Bursty Topic Detection in Cancer Field Based on Multi-features of Burst Words

xiaocui gong (425641137@qq.com)
Chinese Academy of Medical Sciences & Peking Union Medical College Institute of Medical Information
https://orcid.org/0000-0001-6815-3546

xinying an
Chinese Academy of Medical Sciences & Peking Union Medical College Institute of Medical Information

lianhui shan
Chinese Academy of Medical Sciences & Peking Union Medical College Institute of Medical Information

yingxin hao
Tongji Hospital Department of intensive care unit,tongji university,shanghai

yifei li
Chinese Academy of Medical Sciences & Peking Union Medical College Institute of Medical Information

Research article

Keywords: MetaMap,UMLS,Bursty Topic,Cancer Field

Posted Date: August 31st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-64804/v1

Abstract

Background: With the increase in the amounts and types of medical information resources and the interdisciplinarity of the related works,it has become increasingly challenging for researchers and information personnel to grasp the theme development.Considering the prominent position of cancer field among all the subject areas in scientific research, and in order to carry out a new bursty topic detection method, the author proposes a method based on the multi-features of burst words and demonstrates its operating process.

Methods: The main stages in the process include concept mapping,topic words extraction,burst words extraction,burst topics identification.This paper continues to take the study of breast neoplasms treatment research as a field to test the new method for identifying burst topics in medical research.

Result: With this method,we identified four burst topics in breast neoplasms treatment research, including Molecular Targeted Therapy of Triple Negative Breast Neoplasms,molecular pharmacology of Triple Negative Breast Neoplasms,Immunological Antineoplastic Agents therapeutic use of Breast Neoplasms and Surgery therapy of Breast Neoplasms.

Conclusion: The test results are highly concordant with authoritative literature reviews in the field and are further confirmed by interviews with the field’s leading experts,thus verifying the reliability of the techniques and approaches proposed by the study.

Introduction

With the rapid increase in the number and types of medical information resources, researchers and intelligence analysts have more and more challenges in identifying emerging hot topics in the disciplinary field.How to quickly and actively identify and judge the topic development trend from the massive information resources has become a hot topic in the field of intelligence research.

The identification of topic evolution is helpful for researchers to grasp the research trend and realize scientific and technological innovation.Scientific and technological innovation needs the early detection and identification of relatively good emerging topics.By evaluating the development trend of the subject, we can make effective scientific research topic selection and scientific decision-making.Only by accurately grasping the development and changes of the research field and identifying the development vein of the research topics can the research theme be effectively planned and the advantageous development direction be selected as the breakthrough point of China's scientific and technological innovation.

In this study, we focus on the detection of burst topics which are usually difficult to predict and have a certain degree of social heat and burst.The burst topics may be one that splits from other topics, or it may be a new concept that has been put forward recently.The burst topics can exist and evolve for a long time,or it may be suddenly disappeared after existing a period of time.

The identification of burst topics mostly concentrated in the field of online opinions or important events,the study of burst medicine topics is less.Considering the prominent position of cancer field among all the subject areas in scientific research, we proposes a method based on multi-features' burst words detection for judging burst topics in cancer field and demonstrates its operating process, the main stages in the process include concept mapping,topic words extraction,burst words extraction and burst topics identification.

The organization structure of this paper is as follows:The second section introduces related work. In the third section we introduce our research methods. The "Experiment and result" section presents the experimental and the result.The conclusion and prospect are given in the last section.

Related Work

Foreign studies on burst topics started relatively early,their research technology is relatively mature which involving a variety of technical methods,it includes technical methods based on clustering, frequent pattern mining, Exemplar, matrix decomposition and probabilistic topic model[[1]].In China, the related literature on the detection of burst topics of online opinion has been increasing year by year, and has become the research focus in the field of opinion mining in recent years.

Topic Detection and Tracking is one of the study point in the field of Natural Language Processing (NLP) from the beginning[[2]].Topic Detection and Tracking includes two hot categories,one is hot topic detection, another is burst topic detection. The topic with more documents is more likely to become a hot topic, and the topic of breaking out is different from the hot topic, the purpose of its detection is to find new topics that did not exist before or did not get a lot of attention before.There are two main techniques for burst topics recognition: one is bibliometrics, which mainly refers to the analysis of burst keywords; the other is non-supervised technology, which mainly includes clustering methods and topic model methods.

