Exploring Diseases based Biomedical Document Clustering and Visualization using Self-Organizing Maps

Setu Shah
Purdue School of Engineering and Technology
Indiana University–Purdue University Indianapolis
Indianapolis, USA
setshah@iupui.edu

Xiao Luo
Purdue School of Engineering and Technology
Indiana University–Purdue University Indianapolis
Indianapolis, USA
luo25@iupui.edu

***Abstract*—Document clustering is a text mining technique used to provide better document search and browsing in digital libraries or online corpora. In this research, a vector representation of concepts of diseases and similarity measurement between concepts are proposed. They identify the closest concepts of diseases in the context of a corpus. Each document is represented by using the vector space model. A weight scheme is proposed to consider both local content and associations between concepts. Self-Organizing Maps (SOM) are often used as document clustering algorithm. The vector projection and visualization features of SOM enable visualization and analysis of the cluster distribution and relationships on the two dimensional space. The Davies-Bouldin index is used to validate the clusters based on the visualized cluster distributions. The results show that the proposed document clustering framework generates meaningful clusters and can facilitate clustering visualization and information retrieval based on the concepts of diseases.**

I. Introduction

Active research in the medical and biomedical domain has generated pervasive documents and articles. It is estimated that more than 10,000 articles are added to MEDLINE weekly [1]. There is a continuing need for development of techniques to discover and search these documents and articles from the concepts of diseases point of view. Biomedial document clustering based on the concepts of diseases can provide an overview of the literature repository based on the diseases and relationships between the diseases, so that researchers can further explore or review the articles in certain clusters that are related to their research interests. Biomedical document clustering is different from the general text document clustering task, because in the latter, semantic similarities between words or phrases are not usually considered. One medical concept of disease might be represented in different forms, and some medical concepts of diseases might be highly correlated. For example, 'Type 2 Diabetes' is the same concept of disease as 'Diabetes Mellitus Type 2'. 'Hypertension' might co-occur often with 'Stroke'. In order to capture the semantic similarities between words or phrases, previous research on document representation reforming [2] [3] [4] often uses existing ontology such as MeSH or WordNet to identify the semantic relationships. However, ontology doesn't reflect the co-occurrences of medical concepts. This paper focuses on biomedical document clustering based on the concepts of diseases. The proposed similarity measure between the concepts of diseases is based on the Word2Vec model [5]. It identifies the closest concepts based on co-occurrences of the concepts. The proposed concept weighting scheme is a linear combination of the TF-IDF value which reflects the content similarity between documents and the similarity score based on the proposed similarity measurements that reflect the semantic similarity between documents.

Most of the research related to biomedical document clustering focuses on reforming the representation of biomedical documents to improve the clustering performance without investigating the visualization of the clustering results. Visualization of the clustering results can facilitate information retrieval on the biomedical document repository. In this research, the unsupervised learning algorithm Self-Organizing Maps (SOM) [6] is used as the clustering technique. A basic SOM consists of M neurons which can be projected to a low dimensional grid (usually 1 or 2 dimensional) [6]. The algorithm for the formation of the SOM involves three basic steps after initialization: sampling, similarity matching, and updating. These three steps are repeated until formation of the feature map has completed. The most commonly used visualization techniques of SOM are the U-matrix and hit histogram. The U-matrix [6] holds all distances between neurons and their immediate neighbor neurons. The U-matrix gives a direct visualization of the number of clusters and their distribution. The hit histogram of the input data set on the trained SOM map provides a visualization that details the distribution of input data across the clusters. Each input data instance in the data set can be projected to the closest neuron on a trained SOM map. The hit histogram is constructed by counting the number of hits each neuron receives from the input data set.

In this work, Davies-Bouldin index is used to validate the clusters and centroids that are visualized through the U-matrix and hit histogram. The overall clustering framework has been evaluated on a subset of PubMed Central Open Access. The results show that the proposed system can group documents to

This is the author's manuscript of the article published in final edited form as:
Shah, S., & Luo, X. (2017). Exploring diseases based biomedical document clustering and visualization using self-organizing maps. In 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom) (pp. 1–6).
https://doi.org/10.1109/HealthCom.2017.8210791

meaningful and visualizable clusters based on the concepts of diseases. It can be further used to assist information retrieval in large biomedical document repositories based on diseases.

