Review Article

Healthc Inform Res. 2017 July;23(3):141-146.
https://doi.org/10.4258/hir.2017.23.3.141
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Text Mining in Biomedical Domain with Emphasis on Document Clustering

Vinaitheerthan Renganathan, MSc, MCA
Head of Institutional Research, Skyline University College, Sharjah, UAE

Objectives: With the exponential increase in the number of articles published every year in the biomedical domain, there is a need to build automated systems to extract unknown information from the articles published. Text mining techniques enable the extraction of unknown knowledge from unstructured documents. **Methods:** This paper reviews text mining processes in detail and the software tools available to carry out text mining. It also reviews the roles and applications of text mining in the biomedical domain. **Results:** Text mining processes, such as search and retrieval of documents, pre-processing of documents, natural language processing, methods for text clustering, and methods for text classification are described in detail. **Conclusions:** Text mining techniques can facilitate the mining of vast amounts of knowledge on a given topic from published biomedical research articles and draw meaningful conclusions that are not possible otherwise.

Keywords: Text Mining, Cluster Analysis, Classification, Natural Language Processing, Software

I. Introduction

Text mining [1,2] is the process of extracting new information on a particular topic from a set of documents. Text mining is useful where the data is in the form of text (document) which is unstructured and cannot be processed using traditional methods, such as data mining methods [3]. Text mining is different from normal search queries as it is also useful in discovering unknown information from a set of documents. Text mining is based on the natural language processing technique, which helps computers to understand and process human language [4].

Submitted: March 21, 2017
Revised: 1st, May 27, 2017; 2nd, July 14, 2017; 3rd, July 16, 2017
Accepted: July 17, 2017

Corresponding Author
Vinaitheerthan Renganathan, MSc, MCA
Head of Institutional Research, Skyline University College, P.O. Box 1797, University City, Sharjah, United Arab Emirates. Tel: +971-6-5441155, E-mail: vinairesearch@yahoo.com

II. Roles of Text Mining and Its Applications in Biomedical Field

Biomedical researchers have started using text mining techniques [1,5] due to the vast amount of unstructured information available in the biomedical domain in the form of research articles, case reports, Electronic Health Records (EHRs), and so forth. The following section highlights the applications of text mining in the biomedical domain.

1. Extraction of Knowledge from Biomedical Literature

The number of articles and papers published in the biomedical domain is increasing at a fast rate due to the expansion of online publishing. The total number of articles indexed in MEDLINE exceeded 23 million [6], and the number of citations reached more than 806 thousand. Thus, knowledge extraction from this vast collection of articles on a particular topic could be very time-consuming. Text mining techniques

can facilitate the extraction of unknown knowledge from the vast number of articles available. Some of the research carried out in this area has focused on the following applications. (1) In the field of cancer research, it offers a means to improve diagnosis, treatment, and prevention of cancer through mining of cancer literature [7]. (2) In pharmacology, it can be helpful to extract drug-drug interactions, protein interactions, and microbial interactions through mining of biomedical literature

2. Text Mining in Systematic Reviews

Systematic reviews [8] normally involve searching, screening, and synthesizing information from articles meeting the inclusion criteria and combing through the results to address the research problem. The searching, screening, and classification of articles requires enormous time and effort for the researchers involved in systematic reviews. Text mining offers tools for carrying out automatic searching, clustering, and classification of documents and information extraction during various stages of systematic reviews, such as searching, screening, and synthesis of information.

3. Text Mining in Information Extraction from Electronic Health Records

EHR systems store huge amounts of structured and unstructured information. Data mining methods are helpful in analyzing the structured part. Clinical texts [9], such as patient pathology reports, personal medical histories, and notes related to findings during examinations or procedures form the unstructured part of EHRs, and they can be analyzed using text mining techniques to explore hidden information [9]. The following are some of the applications of text mining in EHR systems. (1) EchoInfer [10], a text mining software tool, can be used to extract data pertaining to cardiovascular structures and functions from heterogeneously formatted echocardiographic data sources. (2) A text mining system using Bayesian networks can be used to mine narrative text from mammography reports to aid cancer diagnosis [11].

4. Biomarkers

Text mining techniques are useful for identifying disease-related biomarkers and associated genes from the literature [12]. Text mining is applied through the named entity recognition (NER) method, which is a technique for sub-task information extraction. It identifies entities and classifies them into various classes, such as gene, name of a person, organization, etc. The following are the some of the applications of text mining in the field of biomarkers. (1) MeinfoText [13] is a tool that provides knowledge about associations between gene methylation and cancer through the mining of large amounts of literature. (2) Whatizit [14] is a tool that identifies terms from a web source, such as PubMed abstracts, and then links them to the corresponding entries in bioinformatics databases.