Bibliometrics methods

Word frequency analysis method:

the method based on word frequency analysis mainly analyzes the characteristics of keywords with time through word frequency statistics, word growth rate calculation and word weight calculation, so as to identify the burst words.the method based on word frequency is relatively simple and can reveal the development of the topic directly.For example, Kleinberg et al[[3]]. proposed a more classical method which is based on modeling the stream using an infinite-state automaton, in which bursts appear naturally as state transitions,the hidden markov method is used to analyze the word frequency distribution in the burst state and the transfer probability of the automaton.Mathioudakis et al[[4]] found that the rise of emerging topics would increase the frequency of topic-related keywords in a period of

time. He defined such topic-related words as bursty keywords and proposed QueueBurst algorithm to detect bursty keywords,and presented TwitterMonitor, an analysis system for emerging topics widely used on Twitter.Du,Y et al[[5]], proposed a novel bursty topic detection technique based on an improved method by calculating term weight, they used a novel aging theory to model a term life cycle,then calculated user weight through improved PageRank algorithm to express term weight,at last adopted a unsupervised learning algorithm to detect bursty topic.

Word frequency analysis can only reflect the hot or sudden degree of a topic from the word occurrence frequency, but cannot reveal the semantic correlation analysis between words.Noise removal and accurate extraction of burst words are important factors to improve the detection rate.

Information entropy method:

Calculate the mutation point and identify the evolution of information based on it.For example,taking water resources management papers from 1990 to 2011 as the data source,Wang L.y.[[6]] used the method for identifying mutation point to obtain the literature keyword frequency mutation point in the dissertation collections and based on this divided the evolution process of the research themes of basin water resources management. Another algorithm is the KL algorithm,the Kullback-Leibler divergence (also called relative entropy) is a measure of how one probability distribution diverges from a second, expected probability distribution[[7]] which is a information entropy algorithm.

Co-word analysis:

with its own advantages and characteristics, co-word analysis has become an important means of topics evolution analysis. Based on this method, there are two main ways to identify burst topics: one is to form a co-occurrence network firstly and then identify the burst features;the other is to extract and group the burst characteristics and features firstly and then use co-occurrence method for clustering topics.The co-word analysis is mainly based on the quantitative calculation of the co-occurrence frequency of topic words.Fang,et al[[8]] proposed a bursty keyword discovery scheme using co-occurrence and time information of words.They generate pairs of co-occurring words by applying OpenNLP,and extract bursty keywords by analyzing word clusters within a specified time range.Professor Chen chaomei from drexel university in the United States has designed and developed Citespace software to detect bursty word based on vocabulary increase rate, and construct a co-occurrence word network so as to detect research fronts and bursty topic.

Bibliometrics methods started relatively earlier and it has a limited degree in revealing the connotation of the topics. The application of unsupervised methods such as probabilistic topic model is gradually emerging.

Non-supervised technology

Many studies have used clustering method to identify bursty topics,usually this approach firstly need to generate word-document matrix or co-word matrix and then use the similarity algorithm or distance algorithm to cluster the matrix.

Another unsupervised approach to identify bursty topics is the probabilistic topic model method.Latent Dirichlet allocation (LDA)[[9]],a generative probabilistic model for collections of discrete data such as text corpora, has shown a good performance in general text mining. However,it do not consider the temporal information. In order to model stream data, Some temporal extensions of LDA have been proposed. Blei et al. proposed the DTM[[10]] to capture the Markov property of the popularity and content evolution of topics; Wang and Macallum proposed the TOT[[11]], which used Beta distribution to simulate the distribution of topics over time.The major problem with these models is that they cannot discover the burstiness of topics.

Takahashi et al[[12]] proposed a burst model for detecting bursty topics in a traditional topic model.Based on the burst model that is for burst topics but not for whole topics,they applied the burst model to the topics estimated by the traditional dynamic topic model.

Qi Xiang et al[[13]] propose a new topic model named Burst-LDA, which simultaneously discovers topics and reveals their burstiness through explicitly modeling each topic's burst states with a first order Markov chain and using the chain to generate the topic proportion of documents in a Logistic Normal fashion. A Gibbs sampling algorithm is developed for the posterior inference of the proposed model.

Considering the semantic information of text words,probabilistic topic model method can reveal more abundant knowledge.But it involves the selection of many parameter thresholds which are mostly set according to the experience of scholars.At the same time, noise information removal is also a key factor affecting the result.

This study aims to propose a new method for the recognition of burst topics in medical texts, and strive for a breakthrough in the recognition accuracy and execution efficiency.Combining with the knowledge organization system [[14]] in the medical field, We identify burst words from three characteristics: word frequency characteristic, word frequency increment characteristic and word semantic characteristic,and then get the burst word set. Finally, we study the appropriate clustering algorithm according to the burst word set to obtain the burst topics.