The rest of the paper is organized as follows. In section 2, related work is described. Section 3 demonstrates how the concepts of diseases are extracted by using UMLS MetaMap. Section 4 and 5 detail the measurement of concept similarity and weighting scheme for each concept in the document representation. Experimental settings and results are given in section 6. Section 7 concludes this research and discusses potential future work.

II. Related Work

A lot of research has been done in biomedical document clustering in past decades. Some of it focused on document presentation reforming based on medical ontology or on using different weighting scheme other than TF-IDF, while some others focused on investigating various clustering algorithms. Few of them discussed visualization of the clustering results to facilitate biomedical information retrieval.

Zhang et al. [4] reviewed three different ontology based term similarity measurements: path based [7], information content based [8], and feature based [9] and then proposed their own similarity measurement and term re-weighting scheme. K-means algorithm is used for document clustering. Based on the results comparison, some of them are slightly worse than the word based scheme. The authors mentioned that it might because of the limitation of the domain ontology, term extraction and sense disambiguation. Visualization of the relationships between the clusters was not included in this research.

Yoo et al. [1] used a graphical representation method to represent a set of documents based on the MeSH ontology, and proposed the document clustering and summarization with this graphical representation. The document clustering and summarization model gained comparable results on clustering and also provided some visualization on the documents cluster model based on relationships of the terms. However, this visualization relies largely on the MeSH ontology instead of the document content and relationships themselves.

Logeswari et al. [2] proposed a concept weighting scheme based on the MeSH ontology and tri-gram extraction to extract concepts from a text corpus. The semantic relationship between tri-grams are weighted through a heuristic weight assignment of four predefined semantic relationships. The K-means clustering algorithm results show that concept based representation was better than word based representation. Visualization of the clustering results was not investigated.

Gu et al. [10] proposed a concept similarity measurement by using a linear combination of multiple similarity measurements based on MeSH ontology and local content which included TF-IDF weighting and co-efficient calculation between related article sets. A semi-supervised clustering algorithm was employed at the stage of document clustering. Their focus was not clustering visualization.

Some research has been done about the visualization process to support biomedical literature search. Gorg et al. [11] developed a visual analytics system, named Bio-Jigsaw by using the MeSH ontology. This research demonstrated how visual analytics can be used to analyze a search query on a gene related to breast cancer. Neither document representation nor document clustering were discussed.

To the best of authors' knowledge, this research is the first to present concepts of diseases based on the Word2Vec model without an ontology. The proposed similarity measurement and concept weighting scheme are first applied to document clustering. SOM is then employed to visualize the distribution of document clusters based on the concepts of diseases.

III. Concepts of Diseases Extraction

In this work, the focus is on clustering biomedical documents based on the concepts of diseases that are mentioned in the documents. To extract the concepts of diseases from the documents, Unified Medical Language System (UMLS) MetaMap is used. UMLS MetaMap [12] is a natural language processing tool that makes use of various sources such as UMLS Metathesaurus [13] and SNOMED CT [14] to map the phrases or terms in the narrative text to different semantic types.

Given a sentence 'A retrospective evaluation of Haemophilus influenzae type b meningitis observed over 2-year period documented 86 cases', the phrase 'Haemophilus influenzae type b meningitis' is identified as semantic type 'disease or syndrome' and mapped to phrase 'Type B Hemophilus influenzae Meningitis' based on the lexicon that UMLS MetaMap uses. In this research, if a term or phrase is mapped to semantic types 'Disease or Syndrome' or 'Neoplastic Process', the corresponding phrase in the lexicon produced by MetaMap is extracted.

