Analyzing Research Fronts in Medical Informatics Field Based on Topic Model and Life Cycle Theory

Jia Feng
Jilin University School of Public Health https://orcid.org/0000-0001-9385-3253
Xiaomin Mu
Jilin University School of Public Health
Fangfang Li
Jilin University School of Public Health
Yong Shen
Jilin University School of Public Health
Wei Wang (w_w@jlu.edu.cn)
Ying Xu
University of Georgia

Research article

Keywords: Research fronts, Medical informatics, Life cycle theory, Topic model

Posted Date: January 29th, 2020

DOI: https://doi.org/10.21203/rs.2.22157/v1

Abstract

Background: Medical informatics (MI) is a multidisciplinary field in which researchers pursue scientific exploration, problem-solving, and decision-making to facilitate the effective use of biomedical data, information and knowledge for the improvement of human health. The purpose of this study is to identify research fronts in the field of MI and ultimately elucidate research activities and trends in this field. Methods: This study used topic model to identify research topics in the field of MI based on the latent Dirichlet allocation method (LDA). And the topic cloud is utilized to visualize the research topics. For identifying the research front topics, we proposed the indicators of identifying research front topics. In addition, we investigated how front topics change over time, and divided them into five categories based on the life cycle theory. Results: The data were collected from 35981 published journal abstracts between 2007 and 2016. In the topic distribution of MI, we found that the scope of MI related research has become increasingly interdisciplinary, particular for medical data analysis. Also, in the analysis of research fronts of MI, we found that the use of natural language processing and medical text knowledge extraction play an essential role for systematic analysis and indexing of the underlying semantic contents. Conclusions: By categorizing the research fronts, the results shows that there are twelve growing, five stable and two declining research fronts. We hope that this work will facilitate greater exploration of the method of identifying the research fronts. Moreover, the findings of this study provide an insight on the research fronts and trends in MI.

Background

Medical informatics (MI) is a multidisciplinary field in which researchers pursue scientific exploration, problem-solving, and decision-making to facilitate the effective use of biomedical data, information and knowledge for the improvement of human health[1]. The objective of this study is to identify and analyze the research fronts in the field of MI. Research fronts present the focus and difficulty domain of scientific research. Obtaining the research fronts timely and accurately is of great significance for the country, institutions and researchers.

There are various methods of detecting research fronts, which can be summarized from the qualitative and quantitative perspective. In the qualitative, literature review and brainstorm are commonly used and authoritative methods. Quantitative analysis is mainly based on bibliometrics. We can understand the research status by analyzing the academic literature of a certain field [2]. Numerous quantitative analysis methods have been utilized to study the research front issue, such as citation-related analysis [3–11] and content analysis [12–20]. With ample research results, citation analysis method started early, which can be concluded into co-citation analysis [5–7], bibliographic coupling [10, 11], and direct citation [9]. Content analysis is useful for understanding the knowledge in the literature. It includes term frequency analysis [15, 18], burst term analysis [12], and co-word analysis [17, 20].

The two methods are effective for detecting research topics. However, through the literature review, there are two main problems in the research of front identification:

 It's lack of recognizing indicators of research fronts, though the existing methods of detecting research fronts depend on the accumulation of terms or citation.

 The existing methods neglect the semantic information between texts. Citation analysis and content analysis method can't detect the research fronts semantically, and they neglect the semantic information between texts.

With the rapid development of machine learning technology, examining large collections of literatures can help researchers to understand crytic knowledge. Researchers have proposed some novel methods to detect research fronts, such as topic model, neural network, support vector machine (SVM), the decision tree. Among them, the topic model can extracts valuable potential topic distribution, through semantic analysis of the full text. The most widely applied topic model is latent Dirichlet allocation (LDA) method [21–23].

With purpose of identifying research fronts in the field of MI, this study applied a topic model to quantitatively investigate scientific articles published in 26 MI journals. In addition, to elucidate research activities and trends, the study explores the changes of research topics over time, and divided them into five categories based on the life cycle theory. We hope that this work will facilitate greater exploration of the method of identifying the research fronts.

Moreover, the findings of this study provide an insight on the research front topics and trends in MI.

Methods

Data collection We chose the Web of Science™ as our data source and adopted the Web of Science Category (WC) of MI to retrieve the literature. From 2007 to 2016, there were 35981 records, when WC = " Medical Informatics" (retrieved on January 3, 2017). We adopted abstracts as the analytical corpus. Because the abstract of an article, which is regarded as a condensed representation thereof, has been used to successfully identify and interpret the scientific themes of articles [24]. The topic model and parameters setting The LDA is a generative probabilistic model applicable to collections of discrete data. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words [25]. In fig.1, M is denoted as the total number of articles. N is the total number of words in a article and K is the number of topics. We can view LDA as a dimensionality reduction technique, but with proper underlying generative probabilistic semantics that make sense for the type of data that it models. We used the R package topic models [26] to perform LDA modeling employing the Gibbs sampling method. The LDA model has three key parameters, these are the Dirichlet hyperparameters α and β, and the number of topics K. The α can reflect the distribution degree of documents on each topic. The smaller the α, the better the discrimination among topics [27]. The value of α is related to the K value in the LDA model. The β can reflect the distribution degree of vocabulary in each topic. The value of β can affect model granularity, which means the smaller the β, the more topics. According to Griffiths (2004) and experience in the study, it suggested that the values are α= 0.1 and β= 50/K respectively [28]. Perplexity [29] is a standard measure of performance for statistical models of