Methods

Research framework

Based on the previous research on topic recognition in the medical field, this paper builds a burst topics recognition model in the medical field. The framework is shown in Fig.1.

Burst topics can be defined as sudden increase topics during a period of time,which can help us to notice the most urgent and important research in large collections.In this study,the identification of burst topics in medical field is mainly divided into the following steps:topic words extraction,burst words extraction,burst topics extraction.

Topic words extraction

The interpretation and use of medical information is a very complicated issue.Due to the special nature of medical information data, using KOS to preprocess medical information data is particularly important (Wu et al. 2015)[[1]]. Unifed Medical Language System (UMLS)(Bodenreider 2004)[[2]] is one of the most important KOS in the biomedical field, MetaMap(Aronson and Lang 2010)[[3]] is a tool for obtaining concepts from the text based on the UMLS.This article uses the mesh words from the pubmed medical text as an object of study,and to supplement more meaningful entity words, we use UMLS to map the title and abstract of the text to get the relevant entity words,UMLS can extract concepts and concepts semantic types from biomedical terms,we use the MetaMap tool to complete the mapping.

Firstly, English journal articles from pubmed are selected as data sources, then we search cancer-related literature data in a given field for download. One year is selected as a time window, and mesh words of each time window are extracted as standby.Text mapping tool MetaMap is used to map free text to concept words.Concept words processed by MetaMap are marked with scores and semantic types, we select the result of the first set of meta mapping and store them into the database for later use.

Burst words extraction

Although we have obtained the word set for each document, the word set includes a large number of meaningless words and generic words, such as Adult, Aged, Female and Humans. We define a common word dictionary and filter out meaningless words and generic words.In addition, it is also a very important problem to identify burst words from a large number of word sets. This paper identifies burst words from the multidimensional features of words, including three dimensions:word frequency characteristic, word increment characteristic and word semantic characteristic, in the hope to find more meaningful burst words.

Word frequency characteristic:

word frequency characteristic is the most intuitive reflection of the importance of a word in the data set of the time window. If a word appears frequently, it means that the word is more relevant to the burst words in the time window. This paper uses the method of Wang Jian(2018)[[4]],instead of setting the word frequency threshold directly, it considered the word frequency relative to the highest word frequency in a single time window.The Equation is as follow:

$$C_n(w) = \frac{tf_n(w)}{tf_n^{max}} \qquad (1)$$

In Eq.1,$C_n(w)$ represents the word frequency weight of the word w in time window T_n, $tf_n(w)$ represents the word frequency of the word w in time window T_n, tf_n^{max} represents the max frequency of the word in time window T_n.This method can keep the words with relatively high word frequency while extracting the burst words.

Word frequency increment characteristic:

The weight of word frequency considers only the high frequency words in a time window, but does not considers the changing trend of word frequency.If an event occurs suddenly, the burst words increase sharply in the time window,therefore,the word frequency increment characteristic is introduced to identify the burst words.A burst word may be a new word or an existing word which burst suddenly. Therefore, this paper considers any of the following conditions as a candidate set of burst words:

(1) words do not appear in time window T_n but appear in T_{n+1} and T_{n+2} ;

(2) words appear in both time window T_n and time window T_{n+1} and its word frequency increment is greater than a certain threshold.

Word semantic characteristic:

since medical texts involve words with the entity meaning of genes,proteins,enzymes,drug,etc.,the word frequency of these words may be small, which is easy to be omitted based on the above two methods.Therefore, in this paper, the semantic types of words with entity meaning such as gene, protein and enzyme were selected, and those words with low word frequency but high annual growth rate were reserved as burst words.

Burst topics extraction

In order to better retain effective documents, we reserve documents whose number of burst words is greater than a certain threshold.According to the texts where the burst words are located, we construct the burst words-document matrix. Based on the matrix, Repeated Bisection is used to cluster the matrix with GCLUTO software.

Burst words-document matrix constructing:

each text containing burst words can be represented as a vector:$text_i$={$text_{i,j}$|i=1,2,3...,T,j= 1,2,3,...,N},$text_i$ represents the ith text, $text_{i,j}$ represents the jth word of the ith text.$Text_{i,j}$=1 means the inclusion relation, and $text_{i,j}$=0 means no inclusion relation.

Burst topics identification by Clustering:

based on the matrix of burst words-document, this paper uses repeated bisection to identify burst topics through double clustering.Similarity Function selects cosine function.The cosine function calculation Equation is as follows:

$$\cos ine = \frac{\sum_{k=1}^{m} T_{ki} T_{kj}}{\sqrt{\sum_{k=1}^{m} T_{ki}^{2} \sum_{k=1}^{m} T_{kj}^{2}}} \qquad (2)$$

T_{ki} is the concept k in document i T_{kj} is the concept k in document j.