IV. Concepts Similarity Measure

In the biomedical literature, same concepts of a disease can be presented by different terms or combinations of words. For example, 'cancer of breast' and ' breast cancer' are two phrases that present the concept of the same disease. However, they are treated as two different concepts if typical vector space model and TF-IDF weighting scheme are used for document presentation, and the semantic similarity between them is not measured. In this research, a semantic similarity measure between different concepts of diseases is proposed. Given a total of L concepts of diseases extracted from the raw text corpus, the similarities between any two concepts are stored in the similarity matrix S as presented in Equation 1. Each entry $s_{i,j}$ in the matrix S represents the similarity between concept C_i and C_j.

$$S_{L,L} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,L} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ s_{L,1} & s_{L,2} & \cdots & s_{L,L} \end{pmatrix} \quad (1)$$

To calculate the similarity between two concepts, first, each word is represented by a vector (as proposed in Equation 2). This vector representation is learned by training the Word2Vec model. The Word2Vec training algorithm was developed by a team of researchers at Google led by Tomas Mikolov [5]. It is a computationally-efficient algorithm to generate vectors of real numbers to present words in a given raw text corpus. These vector representations are learned through a three-layer recurrent neural network by using either a continuous bag-of-words approach or a skip-gram architecture. The vectors preserve the distances between words in the vector space so that the words that share common contexts in the raw text corpus are located in close proximity to one another. The dimension of the vector created depends on the number of neurons in the hidden layer of the recurrent neural network when training a Word2Vec model.

$$Word = (wv_1, wv_2, \ldots, wv_m) \tag{2}$$

m: the dimension of the vector.

In this research, a trained Word2Vec model [15] that is created from a subset of PubMed literature database and a subset of PubMed Central (PMC) Open Access database is employed. These two text corpora contain a large number of biomedical documents. The trained model creates 200 dimensional vectors to present the words extracted in the two text corpus. The skip-gram architecture with a window size of 5 is adopted for the learning process [15].

Although some of concepts of diseases contain only one word, many of them include multiple words. In this work, if a concept of disease includes multiple words, a concept vector is generated by aggregating the vectors of all the words in the concept, as shown in Equation 3. For example, for the disease 'diabetes mellitus', the vector for 'diabetes' and the vector for 'mellitus' are aggregated by adding them together.

$$C = \sum_{i=1}^{M} Word_i \tag{3}$$

M: the total number of words in a concept C.

The similarity score between the concepts are calculated using the cosine distance between the vectors as shown in Equation 4.

$$S_{i,j} = \frac{\mathbf{C_i} \cdot \mathbf{C_j}}{\sqrt{\sum_{i=1}^{m} C_i^2} \cdot \sqrt{\sum_{i=1}^{m} C_j^2}} \tag{4}$$

By presenting concepts in vector and using this similarity measure, it is observed that the more the diseases are associated, the higher the similarity scores between them are. Table I provides some examples of concepts of diseases and the their top 3 closest concepts based on the similarity scores. 'Hypertension' is often associated with 'Hyperlipidaemia' in the literature, so the similarity between them is high.

TABLE I
EXAMPLES OF CONCEPTS AND THE TOP 3 CLOSEST CONCEPTS BASED ON THE SIMILARITY SCORES

Concept	**Closest Concepts**	**Score of Similarity**
hypertension	essential hypertension	0.813
	hyperlipidaemia	0.692
	dyslipidemia	0.659
endothelial dysfunction	dysfunction	0.739
	renal dysfunction	0.660
	cortical dysfunction	0.639
carpal tunnel syndrome	bilateral carpal tunnel syndrome	0.970
	cts carpal tunnel syndrome	0.957
	carpal tunnel	0.941
diabetes	diabetes mellitus	0.918
	diabetes mellitus type ii	0.868
	dm diabetes mellitus	0.845
cardiovascular disease	cardiac diseases	0.8181
	metabolic diseases	0.8179
	heart diseases	0.787

V. DOCUMENT REPRESENTATION AND WEIGHTING SCHEME

In this research, the typical vector space model is used to present a biomedical document, each entry of the vector corresponds to a concept of disease which is identified through the UMLS MetaMap. The proposed weight ($Weight_{C_{i,d}}$) that is given to each concept ($C_{i,d}$) is calculated as equation 5:

$$Weight_{C_{i,d}} = \begin{cases} tf_{C_i,d} \times \log \frac{|D|}{df_{C_i}} + \sum_{j=1}^{M} S_{i,j} & tf_{C_i,d} > 0 \\ \sum_{j=1}^{N} \frac{N-(j-1)}{N} S_{i,j} & tf_{C_i,d} = 0 \end{cases} \tag{5}$$

df_{C_i}: the number of documents in which concept C_i occurs at least once

$tf_{C_{i,d}}$: frequency of concept C_i in document d

$|D|$: total number of documents in the corpus

$S_{i,j}$: the similarity between C_i and concept C_j that both occur in document d. C_j is the j^{th} frequent concept in the document d.

