CSCanada
Cross-Cultural Communication
Vol. 13, No. 9, 2017, pp. 1-9
DOI:10.3968/10006

ISSN 1712-8358[Print]
ISSN 1923-6700[Online]
www.cscanada.net
www.cscanada.org

Knowledge Mapping Analysis on Text Mining Research of Medicine Related Fields in Different Regions

GOU Mengye[a]; ZHAO Wenlong[a],*

[a]College of Medical Informatics, Chongqing Medical University, Chongqing, China.
*Corresponding author.

Supported by the National Social Science Programming of China (13BTQ004); Chongqing Science and Technology Commission, China (cstc2015shmszx10004).

Received 24 June 2017; accepted 14 August 2017
Published online 26 September 2017

Abstract
In order to trace the trend of text mining research in medicine related fields through the massive literature, we analyzed the bibliographical reference data of relevant literature in the WOS database with methods of bibliometric and knowledge mapping. We concluded the research state from aspects of time sequence, core authors and institutions, regional and disciplinary distribution; and summarized the research hot points and frontiers through knowledge mapping analysis by using assistant tool CitespaceIII. Our analysis indicates that text mining research in medicine related fields appears a steady-state growth trend and state of multidisciplinary integration; and text mining technology has been widely applied to biomedical field such as named entity recognition task, construction and automatic annotation of gene or protein relating corpus, and biomedical event extraction based on various text mining tools. Besides, the research in recent years turns to the EHR information extraction and knowledge discovery, drug knowledge mining and social media mining, etc. In conclusion, it's worth applying text mining technology to explore medical information, especially clinical information or other aspects more extensively and thoroughly.

Key words: Text mining; Text analysis; Knowledge mapping; Medical information; Biomedical information; Health information

Gou, M. Y., & Zhao, W. L. (2017). Knowledge Mapping Analysis on Text Mining Research of Medicine Related Fields in Different Regions. *Cross-Cultural Communication, 13*(9), 1-9. Available from: http//www.cscanada.net/index.php/ccc/article/view/10006
DOI: http://dx.doi.org/10.3968/10006

INTRODUCTION
With the arrival of the information explosion era, the literature and data documents in the medical field are growing rapidly. These documents are massive in amount, and scattered in distribution. Therefore it's necessary to develop these medical or biomedical documents, trial data, online or media health information and electric health record text into more structured and standardized valuable information, to apply more accurately to clinic or subject knowledge discovery, intelligent decision support and data analysis through methods of text mining or text analysis.

Text mining is also called text analysis, and a well known definition of text mining is "the discovery by computer of new, previously unknown information, by automatically extracting and relating information from different written resources, to reveal otherwise 'hidden' meanings" (Hearst, 1999). The main steps of text mining include information retrieval, named entity recognition, information extraction, and knowledge discovery (Fleuren, 2015).

1. DATA AND METHOD

1.1 Data Resources
Relevant documents were retrieved from the Web of Science (WOS) of core collection database. In order to get more accurate retrieval results, we formulated the search strategies as follows: (theme= "medical" OR "health" OR "biomedical") AND (theme= "text mining"

OR title= "text analysis"), time range from 1998 to 2017 for there were few articles been published before 1998. We set the type of documents as articles, and the language is limited to English. We downloaded the bibliographical reference data and stored them in TXT format.

1.2 Methods

Methods of bibliometrics and knowledge mapping analysis were used in this paper. Bibliometrics is an interdisciplinary science of quantitative analysis of literature information resources by means of mathematics and statistics. Meanwhile, as an emerging approach of scientometrics in recent years, the scientific knowledge mapping analysis is a method showing the structure, laws and distribution of scientific knowledge by means of visualization (Chen et al., 2009). This paper is based on the assistant tool CitespaceⅢ, which is an application for analyzing and visualizing co-citation networks, to facilitate the analysis of emerging trends in a knowledge domain (Chen, 2004).

2. RESULTS AND ANALYSIS

2.1 Results

Following the appropriate search strategies, 1,249 papers were retrieved from the WOS database. The research of text mining in medicine related fields assumed a steady-state growth trend and state of multidisciplinary integration; The United States ranked first in the number of papers issued among different regions, with 532 publications accounting for 42% of the total counts. More detailed measurements are given in the following sections.

2.2 Literature Growth in Recent Years

The changes in the quantity of publications on time sequence can reflect the general research trend in a given field. As shown in Figure 1, numbers of published articles of text mining research in medicine related fields appeared a trend of gentle growth generally, and came to a period of rapid growth after 2008. In the past 3 years, text mining research in the medical field has become increasingly active. There were 174 publications in 2016, reflecting the studies are in a steady growth phase.