5. Disease Surveillance

Web mining, a part text mining which helps us to detect disease outbreaks in disease surveillance systems. (1) Biocaster [15] is a web-based open-source and ontology-based text mining system for detecting and tracking the outbreak of diseases from web-based sources. (2) MedISys and PULS [16] are information retrieval and extraction systems used to analyse disease epidemics.

6. Other Areas

Some recent advances in biomedical text mining are in the areas of pharmacogenomics, toxicology, precision medicine, and drug repositioning [5]. Text mining can help identify named entities, such genes, proteins, drugs, and diseases and identify relations among them. Text mining can help the extraction of genotype and phenotype data for providing care in the field of precision medicine as well as the identification of relations among existing drugs and new diseases by mining the biomedical literature to reposition drugs.

III. Text Mining Software in Biomedical Domain

Currently there are several free and commercial software tools available to carry out text mining on various research databases. Table 1 lists the free and commercial software tools available for text mining in the field, and Table 2 compares the biomedical text mining software tools presented in Table 1.

IV. Text Mining Process

The following processes are involved in text mining: search and retrieval of document [24], creation of corpus of documents, pre-processing of documents [25], preparation of document matrix, clustering of documents [26], finding associations, preparation of word cloud, and processing the language part using natural language techniques [4,10]. Once the process is completed, the next level classifies the documents using a naïve Bayes classifier [1,12] or the support vector machine [1,12] method, or the decision tree

 https://doi.org/10.4258/hir.2017.23.3.141

Table 1. List of software tools available for text mining in the biomedical domain

Software	Description
Dnorm [17]	Uses an automated method to determine which diseases are mentioned in biomedical text and the task of disease normalization
National Center for Text Mining [18]	Several text mining software tools available, such as Part Speech Taggers, Parsers, Named Entity Recognition tools
PESCADOR [19]	Web-based tool that assists text-mining of bio interactions extracted from PubMed queries
BioClass [20]	Parameterizes, trains, and tests various text classifiers to determine which technique performs better according to the document corpus.
BioLMiner [21]	Automatically extracts useful information from biological literature, such as gene mentions, normalized gene mentions, interaction articles, and protein-protein interaction pairs
OntoGene [22]	Detects entities and relationships from selected categories, such as proteins, genes, drugs, diseases, and chemicals
MedMiner [23]	Internet text mining tool for biomedical information, with application to gene expression profiling
IBM SPSS Text Analytics	Commercial general text mining package for mining unstructured text data; organizes key concepts, groups into categories from the web, comment fields, books, and other text sources
SAS Text Miner	Commercial general text mining package; analyses text data from the web, comment fields, books, and other text sources
RapidMiner	Commercial general text mining package; extracts information from publicly available data sources
R software	Open-source software user packages

Table 2. Comparison of biomedical text mining software tools

Software	Search & retrieval and pre-processing of text	Entity recognition	Finding similarities	Classification and categorization	Natural language processing
Dnorm [17]	Yes	Yes	Yes	-	Yes
National Centre for Text Mining [18]	Yes	Yes	Yes	Yes	Yes
PESCADOR [19]	Yes	Yes		-	-
BioClass [20]	Yes	-	Yes	Yes	-
BioLMiner [21]	Yes	Yes	-	-	Yes
OntoGene [22]	Yes	Yes	Yes	Yes	Yes
MedMiner [23]	Yes	Yes	-	-	-
IBM SPSS Text Analytics	Yes	-	Yes	Yes	Yes
SAS Text Miner	Yes	-	Yes	Yes	Yes
RapidMiner	Yes	-	Yes	Yes	Yes
R software	Yes	-	Yes	Yes	Yes

method. The vector space model [1,12] concept takes the centre stage in the text mining process, in which the documents are represented as n-dimensional vectors of terms.

1. Search and Retrieval of Documents

The first step involved in text mining is to search and retrieve documents using the information retrieval process [24], which automatically retrieves documents based on the information need of the user from a large collection of documents, which is usually web-based.

2. Pre-processing of Documents

The pre-processing of documents [25] involves the steps of stop word removal and stemming.

1) Stop word removal

Stop words, such as 'the' and 'a' are removed. There are number of methods available to remove stop words. The classic method removes pre-defined stop words, and Zip's law [25] method removes words with high Term Frequency-Inverse Document Frequency (TF-IDF) value and words appearing only once in the document.

2) Stemming

After the removal of stop words, the next step involves 'stemming', which helps us to use only the roots of terms. For example, the terms 'analyze', 'analytical', and 'analyzing' are represented by the root term 'analysis'.