Tools

(1) MetaMap https://mmtx.nlm.nih.gov/ :MetaMap is a program that matches the biomedical text with the concepts in the UMLS thesaurus. This program can choose parameters to control the internal operation mode and the output form of the results.

(2) Java program and MySQL database: we need to use Java program to complete the format conversion, data processing and calculation.

(3) gCLUTO: it can use co-occurrence matrix or word-document matrix to perform double clustering analysis which can cluster rows and columns at the same time,the clustering methods have four kinds: Repeated Bisection, Direct, Agglomerative and Graph.We can choose according to our needs.It has the following characteristics:management data files, providing visual solutions and presenting clustering tree view of the project.

Results

Topic words extraction results

This study takes the research of breast cancer treatment field as an example, we takes the PubMed database as the retrieval entry, limiting the search strategy to "Breast Neoplasms/therapy"[Mesh],what is more,we restrict the literature type to Journal Article and limit the research object to Humans and record

the retrieval time to May 31, 2019. The retrieval results of each year are shown in the following table and downloaded in MEDLINE format.

Table 1
Literature search results

year	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
num	4803	4944	5793	6235	6576	6635	6718	7063	6930	6641

(1)We store each article in the mysql database and extract the mesh words of each article as a standby, mesh words are shown in Fig. 2.

(2)We use the online tool MetaMap (https://mmtx.nlm.nih.gov/) for concept mapping,choosing parameters:1,No Derivational Variants;2, Ignore Stop Phrases;3,the Use of Word Sense Disambiguation.The first set of Meta Mapping results are selected in the experiment.the umls mapping words are shown in Fig. 3.

Burst words extraction results

We define a common word dictionary and filter out meaningless words and generic words such as Adult, Aged, Female and Humans.After repeated experiments, words with a word frequency characteristic greater than 0.12 and a word frequency increment characteristic greater than 0.60 were selected as burst words.

At the same time, in order to supplement the entity words that have a small word frequency but have a sudden increment, according to the mapping results of MetaMap, specific semantic types with large word frequency increment are selected to be reserved, and the semantic types include Organism Function Congenital Abnormality Therapeutic or Preventive Procedure Indicator, Reagent, or Diagnostic Aid Drug Cell Function Cell or Molecular Dysfunction Therapeutic or Preventive Procedure Body Substance Neoplastic Process Tissue Virus Antibiotic Body Location or Region Pathologic Function Sign or Symptom Physiologic Function Diagnostic Procedure Amino Acid, Peptide, or Protein,Immunologic Factor Gene or Genome Body Part, Organ, or Organ Component etc.Finally, we got 95 burst words.

Burst topics extraction results

In order to further limit and filter out meaningless documents, we select documents with burst words num no less than 2 to retain,we construct a burst words-document matrix, perform double clustering analysis on the matrix. After repeated experiments, when the burst words are grouped into 4 classes, the intra-class similarity is large, and the similarity between classes is small, which is the most suitable. The clustering result is shown in Fig. 4. The burst words are the rows, the article PMIDs and years are the columns,by analyzing the burst words classes and the corresponding document information,we can identify the burst topics and its corresponding documents and years.Burst topics recognition result is shown in Table 2.We reserve the burst words that are most relevant to the topic.