M: the total number of concepts in document d.

N: top N closest concepts of C_i. In this research, $N = 3$.

If a concept occurs in a document, the weighting scheme uses the TF-IDF value to underline the occurrence of the concept in the local content. The $\sum_{j=1}^{M} S_{i,j}$ calculates the sum of similarity scores between the occurred concept $C_{i,d}$ and other concepts ($C_{j,d}$, $j = 1, \ldots, M$) that also occurs within the document. If a concept does not occur in the document, the weight is calculated by a weighted sum of the top 3 closest concepts ($C_{j,d}$, $j = 1, \ldots, 3$) that appear in the document based on the similarities scores. By using this weighting scheme, the representation measures the occurrences of different representations of the same or similar concepts. For example, 'diabetes' occurs in one document, but 'diabetes mellitus' occurs in another document. By using the traditional TF-IDF weighting scheme, their values would be 0 for documents in which the concept does not appear. However, by using the proposed weighting scheme, they are

TABLE II
PMC-OA DATA SET

Name of journal	# of documents
American Journal of Hypertension	13
Augmentative and alternative communication	2
Ancient Science of Life	3
Bioinformatics and biology insights	45
Allergy and asthma proceedings	28
BoneKEy reports	4
Anesthesia, essays and researches	135
Biological trace element research	31
Bone Marrow Research	1
Brain and language	1
American journal of physiology. Endocrinology and metabolism	11
Aphasiology	3
Annals of rehabilitation medicine	323

weighted based on the similarity between the concept and its closest concepts. Thus, for the document that does not contain the concept 'diabetes mellitus', instead of using 0, the similarity score between 'diabetes mellitus' and other concepts that appear in the document is used.

VI. EXPERIMENT SETTING AND RESULT ANALYSIS

To evaluate the proposed biomedical document clustering framework, a subset of a large biomedical document collection – PubMed Central Open Access has been used. Since this data set is not labeled by the concepts of diseases, external clustering validation metric, such as purity or F-measure, is not suitable to validate the clustering results. Hence, one of the internal clustering validation metric - Davies-Bouldin index (DB index) is used to validate the clustering visualization based on the cluster distributions. The DB index calculation is detailed in sub-section VI-B. The details of the document collection and the corresponding clustering results, visualization and validation are detailed in the following subsections.

A. Data Set - PubMed Central Open Access (PMC-OA)

PubMed Central Open Access document set [17] has been used by many research projects to examine tasks of biomedical literature clustering and classification [4] [18]. It is a subset of over 1 million articles from the total collection of articles in PMC. For this research, a set of 600 articles were randomly selected from journals whose names start with letter 'A' or 'B'. The number of selected articles from each journal is shown in Table II. In this research, only content in the 'Title' and 'Abstract' sections from these documents are used. Figure 1 shows the distribution of these concepts based on the number of words in each concept. Over 50% of the concepts of diseases contain two words, about 20% of them have one or three words and less than 10% of them have over four words.

Figure 2 shows the distribution of concepts based on their document frequencies. Over 70% of the concepts have document frequency as 1, and about 2% of the concepts have a document frequency value over 10. The nature of the sparseness makes it hard to retrieve the documents from a

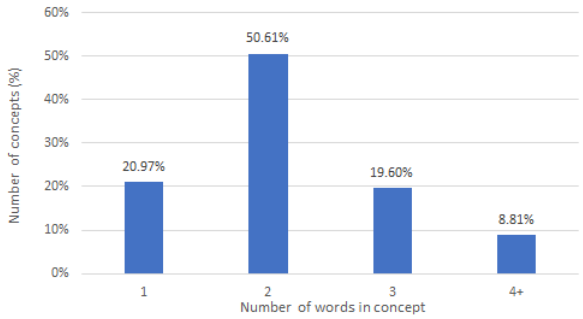

Fig. 1. Distribution of the concepts based on the number of words

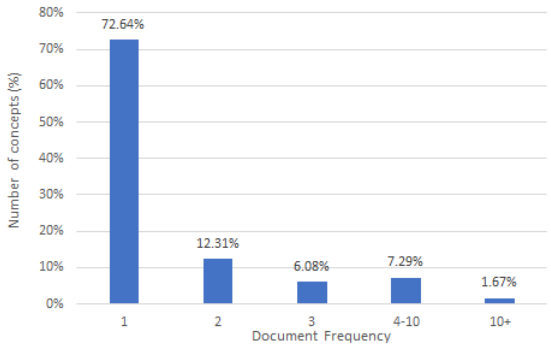

Fig. 2. Distribution of the document frequencies

biomedical document repository using the traditional vector space model representation and weighting scheme.