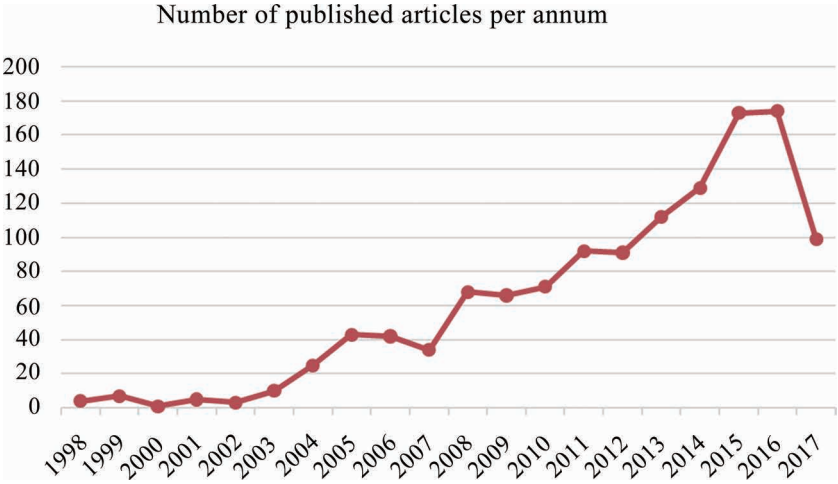

Figure 1
Quantitative Changes in the Literature of Medicine Related Text Mining Research in WOS on Time Sequence

2.3 Distribution of Authors and Institutions

Table 1 shows the statistics of the top 10 high yield authors. As we can see that the top ranked author is Lu, Z.Y., with 31 articles published, and all 10 authors issued a total of 192 articles, accounting for 15.37% of the total number in the WOS database. The ranking on the top ten institutions in the number of publications was displayed in Figure 2. Manchester University and Stanford University have made great contributions to the field of medical text mining research, followed by Dalian University and Columbia University.

Table 1
Top Ten Prolific Authors of Medicine Related Text Mining Research in WOS

Author	Frequency	Percentage (%)
Lu, Z. Y.	31	2.48
Ananiadou, S.	28	2.24
Yang, Z. H.	20	1.60
Lin, H. F.	20	1.60
Nenadic, G.	19	1.52
Wilbur, W. J.	16	1.28
Wei, C. H.	16	1.28
Kostoff, R. N.	15	1.20
Li, Y. P.	14	1.12
Yu, H.	13	1.04
Sum	192	15.37

To be continued

Continued

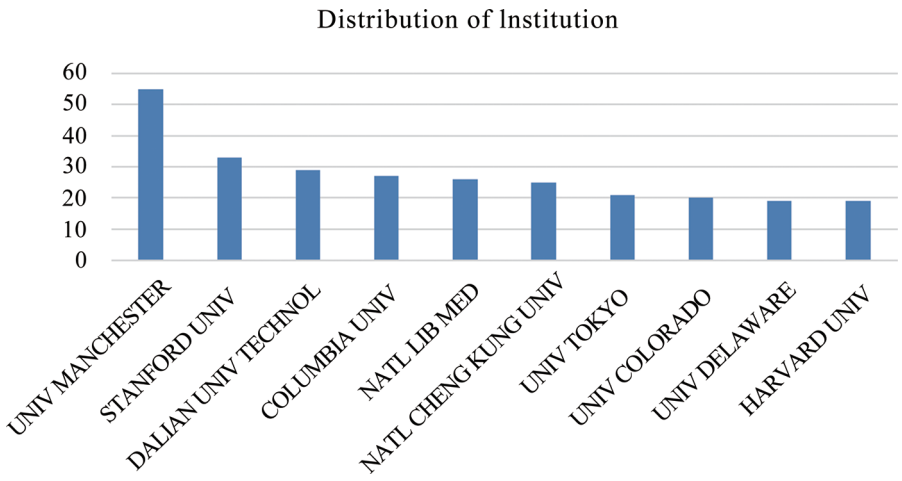

Figure 2
Statistics of Top 10 Research Affiliations Based on the Amount of Publications in WOS

2.4 Distribution of Regions

As shown in Figure 3, the size of the node represents the frequency of article publication in each region. According to the number of publications, The United States ranked first, with 532 publications, accounting for 42% of the total volume. Followed by England and China, each accounted for more than 10%.

