International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 8 Issue 05 Series. IV || May 2019 || PP 28-33

Application of text data mining in biomedical research

Le Anh Vu[1,*], Phan Thi Cam Quyen[2], Nguyen Thuy Huong[3]
[1,3](*Faculty of Chemistry Engineering, Ho Chi Minh City University of Technology, Vietnam)*
[2](*Department of Biotechnology, Kien Giang Seed Research Center, Vietnam)*
Corresponding Author: Le Anh Vu

Abstract: *Data mining of available biomedical data and information has greatly boosted knowledge discovery in the computational biology era. Data mining refers to a bioinformatics approach that combines biological concepts with computer tools or statistical methods that are mainly used to discover biological processes and relevant factors. This review explicates text data mining and their applications to biomedical knowledge discovery. Also discussed are the limitations and future perspectives of this approach.*
Keywords - *bioinformatics, biomedical knowledge, computational biology, data mining, databases.*

Date of Submission: 20-05-2019 Date of acceptance: 03-06-2019

I. Introduction

Currently, there have been increasing in researches to identify biological processes related to human diseases. This trend creates a series of biological data, from gene sequences to protein structures and interactions at molecular-level. Therefore, we are in a period when the ability to create data has exceeded the ability to exploit and analyze data [1]. Recently, the focus of biomedical studies has shifted from data generation to knowledge discovery, or more specifically biomedical data mining, as we enter the new field of deep learning of scientific data. Development is no longer focused on allowing researchers to create data quickly, but in a way that converts them into useful knowledge through data mining. Therefore, efficient biomedical data mining will be important in increasing understanding of pathogenesis mechanisms, aiming to discover new therapeutic treatments as well as new drugs. However, in order to transfer large amounts of biomedical data into useful insights, researchers need to overcome difficulties such as handling data that are noisy or incomplete (such as false positives/negatives data of protein interactions), handle intensive computing tasks (such as indexing, searching and exploiting large-scale graphs), integrating different data sources (such as linking genetic and protein data with clinical data) and biomedical data extraction within the ethical framework and privacy protection [1]. All of these problems pose challenges that need to be addressed for data mining researchers in the computational biology era.

In this review, we focus on the approach to text data mining with the potential of specific applications in the field of biomedical research. In addition, the limitations of the approach and future prospects are also discussed.

II. Biomedical Data Mining

2.1 Introduction to text data mining

The field of biomedical data mining has been given much attention in recent years due to the huge amount of textual data created in various forms such as medical records, health care and scientific research data. Therefore, those data are an invaluable source of information and knowledge. Text data is often in unstructured form, which is considered one of the simplest data types and can be created in most cases. Unstructured documents are easily handled and perceived by humans, but they are much more confusing to computers. Therefore, it is necessary to design algorithms to effectively handle these documents in different contexts. Text mining involves traditional data mining and knowledge discovery, with some unique characteristics.

Knowledge discovery (KD) is a method used to extract legal and useful information from data. Meanwhile, text mining (TM) is a method of applying specific algorithms to extract specific characteristics from data [2]. Based on the above definition, KD refers to the whole process of discovery of useful information from data, while text mining refers to a specific step in this process. Data can be structured like databases, but may not have structure as in a simple text file. Text mining includes a set of related methods and algorithms for text analysis, including information retrieval, preprocessing, natural language processing, information extraction, text summarization, classification, clustering and machine learning are applied in many fields of biomedical sciences. Summary information about these terms is mentioned in Table 1.

Table 1: Methods and algorithms are often used in text data mining.

Method/Algorithm	Aim
Information retrieval	Retrieving valuable information from unstructured text
Preprocessing	Convert raw data into a more understandable format
Natural language processing	Help computers understand human language
Information extraction	Extract information from structured data
Text summarization	Reduce the text size while keeping the main points and the overall meaning of the text
Classification	Assign documents by a set of predefined topics according to their content
Clustering	Find internal structures in the text and organize them into small groups for further research and analysis
Machine learning	A set of statistical techniques to identify aspects of text.

2.1.1 Information retrieval

Information retrieval is the activity of finding information resources (usually documents) from a collection of unstructured datasets that meet searching conditions [3]. Therefore, the retrieval of information mainly focuses on facilitating access to information rather than analyzing information and finding hidden features, which are the main purpose of text mining.