B. Cluster Validation Metric - Davies-Bouldin index

Typically, there are two types of evaluation metrics: internal evaluation and external evaluation. The internal evaluation is to formalize the goal of attaining high intra-cluster similarity and low inter-cluster similarity, whereas, the external evaluation which based on the interest of an application, such as categorization. Since the data set used in this research have no assigned categories, the interal evaluation Davies-Bouldin index (DB index) [19] is used to validate the clustering results. DB index has been used to validate the clustering results of the SOM in the previous research [20] [21]. The calculation of the Davies-Bouldin index is shown in Equation 6, where SD_i is the standard deviation of the distance of samples in a cluster to the respective cluster centroid, $d(CL_i, CL_j)$ is the Euclidean distance between centroids CL_i and CL_j, NC is the total number of centroids. The more distinct the clusters are from each other, smaller the DB index value is.

$$DBindex = \frac{1}{NC}\sum_{i=1}^{NC}\max_{j,j\neq i}\frac{SD_i + SD_j}{d(CL_i, CL_j)} \tag{6}$$

In this research, DB index is calculated to identify the best partitions for the clustering based on the visualized cluster distributions through U-matrix and hit histogram of the SOM.

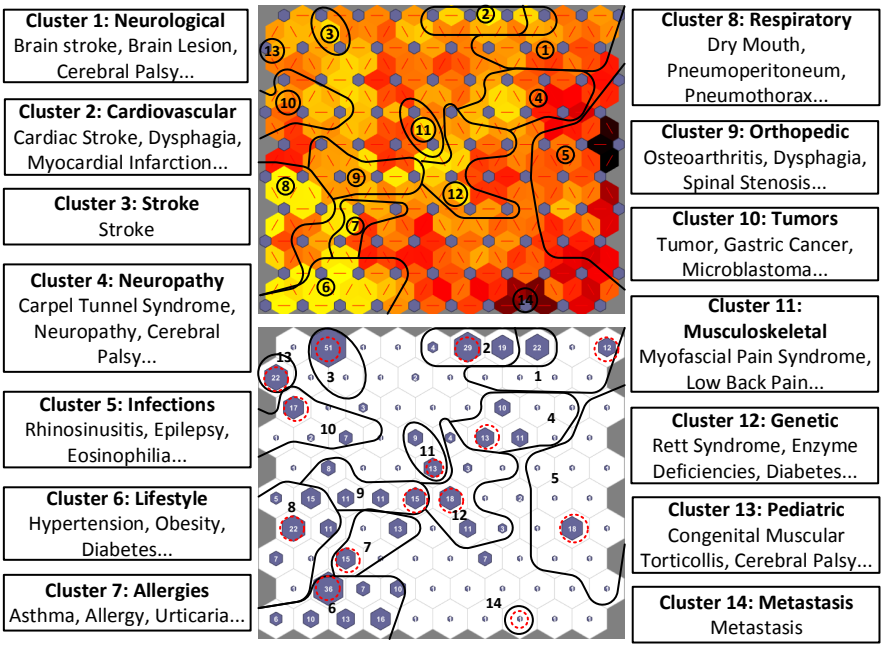

Fig. 3. Clustering results of PubMed Central Open Access

C. Clustering Visualization, Validation and Discussion

SOM has been used for document clustering after concepts extraction and document representation using the proposed weighting scheme. The size of the map is 10 by 10 which contains 100 neurons. The training iterations are set to be 50,000.