2.5 Discipline and Research Direction

Shown as Figure 4, the medicine related text mining studies presented a state of interdisciplinary convergence, and the most active research area is computer science, followed by biochemical research and mathematics or statistics related area. Besides, other major research directions are medical informatics and health care science. Nowadays, text mining technology has been increasingly applied in medical or biomedical researches.

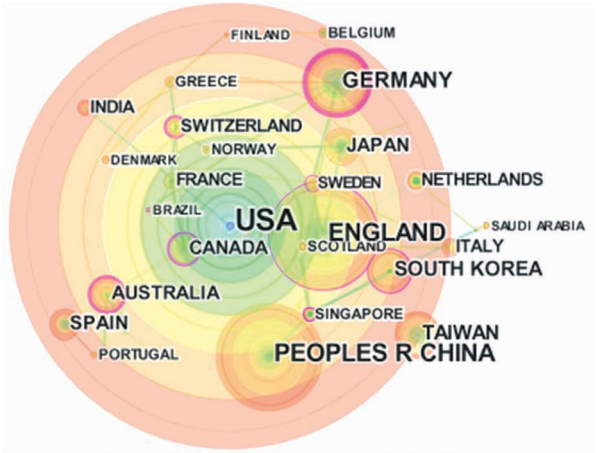

Figure 3
Region Distribution Mapping of Medicine Related Text Mining Research in Web of Science

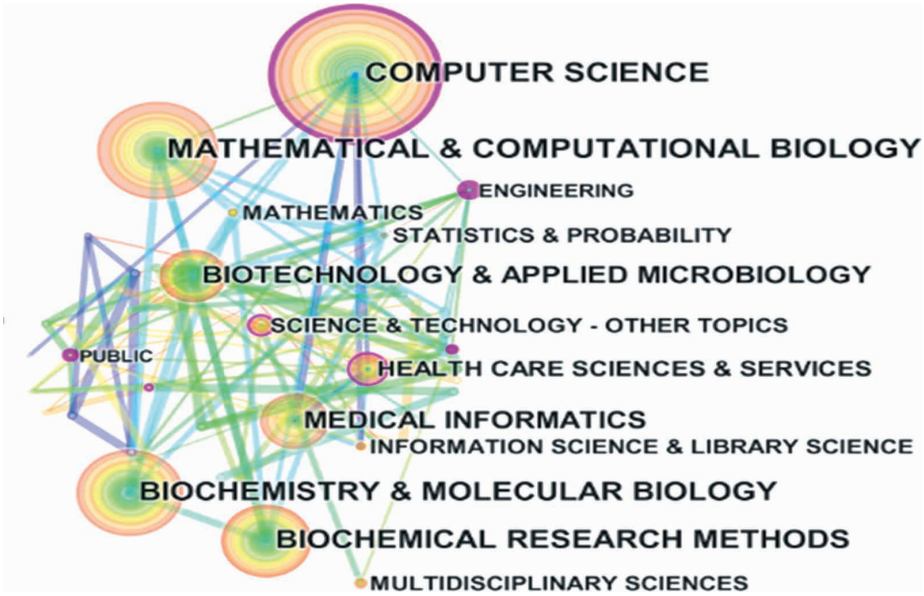

Figure 4
Knowledge Mapping of Research Domain

3. DISCUSSION

3.1 Keywords Co-Appearance Mapping

In this paper, we started discussion with the analysis of keywords of text mining research in the medicine related fields, for keywords are concise and summary of the article themes. With the help of CitespaceIII software, and by setting the time range from 1998-2017, the time slice 2, selecting "Keyword" in the node type and the Pathfinder algorithm, setting the corresponding thresholds, knowledge mapping of keywords co-occurrence was carried out, forming a total of 104 nodes and 307 links. As shown in Figure 5, each node represents a keyword, the size of the node represents the frequency of the keyword appearing in the co-occurrence pair; the centrality of a keyword reflects the probability of the keyword co-existing with other keywords, and its influence in the network (Feng, 2012). We can find the key node based on the size of the node, to reveal the current research state and hot points. Apparently, high frequency keywords are: text, information, database, system, information extraction, natural language processing, biomedical literature, classification, identification, gene, and ontology, etc. And small spots scattered around can also show some fresh research hot words such as: biomedical text mining, named entity recognition, relation extraction, corpus, electric health records, social media etc.