2.2.2 Preprocessing

Preprocessing is one of the main steps in many text mining algorithms. Many traditional text classification processes often include preprocessing stages, characteristic extraction, characteristic selection and classification stages. Although the steps for extracting characteristics, characteristics selection and classification algorithms have been confirmed that have a significant impact on the classification process, the preprocessing stage can significantly affect the success of data mining [4]. Preprocessing steps usually include methods such as tokenization, filtering, lemmatization and stemming (Table 2).

Table 2: Methods used in preprocessing documents.

Method	Aim
Tokenization	Tokenization is a technique for dividing the sequence of strings into smaller parts such as words, keywords, phrases, symbols and other elements. These smaller sections are called tokens. During tokenization, some characters like punctuation are removed. The tokens then become inputs for another process such as parsing and text extraction.
Filtering	A technique used to remove or retain previously selected words.
Lemmatization	Lemmatization is an algorithm used in analyzing the morphology of words. Therefore, a detailed dictionary is required that the algorithm can look at to link the form back to its lemma.
Stemming	The stemming algorithm works by cutting off the end or the beginning of a word, taking into account a list of common prefixes and suffixes that can be found in a deformed word.

2.2.3 Natural language processing (NLP)

Natural language processing is an area that combines the knowledge of computer science, artificial intelligence and linguistics to understand natural language using computer [5]. Many text mining algorithms commonly use NLP techniques, such as part of speech recognition, parsing, and other types of language analysis [6].

2.2.4 Information extraction

Information extraction is a technique that automatically extracts information or events from unstructured or semi-structured documents [7]. It is often used as a starting point for other text mining algorithms. For example, name entity recognition (NER) and their relationships from text can provide useful semantic information.

2.2.5 Text summarization

Many text mining applications need to summarize text documents to get a brief overview of a large document or a set of documents on a specific topic [8]. In general, there are two types of summary techniques: practical summaries in which the summary includes information units extracted from the original text and abstract summaries in which a summary can contain information that will not be exist in the original document [9].

2.2.6 Text classification

Classification is the assignment of common text documents to a set of predefined topics according to their content. The result is a set of text documents, with a specific theme for each document. Automatic text classification has been applied in many contexts from automatic indexing to spam filtering, classification of web

pages by category and detection of text categories [10]. Today, this is a hot topic in the field of machine learning.

2.2.7 Clustering

Clustering is one of the most popular data mining algorithms used in classification, visualization and organization of documents [11]. Clustering is the task of finding similar groups of documents in a document collection. Similarity is calculated using the function function. Clustered text can be at different levels of detail in which clusters can be documents, paragraphs, sentences or terms [12].

2.2.8 Machine learning

Machine learning in the context of text mining is a set of statistical techniques to identify parts of words, entities and other aspects of text. Techniques can be expressed in the form of a model that is then applied to another document, also known as supervised learning. It can also be a set of algorithms that operate on large data sets to extract meaning, called unsupervised learning (Table 3) [13].

Table 3: Two techniques commonly used in machine learning.

Technique	Description
Supervised learning	Supervised learning is the algorithm that predicts output (outcome) of a new data (new input) based on previously known pairs (input, outcomes).
Unsupervised learning	In this algorithm, the computer does not know the output but only the input data. The unsupervised learning algorithm will rely on the structure of the data to perform a certain task, such as clustering or reducing the dimension of data to facilitate storage and computation.

2.2 Application of text mining in biomedical research

One of the areas where text mining is heavily used is biomedical sciences. Biomedical data is growing exponentially [14], which makes it difficult for scientists to assimilate data and keep up with research results. To overcome this information overload, the techniques of text mining along with machine learning algorithms have been widely used in this field. Currently, many software have been developed to analyze text data (Table 4). Text mining techniques have been used in many areas such as protein structure prediction, gene clustering and clinical diagnosis, etc. In this section, we describe some common applications of text mining in biomedical field includes Ontology definition, entity identification, information extraction, correlation extraction, text summary, and question answering.

Table 4: Some tools/software are used in text data mining.

Tool/Software	Description	Internet access
Google Scholar	Search engine	http://scholar.google.com/
GoPubMed	PubMed engine	http://www.gopubmed.org/
Textpresso	Full-text search	http://www.textpresso.org/
BioRAT	Full-text search	http://bioinf.cs.ucl.ac.uk/biorat/
ABNER	Entity taggers	http://pages.cs.wisc.edu/~bsettles/abner/
iHOP	Entity recognition	http://www.ihop-net.org/UniPub/iHOP/
GeneWays	Pathway extraction	http://geneways.genomecenter.columbia.edu/

