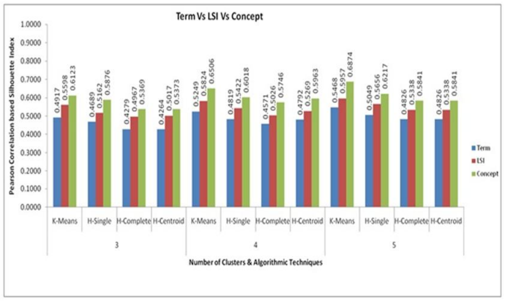


Figure 3. Performance Analysis - Pearson Correlation based Silhouette Index

IV. Conclusion

This paper proposed a semantic relationship-based indexing for clustering medical documents in the PubMed repository. For idea extraction and the identification of semantic links between terms and texts, this method proposes using the MeSH domain ontology as background knowledge. Both hierarchical and partitional clustering algorithms are used to evaluate the suggested approach's performance. The extraction of the nouns expressing the key concepts resolves the sparsity of the core words due to the high dimensionality. The experimental results show that the suggested concept-based indexing is consistently more effective than typical BOW indexing methods.

References

[1] Tar, H. H., & Nyunt, T. T. S. (2011). Ontology-based concept weighting for text documents. World Academy of Science, engineering and Technology, 57, 249-253.

[2] Soe, T. L. (2014). Ontology-Based Indexing and Semantic Indexing in Information Retrieval Systems. International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), 1, 1-9

[3] Shanmugam, L., & Premalatha, K. (2019). Semantic-Based Indexing Approaches for Medical Document Clustering Using Cognitive Search. In Cognitive Social Mining Applications in Data Analytics and Forensics (pp. 41-64). IGI Global.

[4] Alghoson, A. M. (2014, January). Medical document classification based on mesh. In 2014 47th Hawaii International Conference on System Sciences (pp. 2571-2575). IEEE.

[5] Elsayed, A., Mokhtar, H. M., & Ismail, O. (2015). Ontology based document clustering using mapreduce., International Journal of Database Management Systems (IJDMS) Vol.7, No.2, DOI : 10.5121/ijdms.2015.7201.

[6] Ambreen Zafar, Muhammad Awais and Muhammad Ahmad Aftab.(2018).Ontology Based Document Data Analysis, IJCSNS International Journal of Computer Science and Network Security, VOL.18 No.11,pp.42-48.

[7] Pita Costa, J., Stopar, L., Fuart, F., Grobelnik, M., Santanam, R., Sun, C. & Wallace, J. (2018). Text mining MEDLINE to support public health.

[8] Jasila, E. K., Saleena, N., & Nazeer, K. A. (2019, December). Ontology Based Document Clustering-An Efficient Hybrid Approach. In 2019 IEEE 9th International Conference on Advanced Computing (IACC) (pp. 153-157). IEEE.

[9] Shanavas, N., Wang, H., Lin, Z., & Hawe, G. (2020). Ontology-based enriched concept graphs for medical document classification. Information Sciences, 525, 172-181.