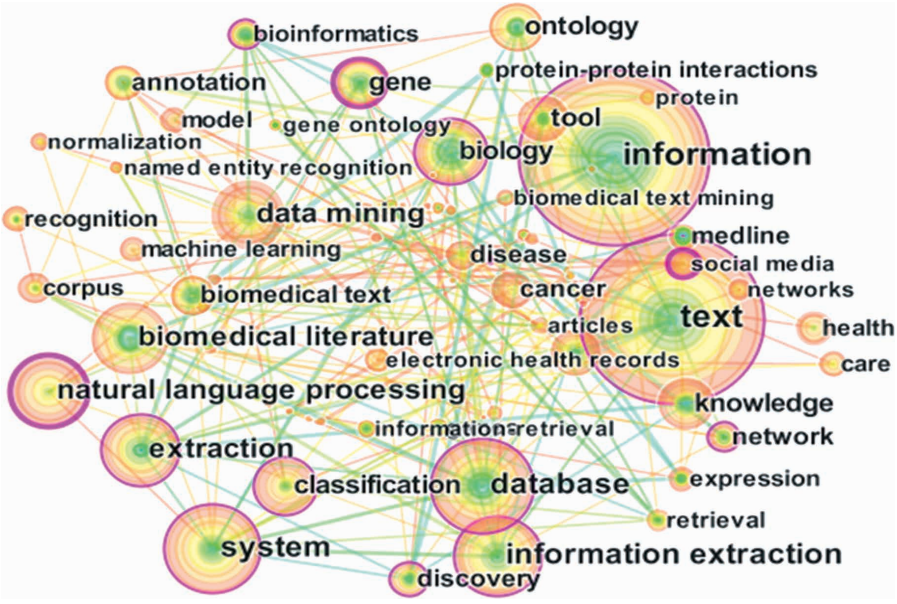

Figure 5
Keyword Co-Appearance Mapping of Medicine Related Text Mining Research in WOS

3.2 Literature Co-Citation Mapping

In a given field, a research front refers to the body of articles that scientists actively cite (Price, 1965), and document co-citation analysis studies a network of co-cited references (Small, 1980). Citespace shows the intellectual base of a research front by citation and co-citation network of scientific publications (Chen, 2006). By selecting "cited reference" in the node type and setting the corresponding thresholds in CitespaceIII, we got the literature co-citation mapping, shown as Figure 6. The centrality of a node is a graph-theoretical property that quantifies the importance of the node's position in a network (Chen, 2006), and it's highlighted by the purple circle. Besides, the color of the circle turns warm represents the cited time was closer. Table 2 listed the title of 12 key documents based on centrality and cited frequency. It can be concluded that the text mining methods are maturely applied to biomedicine or bioinformatics domain. As shown in Table 2, document 1,8 introduced the applications of text mining in biomedical event extraction; document 2,11 focused on the automatic annotation in bioinformatics field based on text mining; document 4 is a biomedical text mining review; document 5,10 introduced the corpus for biomedical information (mainly about PPI or protein, gene, and RNA relationships) extraction; document 6 concentrated on an open-source species name recognition and normalization software system called LINNAEUS; document 9 paid regard to gene normalization task; document 7 introduced the iHOP literature server (http://www.ihop.net.org) aiming at retrieving specified information of genes and proteins interactions or functions by navigating related scientific literature; document 3 refers to an accurate prediction method for protein-protein interaction sites prediction; and document 12 discussed the application of text mining in clinical records as medical concept extraction and assertion or relation classification tasks.