2.2.1 Ontology identification

Ontology is defined as common terms used to describe and represent a specific field of knowledge. Under this definition, ontology has the following characteristics: 1) Ontology is a specific domain, that is, it is used to describe and represent a field of knowledge such as education and medicine; 2) Ontology includes terms and relationships between these terms. The terms are often called classes / concepts and relationships are called attributes [15]. Currently there are many ontology biomedical data stored in databases such as the Open Biomedical Ontologies (OBO) and the National Center for Biomedical Ontology (NCBO). In addition, there are many other ontology databases that focus more on sub-fields of biomedical sciences. For example, the Pharmacogenomics Knowledge database contains clinical information including instructions on dosage and drug labels, the relationship between drugs - genes and genotypic relationships - phenotype [9]. Corresponding ontologies and databases are widely used by different text mining techniques such as information mining and clustering.

2.2.2 Information extraction

As mentioned in the previous section, information extraction is the task of extracting structured information from unstructured text automatically. In the biomedical field, unstructured text includes most scientific articles in scientific databases and clinical information found in clinical information systems.

Information extraction is often considered a preprocessing step in other biomedical text mining applications such as question answering [16], hypotheses generation [17] and text summarization [18].

2.2.3. Named entity recognition (NER)

Named entity identification is the task of exploiting information used to locate and classify biomedical entities into categories such as protein names, gene names or diseases. Ontologies can be used to provide semantic representation for extracted entities. It is important for NER systems to be of high quality and work well when analyzing a large number of documents. Accuracy, recovery and F point are typical assessment methods used in NER systems [19]. NER methods are often grouped into a number of different approaches (Table 5).

Table 5: Common approaches in identifying entity names.

Approach	Description
Dictionary based	Use a dictionary of biomedical terms to locate the entity mentioned in the text. It determines whether a word or phrase in the text matches some biomedical entities in the dictionary.
Principle based	Identify specific characteristic rules of biomedical entities.
Statitical based	Use some machine learning methods to identify biomedical entities.

2.2.4 Correlation extraction

Extraction of correlation in the biomedical field include determining the relationship between biomedical entities. Given two entities, this technique aims to locate the appearance of a specific relationship between them. The correlation between entities is usually a duo, however, it may include more than two entities. For example, in the genomic field, the focus is mainly on the correlation between genes and proteins, such as protein-protein or gene-disease relationships. Correlation extraction encountered similar difficulties as NER, such as creating high quality annotation data to train and evaluate the performance of correlation extraction systems [20]. There are many different approaches to extract correlations between biomedical entities (Table 6).

Table 6: Approaches used in correlation extraction.

Approach	Description
Simultaneous access	If entities are mentioned together often, they may be highly relevant in some way. But this approach cannot recognize the type of correlation when only using statistical methods.
Principle based	Manually identified by experts in the field or automatically using machine learning techniques from a corpus of notes.
Classification based	Use supervised machine learning algorithms to detect and discover different types of relationships.

2.2.5 Text summarization

Summary is the task of identifying important aspects of one or more documents and expressing them automatically [21]. Recently it has received attention for the development of unstructured information in the field of biomedical. Biomedical summary text is often application-oriented and can be applied for different purposes. Depending on the purpose, a series of summaries can be created such as a single document summary aimed at the content of individual documents and a summary of multiple documents that contain the information content of many viewed documents review Therefore, assessing the performance of summary methods is a challenge in the field of text mining. The reason is that decisions are often subjective and at the same time manual reviews of the summaries are often time-consuming to implement. Recently, there is a popular automated evaluation technique for abstracts that have been developed called ROUGE (Recall-Oriented Understudy for Gisting Evaluation). ROUGE measures the quality of an automatically generated summary by comparing it to the ideal man-made summaries. This technique is scored by counting words that overlap between a computer-generated summary and an ideal human-generated summary [21].

2.2.6 Question answering

Question answering is defined as the process of creating accurate answers to questions posed by humans in natural language [21]. In order to create accurate answers, the systems that answer questions use widely natural language processing techniques. The main steps of the question answering system can be summarized as follows: a) The system of receiving natural language texts as input; b) Using language analysis and question classification algorithms, the system determines the type of questions to be asked and the answers to be created; c) It then creates a query and passes it to the data processing stage; d) During the data processing phase, the system will send queries to the search engine, retrieve the retrieved data and extract the relevant paragraphs as a potential answer and send them to the processing stage; e) In the answer processing phase, the system analyzes potential answers and ranks them according to the level of the expected answer set in the

question processing step; f) The top rated answer is chosen as the output of the question answering system. Recent biomedical answering systems have begun to use and incorporate semantic knowledge throughout their processing steps to create more accurate answers. These systems based on semantic knowledge use different semantic components such as semantic metadata in databases, Ontology and semantic relationships to provide answers [21].