3. Term Document Matrix

Once the pre-processing has been completed, the next step is to prepare the term document matrix (TDM), in which terms are represented by rows, and documents are represented by columns. TF-IDF [27] are important measures in the text mining process.

4. Natural Language Processing

Natural language processing [4,10] is a tool that is used to analyze the language part of text documents through automated systems. Basically, language processing is divided into the processing of words (morphology), their different forms (lexicon), sentence structures (syntax), and sentence meanings (semantics), conference analysis, and the relationships between sentences (discourse). Natural language processing systems widely use statistical techniques to remove the ambiguity present in the processing of texts. It is used to automatically process text using a probabilistic approach and to carry out tasks such as segmentation of sentences into words, named entity recognition, parts of speech tagging, conference resolution [4,28], etc.

5. Methods for Text Clustering

Documents are grouped according to their document vector, and each cluster is denoted by the document vector name. Clustering [26] of documents can be carried out using techniques such as hierarchical clustering and portioning clustering (K-means clustering).

1) Hierarchical clustering technique

Hierarchical clustering [29] creates a hierarchy of clusters of documents using a top-down (divisive) or bottom-up approach (agglomerative). In the agglomerative method, clustering starts with each document as a single cluster, and in the next step, each cluster is combined with another cluster to form a new cluster based on the closest distance or similarity between the two clusters. This process is repeated until a single cluster is formed. In the divisive method, initially all the documents are combined to form a single cluster, and the cluster is divided into two sub-clusters which have maximum distance or dissimilarity between them. This process will continue until each document forms its own cluster. In hierarchical clustering, previous knowledge about the number of clusters is not required. The outcome of hierarchical clustering is a graphical representation called a dendrogram, in which the documents are represented in a hierarchical tree structure representing the documents as its branches.

2) Partitioning (K-means) clustering

K-means clustering [30] starts with a predefined number of clusters of documents, for instance, k clusters. Documents will be relocated to different clusters based on the nearness to the cluster centroid (mean). At each partition, the cluster centroid is recalculated recursively after the relocation of documents based on nearness to the cluster centroid. This process is repeated until there is no change in the cluster means or centroid due to the relocation of documents. Generally the K-means clustering algorithm is faster than the hierarchical clustering algorithm.

3) Similarity measures

Clustering process efficiency depends on the choice of similarity measures, such as cosine, Euclidean, Manhattan, and Mahalanobis [28]. Cosine measure is the simplest and easiest method for clustering documents. Cosine measure calculates the normalized dot products of two document vectors. The cosine values range from 0 to 1, and when the two documents do not share any words, the cosine value is 0.

6. Methods for Text Classification

Documents can be automatically classified into specific categories using classifier algorithm such as naïve Bayes, support vector machine, and decision tree.

1) Naïve Bayes

The naïve Bayes classifier can be used to classify documents based on a probabilistic concept by which terms in each document are assigned specific probabilities based on their frequency in the document corpus. During a supervised training process, the naïve Bayes classifier assigns documents in the training document set to predefined categories based

https://doi.org/10.4258/hir.2017.23.3.141

on set of terms in the document whose probability of occurrence is maximum in the predefined category in relation to other category. This training document set can be used to classify a new set of documents based on the posterior probabilities that the set of documents will have terms with similar probabilities and can be classified to a predefined category.

2) Support vector machine

Support vector machine is used to find separators to separate two document categories. Documents are assumed to take the form of linear space, and the separator will be a hyper plane (a subspace in a two-dimensional space) that separates the two categories of documents.

3) Decision tree

The decision tree method represents category conditions as nodes and documents categories as leaves. The decision tree method works recursively and classifies documents into categories based on conditions.

4) Evaluation of classification

The evaluation of the classification process is carried out using recall, precision, and F measures:

$$\text{Precision class} = TP / (TP + FP),$$

$$\text{Recall class} = TP / (TP + FN),$$

where

TP (true positive): number of correctly classified instances to a class;

FP (false positive): number of falsely classified instances, as belonging to a class;

FN (false negative): number of instances belonging to a class, not correctly classified.