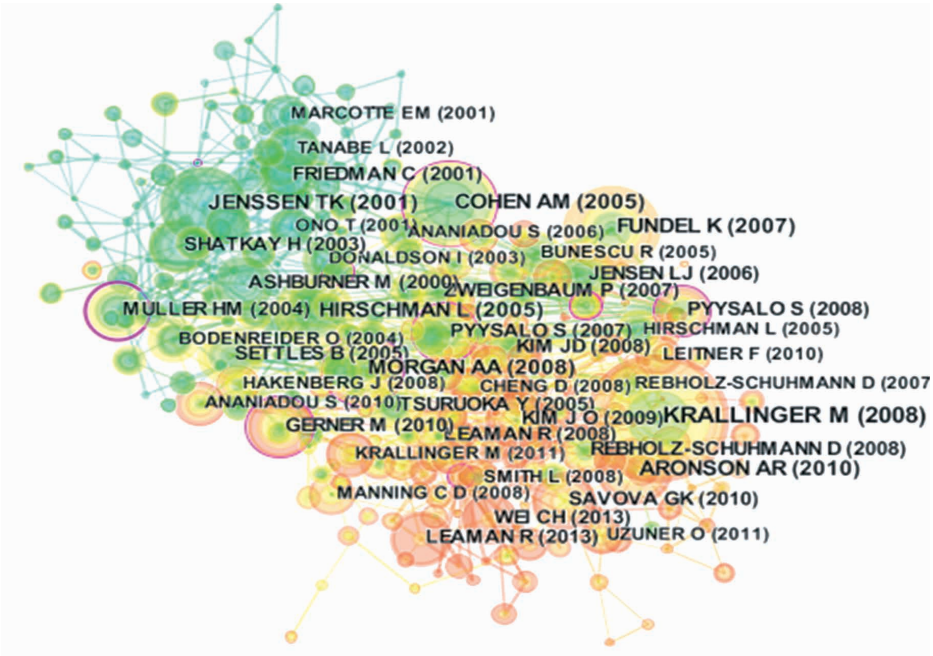

Figure 6
Literature Co-Citation Mapping of Medicine Related Text Mining Research in WOS

Table 2
12 Key Documents of Text Mining Research in Medicine Related Fields Based on Cited Frequency and Centrality in WOS

Sequence	Cited frequency	Centrality	Year	Author	Title
1	33	0.24	2004	Muller, H. M.	Textpresso: An ontology-based information retrieval and extraction system for biological literature
2	14	0.18	2011	Névéol, A.	Semi-automatic semantic annotation of PubMed queries: A study on quality, efficiency, satisfaction
3	17	0.16	2005	Ko J., Murga L. F., Ondrechen M. J.	Prediction of active sites for protein structures from computed chemical properties
4	50	0.15	2005	Cohen, A. M.	A survey of current work in biomedical text mining
5	30	0.15	2008	Pyysalo, S.	Comparative analysis of five protein-protein interaction corpora
6	35	0.14	2010	Gerner M., et al.	Linnaeus: A species name identification system for biomedical literature
7	18	0.13	2005	Hoffmann, R., Valencia, A.	Implementing the iHOP concept for navigation of biomedical literature.
8	28	0.12	2010	Ananiadou, S.	Event extraction for systems biology by text mining the literature
9	43	0.10	2008	Morgan, A. A.	Overview of BioCreative II gene normalization
10	35	0.10	2007	Pyysalo, S.	BioInfer: A corpus for information extraction in the biomedical domain
11	73	0.09	2008	Krallinger, M.	A sentence sliding window approach to extract protein annotations from biomedical articles.
12	15	0.08	2010	Uzune, Ö.	2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text

3.3 Keywords Evolution Mapping in Recent Years

CiteSpace supports a time-zone view to highlight temporal patterns between a research front and its intellectual base (Chen, 2006). By setting the time range from 2011-2017 in Citespace, and selecting the timezone display mode, we got the keywords evolution mapping on the time path in recent years. From Figure 7 we can clearly be informed of the evolution of keywords from 2011 to 2017. The studies focused on keywords such as text mining, information extraction, natural language processing, biology, gene, protein, annotation, corpus in the first time span; then transited to biomedical literature, ontology, identification, classification, recognition and normalization area; afterwards concentrated on electronic health record, knowledge discovery of diseases, as

well as drug discovery and pharmacovigilance; and finally to social media, relation extraction related themes.

Table 3 listed the keywords within 5 years based on cited frequency and centrality. More clearly, studies in recent years have been focused on association extraction, clinical text mining, social media mining and text mining of pharmacovigilance.