Through the U-matrix and hit histogram, 11 clusters (Clusters 2, 3, 4, 6, 7, 8, 9, 10, 11, 12 and 13 in Figure 3) are identified initially. A neuron of each cluster is selected as centroid, then DB index based on the partitions is calculated to decide whether the partition is optimal. The lower the DB index value is, the better the partition is. It is observed that lower DB index value is returned when the neurons with highest number of hits are chosen as the centroids. Figure 3 shows the major clusters visualized on the SOM map. The centroids which are selected through the calculation of the DB indexes are marked with a red dotted circle. The clusters are marked with a black boundary. The visualized major clusters do not include all the input data. Some of the data hit the neurons that are far away from the 14 centroids as visualized. In order to fully evaluate the best number of clustering partitions to cover all the input data instances, we have increased the number of clusters by adding the neurons that are not covered by the 14 clusters as centroids. Figure 4 shows that the DB index value decreases as the number of clusters are increased. That is because adding clusters to separate the data that is far away from the existing centroids creates better cluster partitions.

Through further analysis on the major clusters, we have discovered the concepts of diseases of each cluster as shown in Figure 3. A majority of documents in cluster 1 are articles that discuss neurological diseases like brain strokes, brain lesions, cerebral palsy and diseases that lead to speech disorders; most of the documents in cluster 2 are related to cardiovascular diseases such as 'hypertension', 'coronary artery disease', 'ischemic strokes' and so on. Cluster 1 and cluster 2 have one over lapped neuron on the top right of the map, it is because over half of the documents that hit this neuron discuss both neurological and cardiological 'strokes' concepts. The distances between the neurons within cluster 5 are larger than that of the other neurons. The diseases discussed in the documents in this cluster include infections like 'rhinosinusitis', 'epilepsy', 'eosinophilia' and so on. Other concepts that are found in this cluster are 'seizures' and obstructions of intestines and throat. Although these concepts are not very closely related, they are more closely related to each other than to the concepts in other clusters. Cluster 6 has documents related to 'obesity', 'diabetes', 'hypertension'

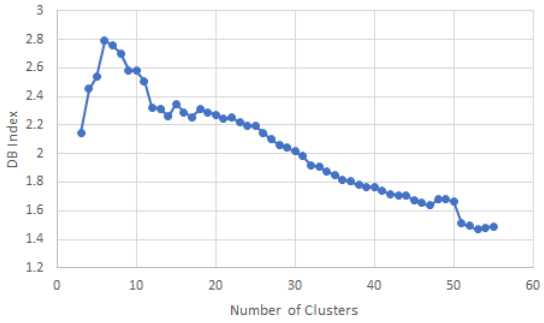

Fig. 4. DB index over the changing number of clusters.

and 'hyperglycemia'. The concept 'coronary artery disease' is also discussed in some documents of this cluster. We discovered that it is because some articles discuss 'coronary artery disease' as a possible outcome of 'hypertension', 'hyperglycemia' or their combination. Cluster 11 has neurons within short distances of cluster 9. This proximity is also seen in the form of the diseases discussed by the documents of these clusters, since muscle pain and orthopedic concepts are highly related to each other. Cluster 12 is very closely located to cluster 4, 9 and 11. We analyzed the documents in this cluster and found that genetic disorder related diseases that are discussed in cluster 12 are related to neurological, paralytic and orthopedic concepts which are discussed in cluster 4, 9 and 11 respectively. Cluster 14 has only 1 document. It is identified that this document is related to 'metastasis'.

It is worth mentioning that we found that cluster 3 contains all the documents in which the concepts of diseases tagged by UMLS MetaMap is 'stroke'. However, further analysis shows that most of the documents are not related to cardiological or neurological 'stroke'. This is also reflected on the u-matrix that cluster 3 is not close to cluster 1 and 2. It confirmed that the proposed document presentation and weight scheme based on the concept similarity measure can effectively differentiate documents based on the concepts of diseases. On the other side, it also shows that the UMLS MetaMap might not accurately map all concepts to the corresponding phrases through the lexion.

Overall, the proposed document clustering and visualization framework works well on the representative data set used in this research. The clustering visualization based on the concepts of diseases can facilitate biomedical document retrieval based on diseases.