2.3 Database Standards

In order to effectively manage the data sets created in biomedical studies, standards are required for initiating, evaluating quality, storing and exchanging data. Standardization is necessary to ensure the accuracy and reproducibility of data. Specifically, the sharing and exchange of research data requires standard formats at all levels, from raw data to processed data and ultimately results, as well as detailed information about referrals, numbers and protocols used in sample processing and data analysis. Standard formats for data sets and results of computational analysis with converters and other software tools that interact with these standards have been proposed [22]. The need for robust and easily accessible data storage facilities has been widely recognized for the benefit of proven data sharing in research areas such as genomic, transcriptomic, proteomic and metabolomic [23]. Currently, there are many databases integrating data on many laboratories, allowing data analysis with new improved tools and algorithms and avoiding rediscovering known results. Major databases include PubMed Central, UniProt, PDB Data Bank (https://www.rcsb.org/pdb/) and many smaller-scale storage databases (Table 7) [24].

Table 7: Some popular biomedical text databases.

Database	Description	Internet access
PubMed Central	Full text of scientific reports	http://www.pubmedcentral.nih.gov/
HighWire Press	Full text of scientific reports	http://highwire.stanford.edu/
UniProt	Protein information	http://www.uniprot.org/
PDB	Protein information	
InterPro	Information about protein domain	http://www.ebi.ac.uk/interpro/
KEGG	Kyoto Encyclopedia of Genes and Genomes	https://www.genome.jp/kegg/
DEG	Database of Essential Genes	www.essentialgene.org/
TTD	Theraeutic Target Database	http://xin.cz3.nus.edu.sg/group/ttd/ttd.asp
MetaCyc	Database of Metabolic Pathways	https://metacyc.org/
LIGAND	Chemical Compound Database	http://www.genome.ad.jp/ligand/
OMIM	Online Mendelian Inheritance in Man	https://www.omim.org/
HPID	Human Protein Interaction Database	http://wilab.inha.ac.kr/hpid/
BIND	Biomolecular Interaction Database	http://bind.ca
PharmGKB	National Institutes Of Health Pharmacogenetics Research Database	http://www.pharmgkb.org/
STRING	Database of Protein - Protein Interaction Network	https://string-db.org/

2.4 Limitation

The limit of text mining arises from the complexity of natural language. Natural language often encounters ambiguous situations. Ambiguity means the ability to be understood in two or more ways. Ambiguity gives natural language flexibility and usability, and therefore, it cannot be completely removed from natural language. A word can have many meanings and a phrase or sentence can be interpreted in different ways, so they can have many different meanings [25]. Another difficulty in exploiting biomedical data is to identify the entity named (NER). This is due to the large number of semantic-related entities in the biomedical field and is rapidly increasing with new discoveries made in this area. The continuous growth of these entities' volumes is problematic for NER systems, because they depend on the dictionary of terms that are difficult to accomplish due to continuous progress in scientific studies. In addition, in the field of biomedical, the same concept can have many different names (synonyms). For example, "heart attack" and "myocardial infarction" have the same concept and NER systems need to recognize the same concept regardless of the different expression. Finally, the use of synonyms and the initials of other words (acronym) that are very common in biomedical documents have made it impossible to define the concepts due to the complex of these terms [9].

III. Conclusion and Future Perspectives

Text mining is essential for scientific research due to the vast amount of scientific documents updated annually. Although this growth allowed researchers to easily access more information, it also difficult for them to identify suitable data to meet research purposes. Therefore, the processing and exploitation of this huge amount of text is now very interested by researchers. In this article, we have provided an overview of data mining, text mining techniques and their application in the field of biomedical. Although these techniques are increasingly being adopted, they need to be explained carefully. The knowledge and model discovered by computers need to be authenticated either clinically or practically, just like any human-generated knowledge. Errors and inaccurate descriptions can spread quickly through electronic means, especially when large databases

and powerful computational techniques are involved. However, these methods of data management and exploitation are changing the way of discovering, organizing, applying and disseminating new knowledge. With the increasing speed of computer hardware, the popularity of the Internet, abundance of biomedical data and advances in bioinformatics research, text data mining will continue to create, governing Reasoning and extracting biomedical knowledge effectively, allowing researchers to better understand complex biological processes and support solving human health problems.