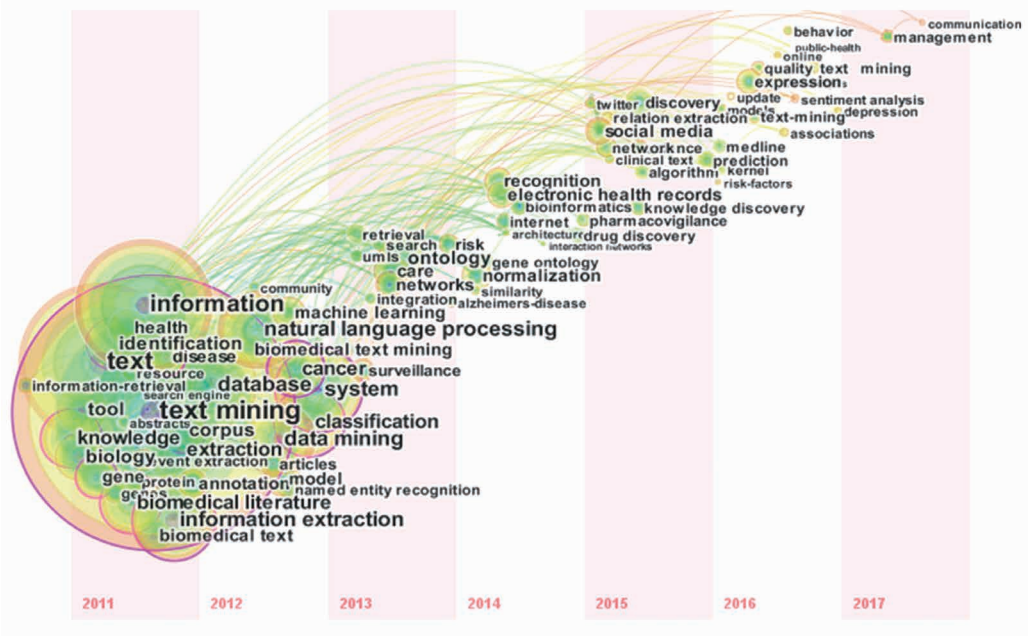

Figure 7
Keywords Evolution Mapping of Medicine Related Text Mining Studies in WOS on Time Path

Table 3
Keywords Statistics of Medicine Related Text Mining Research Based on Centrality and Cited Frequency in Web of Science Last 5 Years

Year	Cited frequency	Centrality	Keywords
2016	11	0.05	Sentiment analysis
	11	0.05	Associations
	10	0.01	Depression
2015	26	0.08	Social media
	11	0.05	Clinical text
	25	0.04	Discovery
	27	0.05	Electronic health records
2014	17	0.03	Bioinformatics
	16	0.00	Pharmacovigilance
	26	0.04	Normalization
2013	21	0.04	Risk
	38	0.03	Ontology
	43	0.18	Cancer
2012	22	0.09	Surveillance
	62	0.05	Natural language processing
	48	0.17	Biomedical literature
2011	33	0.13	Gene
	53	0.11	Classification
	37	0.10	Corpus

3.4 Burst Terms Mapping

Burst detection algorithm can be adapted for detecting sharp increases of interest in a specialty (Kleinberg, 2002). In CiteSpace, a current research front is identified based on such burst terms extracted from bibliographic records (Chen, 2006). By choosing the node type as "Terms" and terms type as "Detect Burst", we got the burst terms mapping shown as Figure 8. Table 4 listed the burst terms within latest 10 years, and we can infer from the data that terms as: annotation, genes, BioCreative II, clinical notes, pharmacovigilance, electronic health record, social media, big data, and relation extraction have strong bursts, which represent the research frontiers in recent years.

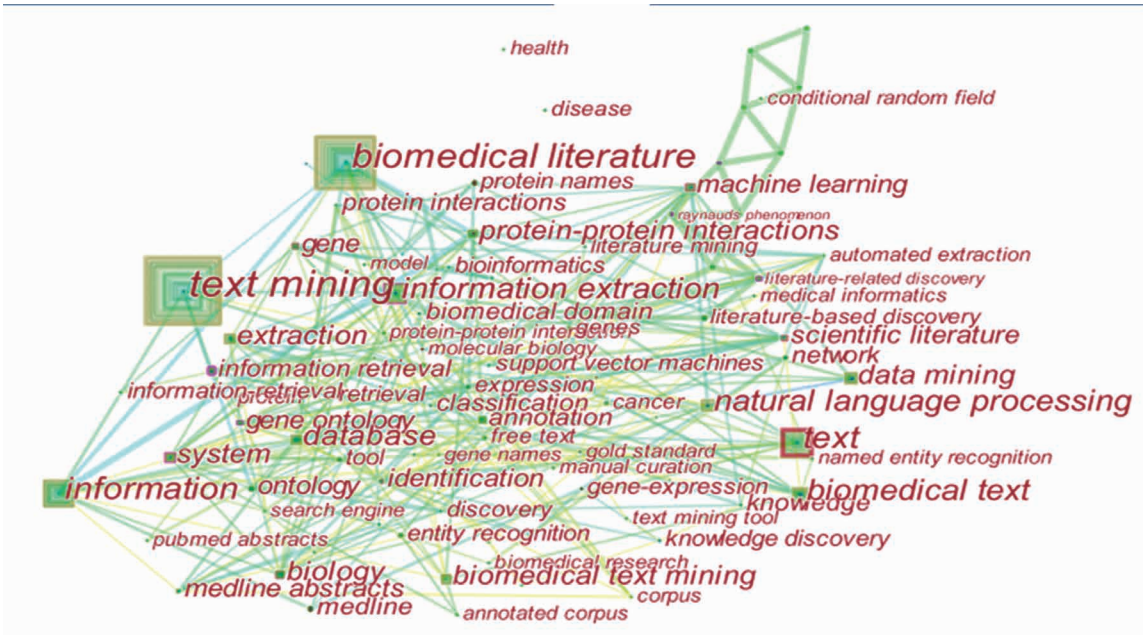

Figure 8
Burst Terms Mapping of Text Mining Research in Medicine Related Fields of WOS