The F-measure is the harmonic mean of precision and recall. It is calculated by

$$F = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall}).$$

V. Conclusion

This paper provided a comprehensive overview of text mining methods. The paper discussed the roles of text mining in biomedical applications and presented software available to carry out text mining. The paper also presented an overview of techniques to find similarities between studies on a given topic from available research articles.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

1. Ananiadou S, McNaught J. Text mining for biology and biomedicine. London: Artech House; 2006.
2. Hearst MA. Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th Conference on Computational Linguistics; 1992 Aug 23-28; Nantes, France. p. 539-45.
3. Dorre J, Gerstl P, Seiffert R. Text mining: finding nuggets in mountains of textual data. Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 1999 Aug 15-18; San Diego, CA. p. 398-401.
4. Chowdhury GG. Natural language processing. Annu Rev Inf Sci Technol 2003;37(1):51-89.
5. Gonzalez GH, Tahsin T, Goodale BC, Greene AC, Greene CS. Recent advances and emerging applications in text and data mining for biomedical discovery. Brief Bioinform 2016;17(1):33-42.
6. National Institute of Health. Fact sheet MEDLINE [Internet]. Bethesda (MD): National Institutes of Health; c2017 [cited at 2017 Jul 15]. Available from: https://www.nlm.nih.gov/pubs/factsheets/medline.html.
7. Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, et al. Biomedical text mining and its applications in cancer research. J Biomed Inform 2013;46(2):200-11.
8. Tari L, Anwar S, Liang S, Cai J, Baral C. Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. Bioinformatics 2010;26(18):i547-53.
9. Zhou D, He Y. Extracting interactions between proteins from the literature. J Biomed Inform 2008;41(2):393-407.
10. Lim KM, Li C, Chng KR, Nagarajan N. @MInter: automated text-mining of microbial interactions. Bioinformatics 2016;32(19):2981-7.
11. Higgins JP, Green S. Cochrane handbook for systematic reviews of interventions. 1st ed. Chichester: Wiley-Blackwell; 2008.
12. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle

JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008;35(8):128-44.
13. Nath C, Albaghdadi MS, Jonnalagadda SR. A natural language processing tool for large-scale data extraction from echocardiography reports. PLoS One 2016;11(4): e0153749.
14. Bozkurt S, Gimenez F, Burnside ES, Gulkesen KH, Rubin DL. Using automatically extracted information from mammography reports for decision-support. J Biomed Inform 2016;62:224-31.
15. Bravo A, Cases M, Queralt-Rosinach N, Sanz F, Furlong LI. A knowledge-driven approach to extract disease-related biomarkers from the literature. BioMed Res Int 2014;2014:253128.
16. Fang YC, Huang HC, Juan HF. MeInfoText: associated gene methylation and cancer information from text mining. BMC Bioinformatics 2008;9:22.
17. Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A. Text processing through Web services: calling Whatizit. Bioinformatics 2008;24(2):296-8.
18. Collier N, Doan S, Kawazoe A, Goodwin RM, Conway M, Tateno Y, et al. BioCaster: detecting public health rumors with a Web-based text mining system. Bioinformatics 2008;24(24):2940-1.
19. Steinberger R, Fuart F, van der Goot E, Best C, von Etter P, Yangarber R. Text mining from the web for medical intelligence. In: Francoise FS, editor. Mining massive data sets for security. Amsterdam: IOS Press; 2008. p. 295-300.
20. National Institutes of Health. DNorm [Internet]. Bethesda (MD): National Institutes of Health; c2016 [cited at 2017 Jul 15]. Available from: https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/dnorm/.
21. National Centre for Text Mining. NaCTeM software tools [Internet]. Manchester: National Centre for Text Mining; c2016 [cited at 2017 Jul 15]. Available from: http://www.nactem.ac.uk/software.php.
22. Romero R, Vieira AS, Iglesias EL, Borrajo L. BioClass: a tool for biomedical text classification. Proceedings of 8th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2014); 2014 Jun 4-6; Salamanca, Spain. p. 243-51.
23. Barbosa-Silva A, Fontaine JF, Donnard ER, Stussi F, Ortega JM, Andrade-Navarro MA. PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries. BMC Bioinformatics 2011;12:435.
24. Chen Y, Liu F, Manderick B. BioLMiner system: interaction normalization task and interaction pair task in the BioCreative II.5 challenge. IEEE/ACM Trans Comput Biol Bioinform 2010;7(3):428-41.
25. Rinaldi F, Clematide S, Marques H, Ellendorff T, Romacker M, Rodriguez-Esteban R. OntoGene web services for biomedical text mining. BMC Bioinformatic. 2014;15 Suppl 14:S6.
26. Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. Biotechniques 1999;27(6):1210-4, 1216-7.
27. Salton G, McGill MJ. Introduction to modern information retrieval. New York (NY): McGraw-Hill; 1986.
28. Manning CD, Schutze H. Foundations of statistical natural language processing. Cambridge (MA): MIT Press; 1999.
29. Vijayarani S, Ilamathi MJ, Nithya M. Preprocessing techniques for text mining-an overview. Int J Comput Sci Commun Netw 2015;5(1):7-16.
30. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manag 1988;24(5): 513-23.

 https://doi.org/10.4258/hir.2017.23.3.141