VII. Conclusion and Future Work

In this paper, a biomedical document clustering framework based on concepts of diseases is proposed. The concepts of diseases are identified by using UMLS MetaMap. Instead of using an existing ontology to generate concept representation, the concepts of diseases are represented by using vectors based on the Word2Vec model. By using the proposed vector presentation of the concepts of diseases, the proposed similarity measure shows that closely associated concepts of diseases have higher similarity scores than others. A proposed representation of documents that considers the local content and semantic similarity between the concepts within the documents is used. Self-Organizing Map is a clustering algorithm that provides a visualization which can aid in understanding the clusters and distribution of the clusters. The internal clustering validation metric - Davies-Bouldin index is used to evaluate the visualized clusters to identify the best partitions. The results show that the clustering occurs based on the concepts of similar nature, similar area and organs of the body, and concepts which are synonymous to one another. Nearby clusters are related in most cases, as well. This kind of visualization will help researchers explore related articles based on concepts of diseases.

Potential future work includes visualizing clusters of larger corpora by using a hierarchical clustering architecture, evaluating this visualization aid for the task of biomedical document search and extending this framework to biomedical document clustering based on concepts of symptoms and treatments.

References

[1] I. Yoo, X. Hu, and I.-Y. Song, "A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method," in *First International Workshop on Text Mining in Bioinformatics Proceedings*, November 2006, pp. 84–89.

[2] S. Logeswari and K. Premalatha, "Biomedical document clustering using ontology based concept weight," in *International Conference on Computer Communication and Informatics Proceedings*, Jan. 2013, pp. 1–4.

[3] I. Yoo and X. Hu, "A comprehensive comparison study of document clustering for a biomedical digital library medline," in *Joint Conference on Digital Library Proceedings*, June 2006, pp. 220–229.

[4] X. Zhang, L. Jing, X. Hu, M. Ng, and X. Zhou, "A comparative study of ontology based term similarity measure on pubmed document clustering," in *International Conference on Database Systems for Advanced Applications Proceedings*, 2007, pp. 115–126.

[5] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *International Conference on Neural Information Processing Systems*, 2013, pp. 3111–3119.

[6] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 13, pp. 1 – 6, 1998.

[7] Z. Wu and M. Palmer, "Verb semantics and lexical selection," in *the 32nd Annual Meeting of the Associations for Computational Linguistics Proceedings*, 1994, pp. 133–138.

[8] O. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity and natural language," *Journal of Artifical Intelligence Research*, vol. 11, pp. 95–130, 1999.

[9] R. Knappe, H. Bulskov, and T. Andreasen, "Perspectives on ontology-based querying," *International Journal of Intelligent Systems*, vol. 22, no. 7, pp. 739–761, 2007.

[10] J. Gu, W. Feng, J. Zeng, H. Mamitsuka, and S. Zhu, "Efficient semisupervised medline document clustering with mesh-semantic and global-content constraints," *IEEE Transactions on Cybernetics*, vol. 43, no. 4, pp. 1265–1276, August 2013.

[11] C. Gorg, H. Tipney, K. Verspoor, W. K. Baumgartner, K. B. Cohen, J. Stasko, and L. E. Hunter, "Visualization and language processing for supporting analysis across the biomedical literature," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems Proceedings*, 2010, pp. 420–429.

[12] *MetaMap - A Tool For Recognizing UMLS Concepts in Text*, https://metamap.nlm.nih.gov/.

[13] *Fact Sheet - UMLS Metathesaurus*, https://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html.

[14] *SNOMED CT*, https://www.nlm.nih.gov/healthit/snomedct/.

[15] S. Moen and T. S. S. Ananiadou, "Distributional semantics resources for biomedical text processing," 2013.

[16] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, "Self organization of a massive document collection," *IEEE transactions on neural networks*, vol. 11, no. 3, pp. 574–585, 2000.

[17] *Open Access Subset*, https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/.

[18] S. Zhu, J. Zeng, and H. Mamitsuka, "Enhancing medline document clustering by incorporating mesh semantic similarity," *Bioinformatics*, vol. 25, no. 15, pp. 1944–1951, 2009.

[19] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[20] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Transactions on neural networks*, vol. 11, no. 3, pp. 586–600, 2000.

[21] J. R. Millar, G. L. Peterson, and M. J. Mendenhall, "Document clustering and visualization with latent dirichlet allocation and self-organizing maps." in *FLAIRS Conference*, vol. 21, 2009, pp. 69–74.