Table 4
Burst Terms of Text Mining Research in Medicine Related Fields Last 10 Years

Burst terms	Burst Rate	Time range
Annotation	5.08	2008-2012
Literature based discovery	3.24	2008-2012
Gene names	3.14	2008-2011
Genes	4.23	2010-2012
Biocreative ii	4.53	2010-2011
Protein-protein interaction	3.77	2010-2011
Biomedical literature background	3.65	2010-2012
Protein-protein interaction extraction	3.24	2011-2012
Community	3.24	2011-2012
Clinical notes	4.23	2013-2015
Umls	3.76	2013-2014
Pharmacovigilance	3.84	2014-2015
Adverse drug reactions	3.15	2014-2017
Electronic health record	6.41	2015-2017
Social media	5.75	2015-2017
Big data	5.39	2015-2017
Relation extraction	4.26	2015-2017

3.5 Research Hot Points and Frontiers Analysis

After having analyzed and discussed the knowledge mapping above, we have learned that the studies revolved around heated topics listed below.

3.5.1 Text Mining Methods and Applications

The topic contains hot points like: NLP, named entity recognition, gene, disease, corpus, NCBI disease corpus, annotation, Pubmed, text mining tool, relation extraction, Medline annotation, literature curation, etc. Relevant studies discussed the problems of named entity recognition, genome and gene expression annotation, construction and annotation of gene or protein or disease corpus, and introductions or applications of various text mining tools. The future research directions may be automatic annotation methods and optimization, relevant ontology construction, and the standardization of text mining assistant tools.

3.5.2 Protein-Protein Interaction Extraction or Biomedical Network Construction

Relevant hot points include gene, biocuration, comprehensive bench mark, kernel method, full-text biomedical article, etc. For instance, Tikk (2010) performed a comprehensive benchmarking of nine different methods for PPI extraction that use convolution kernels on rich linguistic information, and concluded that for most kernels no sensible estimation of PPI extraction performance on new text is possible, and three kernels are clearly superior to the other methods. Papanikolaou (2014) reviewed the text-mining technology aiming to extract information for proteins and their interactions from literature and various biological databases.

3.5.3 Research on Text Mining of Drug Knowledge Discovery

Highlighted terms are pharmacovigilance, drug-drug interactions, adverse drug events extraction, pharmacogenomic, etc. Rave (2014) provided an overview of text mining application in pharmacovigilance, and discussed data sources such as biomedical literature, product labeling, clinical narratives, social media, and web search logs, which are amenable to text mining for pharmacovigilance.

3.5.4 EHR (Electronic Health Record) Text Mining

Contents cover hot points like: clinical information extraction, classification, knowledge discovery, clinical decision support systems, artificial intelligence etc. Uzuner et al. (2008) organized a Natural Language Processing (NLP) challenge on automatically determining the smoking status of patients from information found in their discharge records. Cohen (2013) analyzed a large-scale EHR corpus and quantified redundancy both in terms of word and semantic concept repetition. And found that while the importance of data cleaning has been known for low-level text characteristics, high-level characteristics such as naturally occurring redundancy, can also hurt text mining. Zheng et al. (2016) developed and tested a NLP-based diabetes case finding algorithm using both structured and unstructured electronic medical records (EMRs). Jonnalagadda et al. (2017) described an information extraction-based approach that automatically converts unstructured text into structured data, which is using a rule-based system to identify patients with specific subtypes of heterogeneous clinical syndromes, in order to determine patients qualified for the clinical trial required by precision medicine.

3.5.5 Other Bioinformatic Topics

The other views involve gene regulatory network, Non-coding RNA, gene-disease-drug associations extraction, biomedical event extraction. For example, Pletscher-Frankild et al. (2015) presented a system for extracting disease-gene associations from biomedical abstracts. They considered that text mining should not stand alone, but be combined with other types of evidence. As a consequence, they have developed the disease resource, which integrates the results from text mining with manually curated disease–gene associations, cancer mutation data, and genome-wide association studies from existing databases. Besides, text mining is applied to social media health data mining as well. Marshall et al. (2015) compared and contrasted symptom cluster patterns derived from messages on a breast cancer forum with those from a symptom checklist completed by breast cancer survivors, using methods of co-occurrence and k-medoid clustering. And the study suggests that the copious amount of data generated by social media outlets can augment findings from traditional data sources.