Table 2
Burst topics recognition result

Topic name	Topic words
Molecular Targeted Therapy of Triple Negative Breast Neoplasms	Triple Negative Breast Neoplasms/*drug therapy/metabolism/pathology□Molecular Docking Simulation□Cell Cycle Checkpoints/drug effects□Autophagy/drug effects□ Triple Negative Breast Neoplasms/*drug therapy/metabolism□Molecular Targeted Therapy□Signal Transduction/drug effects□Xenograft Model Antitumor Assays□ Apoptosis/drug effects□MCF-7 Cells□Real-Time Polymerase Chain Reaction□ MicroRNAs/*genetics□Cell Proliferation/genetics□Triple Negative Breast Neoplasms/*drug therapy/genetics/pathology□RNA Interference□HEK293 Cells
molecular pharmacology of Triple Negative Breast Neoplasms	Triple Negative Breast Neoplasms/*drug therapy/pathology□Molecular Targeted Therapy/methods□Cyclin-Dependent Kinase 4/antagonists & inhibitors□Cyclin-Dependent Kinase 6/antagonists & inhibitors□Biomarkers□NPS gene_Gene or Genome□Cell Cycle Checkpoints_Cell Function□Blood group antibody A_Amino Acid, Peptide, or Protein,Immunologic Factor□Phyllodes□PIP gene_Gene or Genome□ Mutant_Cell or Molecular Dysfunction□Recombination, Genetic_Genetic Function□3-hydroxyflavone_Organic Chemical,Pharmacologic Substance□pertuzumab_Amino Acid, Peptide, or Protein,Immunologic Factor,Pharmacologic Substance□ Nanomedicine_Occupation or Discipline□Lymphocyte_Cell
Immunological Antineoplastic Agents therapeutic use of Breast Neoplasms	Trastuzumab/therapeutic use□Antineoplastic Agents, Immunological/therapeutic use□Trastuzumab/*therapeutic use□Antineoplastic Agents, Immunological/*therapeutic use□Receptor, ErbB-2/genetics□ Trastuzumab/administration & dosage/adverse effects□Cancer Survivors□ Bevacizumab/administration & dosage□Capecitabine/administration & dosage□ Trastuzumab/administration & dosage□Biomarkers, Tumor□Neoplasm Grading□ Kaplan-Meier Estimate□Receptor, ErbB-2/metabolism□Cardiotoxicity□ Trastuzumab/*administration & dosage/adverse effects□Receptor, ErbB-2/*genetics□Trastuzumab/*administration & dosage
Surgery therapy of Breast Neoplasms	Breast/diagnostic imaging/pathology/surgery□Biopsy, Large-Core Needle□*Margins of Excision□*Prophylactic Mastectomy□Breast/diagnostic imaging/pathology□ Sentinel Lymph Node/*pathology□*Acellular Dermis□Mastectomy□Patient Reported Outcome Measures□Margins of Excision□*Radiation Dose Hypofractionation□ Cancer Survivors/*psychology

Discussion

This article explores the method of identifying burst topics in the cancer field and validates it in the breast cancer field. The experiment process includes topic words extraction,burst words extraction,and burst topics extraction.The topic words extraction process combines with the knowledge organization system(KOS) in the medical field to identify more meaningful topic words. The burst words extraction process includes three dimensions:word frequency characteristic,word frequency increment characteristic and words semantic characteristic which identify more meaningful burst words.The burst topics extraction process constructs a burst words-document matrix and applies the method of double clustering to identify burst topics,in the end we got four burst topics:

Topic one:Molecular Targeted Therapy of Triple Negative Breast Neoplasm.Drug therapy for triple negative breast cancer, especially molecular targeted therapy, has been increasingly studied in recent years.

Topic two:molecular pharmacology of Triple Negative Breast Neoplasms.The research on the molecular pharmacology of triple negative breast cancer is gradually increasing, especially the research on its related genes, proteins and enzymes.

Topic three:Immunological Antineoplastic Agents therapeutic use of Breast Neoplasms.The research on immunotherapy of breast cancer has been increasing year by year, and some related genes, drugs and receptors have been explored.

Topic four:Surgery therapy of Breast Neoplasms.Surgical treatments for breast cancer are also being explored, aiming to improve survival.

To sum up, the treatment of breast cancer has been in the exploration and research, the innovative treatment plan is constantly trying, the treatment method is more and more diversified, scientific.Through relevant literature research and expert consultation, the results of this study are reliable and prove the feasibility of this method.

This article has a lot of theoretical and practical significance. It can help scientific researchers quickly grasp the current hot topics, grasp the research direction, and it can help medical managers quickly identify the current development frontiers and make decisions.However,the topic words extraction process uses the knowledge organization system(KOS) in the medical field to identify more meaningful topic words,because the KOS is lagging, how to combine the KOS and identify new words is a problem that needs to be solved. At the same time, the identification of burst topics requires experts interpretation. How to automatically summarize and interpret burst topics is a question worth exploring.

Conclusion

This article explores the method of identifying burst topics in the cancer field and validates it in the breast cancer field and finally identified four burst topics in breast cancer treatment area, Through the method and specific practice in this paper, we can provide reference for the identification of burst topics in the medical field.which can provide decision support for medical managers.

Declarations

Acknowledgements

The authors wish to thank the National Key Research and Development Project(Grant No. 2016YFC0901902-2) and the Project of Fundamental Research Funds for the Central Universities (Grant No. 2018TX63002, 2016-12M-3-018) for their financial support.

Authors’ contributions

All authors contributed to the work described in this manuscript. All authors

have approved the final version of the manuscript. The detailed division of

labor was as follows: XCG provided the original research idea and creation of the

manuscript. XYA and LHS provided advice and expertise throughout the research. YXH performed the data analysis. YFL prepared the tools and wrote part of the manuscript.