Acknowledgements

This paper is funded by Ho Chi Minh City University of Technology – VNU-HCM, under grant number TNCS-KTHH-2017-12.

References

[1]. F. Wang, X. Li, J.T.L. Wang and S. Ng, Guest editorial: special section on biological data mining and its applications in healthcare. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14(3), 2017, 501-502.
[2]. U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, Knowledge discovery and data mining: towards a unifying framework, Proc KDD-96: Second International Conference on Knowledge Discovery & Data Mining Menlo Park, CA: AAAI Press, 1996, 82–88.
[3]. C.D. Manning, P. Raghavan and H. Schütze, Introduction to information retrieval (Cambridge University Press, Cambridge, 2008).
[4]. A.K. Uysal and S. Gunal, The impact of preprocessing on text classifcation, Information Processing & Management, 50(1), 2014, 104–112.
[5]. E.D. Liddy, Natural language processing, Encyclopedia of Library and Information Science, 2nd Ed (Marcel Decker Inc, New York, 2001).
[6]. A. Kao and S.R. Poteet, Natural language processing and text mining (Springer, New York, 2007).
[7]. S. Sarawagi, Information extraction, Foundations and Trends® in Databases, 1(3), 2008, 261–377.
[8]. A. Hotho, A. Nürnberger and G. Paass, A brief survey of text mining, LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, 20, 2005, 19-62.
[9]. M. Allahyari, S.A. Pouriyeh, M. Assefi, S. Safaei, E.D. Trippe, J.B. Gutierrez and K.J. Kochut, Text summarization techniques: a brief survey, International Journal of Advanced Computer Science and Applications (IJACSA,) 8(10), 2017, 397-405.
[10]. M. Allahyari, S.A. Pouriyeh, M. Assefi, S. Safaei, E.D. Trippe, J.B. Gutierrez and K.J. Kochut, A brief survey of text mining: classification, clustering and extraction techniques, CoRR, abs/1707.02919, 2017.
[11]. R. Bekkerman, R. El-Yaniv, N. Tishby and Y. Winter, On feature distributional clustering for text categorization, Proceedings Of The 24th Annual International ACM Sigir Conference on Research and Development in Information Retrieval, ACM, 2001, 146–153.
[12]. L. Kaufman and P.J. Rousseeuw, Finding groups in data: an introduction to cluster analysis (John Wiley & Sons, New Jersey, 2009).
[13]. F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys (CSUR), 34(1), 2002, 1–47.
[14]. K.B. Cohen and L. Hunter, Getting started in text mining, PLOS Computational Biology, 4(1), 2008, e20.
[15]. L. Yu, A developer's guide to the semantic web (Springer, New York, 2011).
[16]. S.J. Athenikos and H. Han, Biomedical question answering: a survey, Computer Methods and Programs in Biomedicine, 99(1), 2010, 1–24.
[17]. A.M.L. Liekens, J. De Knijf, W. Daelemans, B. Goethals, P. De Rijk and J. Del-Favero, BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation, Genome Biology, 12(6), 2011, R57.
[18]. W. Hersh, Information retrieval: a health and biomedical perspective (Springer, New York, 2008).
[19]. U. Leser and J. Hakenberg, What makes a gene name? named entity recognition in the biomedical literature, Briefings in Bioinformatics, 6(4), 2005, 357-369.
[20]. S. Ananiadou, S. Pyysalo, J. Tsujii, D.B. Kell, Event extraction for systems biology by text mining the literature, Trends in Biotechnology, 28(7), 2010, 381–390.
[21]. C.C. Aggarwal and C.X. Zhai, Mining text data (Springer, New York, 2012).
[22]. M. Eisenacher, L. Martens, T. Hardt, M. Kohl, H. Barsnes, K. Helsens, et al., Getting a grip on proteomics data - proteomics data collection (ProDaC), Proteomics, 9, 2009, 3928–3933
[23]. J.T. Prince, M.W. Carlson, R. Wang, P. Lu and E.M. Marcotte, The need for a public proteomics repository, Nature Biotechnology, 22, 2004, 471–472.
[24]. J.A. Mead, L. Bianco and C. Bessant, Recent developments in public proteomic ms repositories and pipelines. Proteomics, 9, 2009, 861–881.
[25]. Y. Yang, S.J. Adelstein and A.I. Kassis, Target discovery from data mining approaches, Drug Discovery Today, 14(3-4), 2009, 147-154.

Le Anh Vu et al. "–Application of text data mining in biomedical research" International Journal of Engineering Science Invention (IJESI), Vol. 08, No. 05, 2019, PP 28-33