CONCLUSION

Through the knowledge mapping and statistical analysis of medicine related text mining research literature in the WOS database, we can draw the following conclusions: as a promising research field, the literature base shows a steady growth and state of multidisciplinary integration; the United States ranked first by the number of papers issued among different regions; the application of text mining is actively applied to biomedicine, and mainly focused on: gene or other biomedical named entity recognition task, the construction and automatic annotation of gene and protein corpus, protein interaction network construction, protein function prediction, drug interactions or adverse event extraction. There are some studies about text mining applied in clinical notes and disease knowledge discovery as well. Nevertheless most of them are messy and have not formed a mature research system. For the future trend of medicine related text mining research, except for optimizing its own research methods and standardizing the applications of relating tools, text mining technology can also be applying to the mining and extraction of disease and health information,

following by its relatively mature development in bioinformatics; or even combining these research areas to discover the relationships between disease, gene, protein and drug. In addition, the future medical text mining research directions may include as follows: Automatic construction and annotation of disease corpus, construction of medical ontology and standardization of thesaurus, development of intelligent diagnosis based on clinical information extraction and domain knowledge base building, along with text mining of drug interactions and pharmacovigilance, or social media text mining.

REFERENCES

Chen, C. M. (2004). Searching for intellectual turning points: Progressive knowledge domain visualization. *Proceedings of the National Academy of Sciences of the United States of America (PNAS), 101*(1), 5303-5310.

Chen, C. M., Chen, Y., & Horowitz, M., et al. (2009). Towards an explantory and computitional theory of scientific discovery. *Journal of Informetrics, 3*(3), 191-209.

Chen, C. M. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359-377.

Cohen, R., & Elhadad, M., et al. (2013). Redundancy in electronic health record corpora: Analysis, impact on text mining performance and mitigation strategies. *BMC Bioinformatics, 14,* 10. doi:10.1186/1471-2105-14-10

Feng, S. J., Zhao, W. L., & Li, Z. (2012). Research hotspots and evolution path of clinical pathway. *Research on Science and Technology Management, 32*(10), 62-65.

Fleuren, W. W. M., & Alkema, W. (2015). Application of text mining in the biomedical domain. *Methods, 74,* 97-106.

Harpaz, R., & Callahan, A., et al. (2014). Text mining for adverse drug events: The promise, challenges, and state of the art. *Drug Safety, 37*(10), 777-790.

Hearst, M. (1999). Untangling text data mining. *Proc. Assoc. Comput. Linguist., 37,* 3-10.

Jonnalagadda, S. R., Adupa, A. K., Garg, R. P., Corona-Cox, J., & Shah, S. J. (2017). Text mining of the electronic health record: An information extraction approach for automated identification and subphenotyping of HFPEF patients for clinical trials. *Journal of Cardiovascular Translational Research*, 1-9.

Kleinberg, J. (2002). *Bursty and hierarchical structure in streams* (p.91, 101). Proceedings of the 8th ACMSIGKDD Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada: ACM Press.

Marshall, S. A., Yang, C. C., Ping, Q., Zhao, M., Avis, N. E., & Ip, E. H. (2015). Symptom clusters in women with breast cancer: An analysis of data from social media and a research study. *Quality of Life Research an International Journal of Quality of Life Aspects of Treatment Care & Rehabilitation*, *25*(3), 1-11.

Papanikolaou, N., et al. (2015) Protein–protein interaction predictions using text mining methods. *Methods, 74,* 47-53.

Pletscher-Frankild, S., et al. (2015). DISEASES: Text mining and data integration of disease–gene associations. *Methods*, *74,* 83-89.

Price, D. D. (1965). Networks of scientific papers. *Science, 149,* 510-515.

Small, H. (1980). Co-citation context analysis and the structure of paradigms. *Journal of Ocumentation*, *36*(3), 183-196.

Tikk, D., Thomas, P., Palaga, P., Hakenberg, J., & Leser, U. (2010). A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Comput Biol, 6*(7), e10.

Uzuner, O., et al. (2008). Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc, 15*(1), 14-24.

Zheng, L., Wang, Y., & Hao, S., et al. (2016). Web-based real-time case finding for the population health management of patients with diabetes mellitus: A prospective validation of the natural language processing-based algorithm with statewide electronic medical records. *Jmir Medical Informatics, 4*(4), e37.