Funding

This work is supported by the National Key Research and Development Project(Grant No. 2016YFC0901902-2) and the Project of Fundamental Research Funds for the Central Universities (Grant No. 2018TX63002, 2016-12M-3-018).

Availability of data and materials

The datasets used in the current article are available from the corresponding

author on reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Ibrahim R,Elbagoury A,Kamel M S,et al.Tools and approaches for topic detection from Twitter streams: survey[J].*Knowledge and Information Systems*,2017:1-29.
2. Chen,Y.N.,Liu,L.Z.,& IEEE.(2016).Development and research of topic detection and tracking.In Proceedings of 2016 IEEE 7th international conference on software engineering and service science.*International conference on software engineering and service science* (pp.170–173).New York:IEEE.

3. Kleinberg J.Bursty and hierarchical structure in streams[J].*Data Mining and Knowledge Discovery*,2003,7(4):373-397.
4. Mathioudakis M,Koudas N.Twitter Monitor:trend detection over the twitter stream[C]. *ACM SIGMOD International Conference on Management of Data*, *SIGMOD2010*,Indianapolis,Indiana,Usa,June.DBLP,2010,1155-1158.
5. Y. Du,Y. He,Y. Tian,Q.Chen and L.Lin,"Microblog bursty topic detection based on user relationship",in Information Technology and Artificial Intelligence Conference (ITAIC),2011 6th IEEE Joint International,vol.1,(2011),pp.260-263.
6. Wang,L.Y.(2013).Subject mutation research based on keyword mutation.Information Studies:Theory & Application,36 (11):45-48.
7. Kullback,S.,& Leibler,R.A.(1951).On information and sufciency.Annals of Mathematical Statistics,22(1),79-86.https://doi.org/10.1214/aoms/1177729694.
8. F.Fang,N.Pervin,A.Datta,K.Dutta and D.VanderMeer, "Detecting Twitter Trends in Real-Time",Proceedings of the 21st Workshop on Information Technologies and System (WITS), (2011).
9. D.M.Blei,A.Y.Ng and M.I.Jordan,"Latent dirichlet allocation",J.Mach.Learn.Res., vol.3,(2003) March,pp.993-1022.
10. D.M.Blei and J.D.Lafferty.Dynamic topic models.Proceedings of the 23rd International Conference on Machine Learning,Pittsburgh,USA,June 2006,113–120.
11. X.Wang and A.McCallum.Topics over time: a non Markov continuous-time model of topical trends.Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,Philadelphia,USA,August 2006,424–433.
12. Y.Takahashi,T.Utsuro,M.Yoshioka,N.Kando,T.Fukuhara,H.Nakagawa and Y.Kiyota,"Applying a Burst Model to Detect Bursty Topics in a Topic Model", in Advances in Natural Language Processing,Springer Berlin Heidelberg,(2012),pp.239-249.
13. Qi Xiang,Huang Yu,Chen Ziyan et al.(2014).BURST-LDA:A NEW TOPIC MODEL FOR DETECTING BURSTY TOPICS FROM STREAM TEXT.JOURNAL OF ELECTRONICS(CHINA),31(6):565-575.
14. Mayr, P.,Tudhope,D.,Clarke,S.D.,Zeng,M.L.,& Lin,X.(2016).Recent applications of Knowledge Organization Systems:introduction to a special issue.International Journal of Digital Library Systems,17(1),1-4. https://doi.org/10.1007/s00799-015-0167-x.
15. Wu,Q.Q.,Zhang,H.B.,& Lan,J.(2015).K-State automaton burst detection model based on KOS:Emerging trends in cancer field.Journal of Information Science,41(1), 16–26.https://doi.org/10.1177/0165551514551500.
16. Bodenreider,O.(2004).The Unifed Medical Language System (UMLS): integrating biomedical terminology.Nucleic Acids Research,32(Database issue),D267-D270. https://doi.org/10.1093/nar/gkh061.
17. Aronson,A.R.,& Lang,F.M. (2010).An overview of MetaMap:historical perspective and recent advances.Journal of the American Medical Informatics Association,17(3),229-

236.https://doi.org/10.1136/jamia.2009.002733.
18. Wang,J.Research on the detection method of microblog Emergencies Based on multi-features fusion(Master).Beijing Information Science and Technology University, Beijing, China, 2018.

Figures

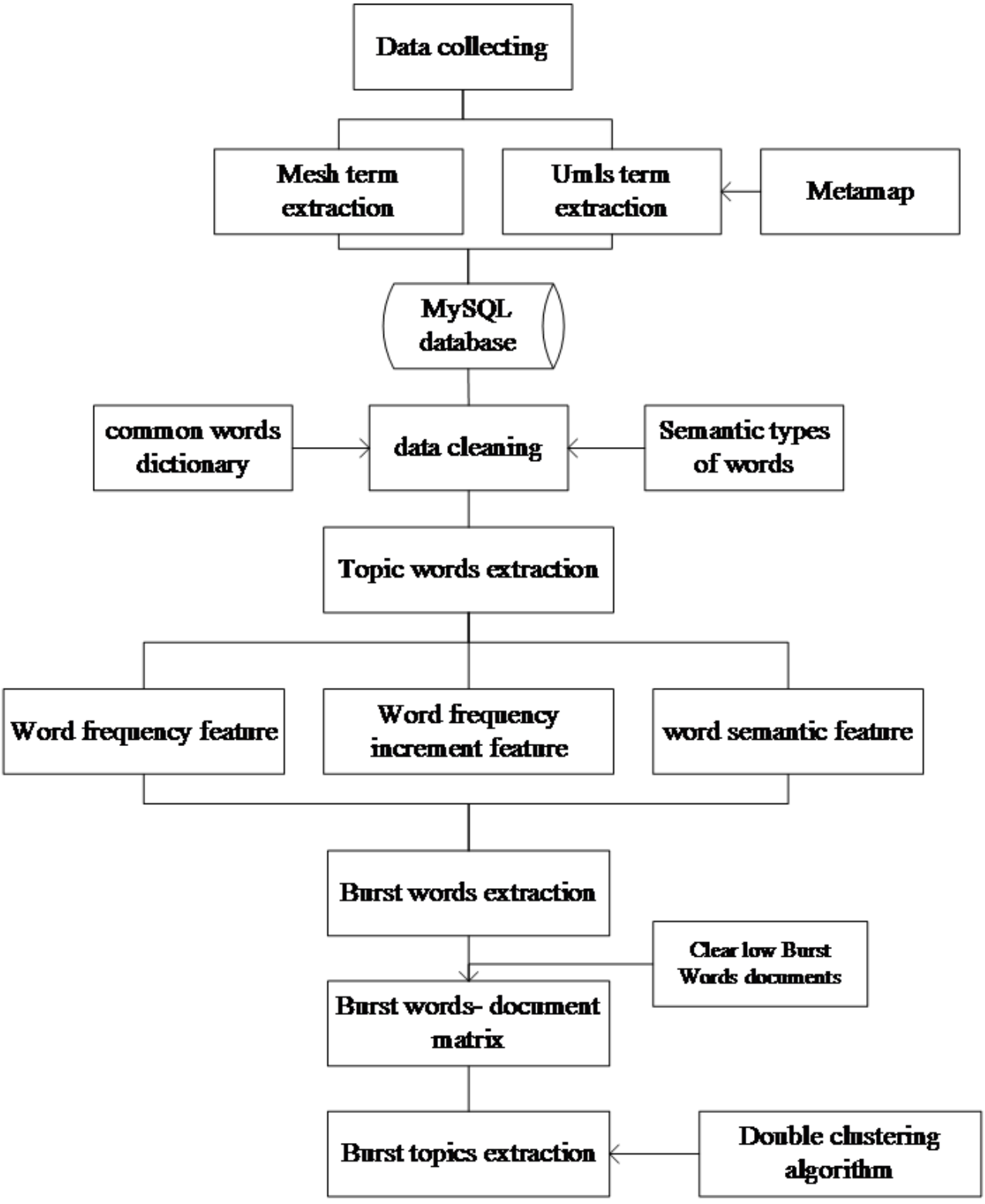

Figure 1

Framework of burst topics discovery model

PMID	MH	TI	AB
29935901	Cyclin-Dependent Kinase 4/*antagonists & inhibitors	CDK4/6 Inhibition as a t	With 40,920
29935901	Cyclin-Dependent Kinase 6/*antagonists & inhibitors	CDK4/6 Inhibition as a t	With 40,920
29935901	Humans	CDK4/6 Inhibition as a t	With 40,920
29935901	Molecular Targeted Therapy	CDK4/6 Inhibition as a t	With 40,920
29935901	Piperazines/therapeutic use	CDK4/6 Inhibition as a t	With 40,920
29935901	Protein Kinase Inhibitors/*therapeutic use	CDK4/6 Inhibition as a t	With 40,920
29935901	Purines/therapeutic use	CDK4/6 Inhibition as a t	With 40,920
29935901	Pyridines/therapeutic use	CDK4/6 Inhibition as a t	With 40,920
29935900	Aminopyridines/therapeutic use	Mechanisms of therape	Cyclin depe

Figure 2

Mesh words

pmid	mapping	class	location	sentence
28988530	Administration procedure	Therapeutic or Preventive Procedure	ab	1:
28988530	TP53 gene	Gene or Genome	ab	2:
28988530	AKT1 gene	Gene or Genome	ab	2:
28988530	GPR162 gene	Gene or Genome	ab	3:
28988530	Treated with	Therapeutic or Preventive Procedure	ab	7:
28988530	AKT1 gene	Gene or Genome	ab	8:
28988530	Neoplasm	Neoplastic Process	ab	9:
28988530	Neoplasm	Neoplastic Process	ab	10:
28988530	Therapeutic procedure	Therapeutic or Preventive Procedure	ab	10:

Figure 3

Umls concept mapping words

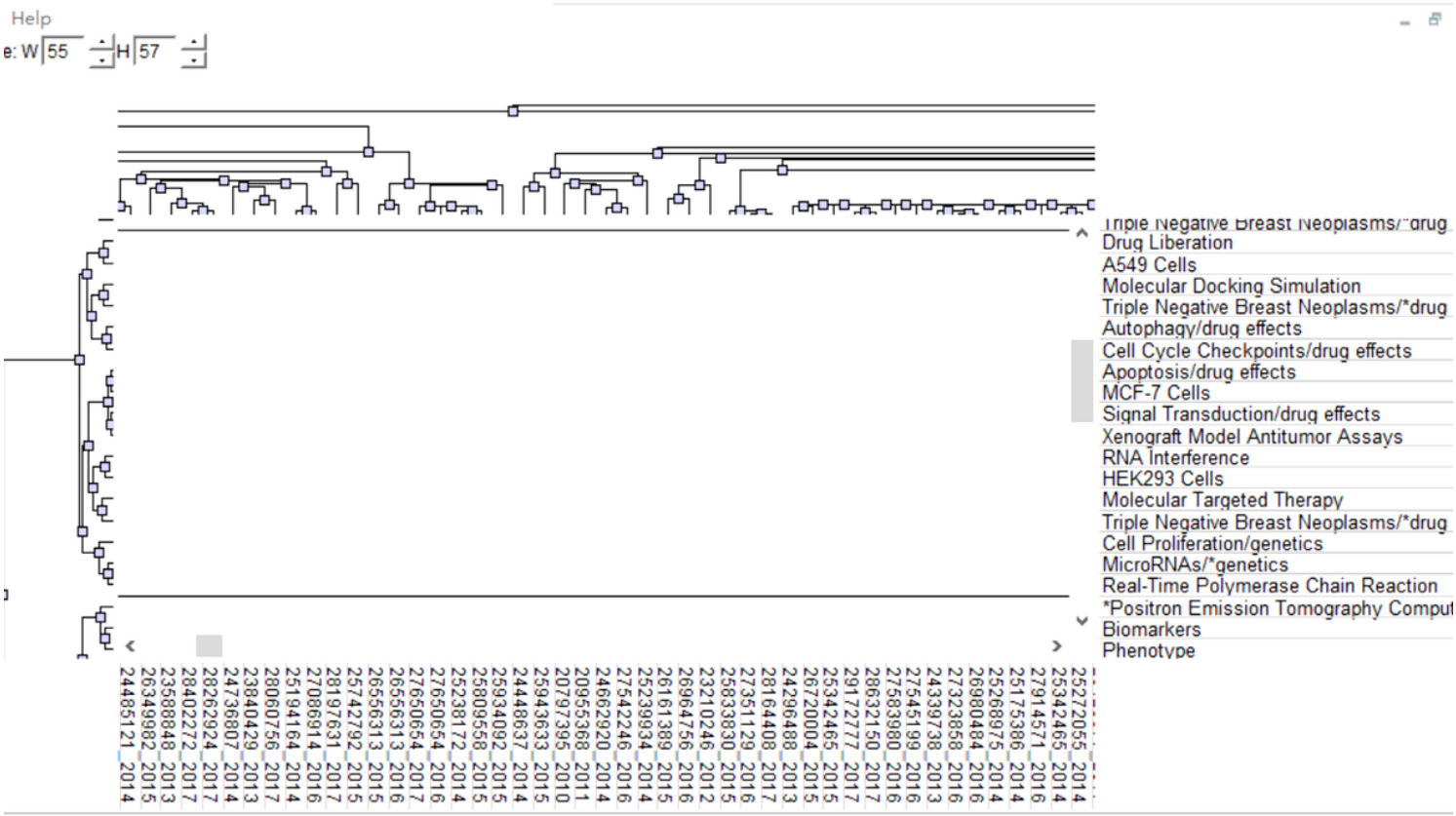

Figure 4

gCUTO double clustering result