natural language and is defined as exp{-log P(wtest|ϕ)/ntest}, where wtest and ntest indicate the identities and number of words in the test set, respectively. Perplexity indicates the uncertainty in predicting a single word; lower values are better. In this paper, K was set by the perplexity of the topic model, and when the value of perplexity is smallest, the value of K is optimal. Indicators of Identifying the Research Fronts This study hypothesized that the topics derived from LDA model, which contain many research front topics. So how can we detect them accurately and quickly? The paper proposed the strength and novelty of topics as the indicators. Strength can reflect the high attention by researchers, and novelty can reflect the latest studies. Strength of topic is the indicator to access the topics whether it can be hot topics. In the process of calculating the topic strength, the supporting documents of a topic is composed of the probability distribution of the documents. The probability value of a document can be understood as the strength of the document. In general, topic strength is the sum of probabilities of the overall supporting documents. The formula is following:

$$S_{\theta_k} = \frac{\sum_m \theta_k^{(m)}}{M} \qquad \text{Formula (1)}$$

$S\theta_k$ is the value of strength of one topic. M represents the total amount of documents. $\theta_k^{(m)}$represents the weight of topic k in the document m. The weights are more high, the more important. Hence, we can calculate the strength of the topic and plot the trends to illustrate the topic.

In order to analyze the novelty of the research topics quantitatively, the average publication time is an important identification index. The more one topic publishes recently, the higher degree of the novelty. Research fronts are the latest scientific findings and research issues, so they are more novel than other topics. For giving more weight to highly probability documents, this paper calculated the average publication time by capturing the top 50 documents in descending order of probability, and then calculated these average publication time. The formula is following:

$$Y_{\theta_k} = \frac{\sum_1^{50} \delta_k^{(m)}}{50} \qquad \text{Formula (2)}$$

Y_{θ_k} is the value of novelty of one topic. $\delta_k^{(m)}$ represents the year of document m in topic k. The strength and novelty of the topics were measured to determine whether they were research fronts.

Types of Research Fronts based on the life cycle theory

Braun et al. (2000) proposed the life cycle theory of the research field. He believed that the development process of a research field can be divided into five stages, including the incubation period, early period, expansion period, mature period and decline period [30]. The evolution of subject topics also follows this rule. With the development of scientific research, research fronts go through a process of development, maturity and decline. Based on the life cycle theory, this study classified fronts into emerging, growing, stable, declining and exiting types, which shown in Fig.2.

With the curve of literature, emerging fronts are newborn. In two adjacent time windows, a front topic does not appear in the previous time window (T1), but it appears in the current time window (T2), and the topic is considered an emerging front. On the opposite, exiting fronts are disappear in T2, which are appear in T1. To explore the trends in nascent research we measured the growth rates of fronts, from T1 to T2 and categorize them as growing, stable, or declining. Growing fronts are those that have more papers in our T2 period than the sum of all of their contributing fronts in the T1 period. Similarly, declining fronts are those that are smaller than the sum of all their contributing fronts in the previous time period. Moreover, stable fronts are those for which the sum of all contributing fronts changes little in the adjacent time windows.

Results

Overall topic distribution

Fig.3 shows the t perplexity versus number of topics. Following the increase in topic number, the perplexity of the model gradually decreases. When the number of topics is 62, the model has the lowest perplexity, which is the best fit.

The LDA model can produce document-topic distributions θ_m and topic-word distributions φ_k, which are difficult to understand intuitively the content of topics. Therefore, we utilize the wordclouds to visualize the each topic-word distribution φ_k, with the top 10 words in each topic. The sizes of words in wordclouds was proportional to probabilities yielded by the topic model.

We believe that each topic is a bag of words related to semantic content. In each topic, the words with higher probability reflect the content of the topic. Thus, we assign each topic a label or research subfield based on the high probability words (see Fig.4 and Fig.5).

For example, T1 is associated with words such as image, tumor, MRI, detection, CT, 3D… which are generally related to tumor image detection. T7 includes words such as sleep, assessment, respiratory, lung… these words are associated with the sleep monitoring analysis. We found some words such as information, model, medical, health… which were used in many topics.

Moreover, the topic word distribution can reveal more details about what a particular topic focus on. For instance, T11: Medical statistical model, which focus on reducing error of model. Analogously, T12： Model prediction, which focus on the parameter setting of model.

Identification of research fronts

Following the extraction of 62 topics, we calculated the strength of each topic, the results of which are shown in Fig.6.

In Fig.6, the height of the bar represents the strength of topics and the dotted line is the average of strength across the 62 topics. The average strength is 2.17. There are 22 topics with a strength value

higher than the average. Topic 4, which is “Medical text knowledge extraction,” has the highest strength value of the set.

With regard to novelty, we created a boxplot to visualize the values. In Fig.7, the boxplot contains one rectangle, a dotted line, and two borderlines. The spacing between the different parts of the box indicates the degree of dispersion spread and skewness in the data. In addition to the points themselves, they provide a visual estimate of various L-estimators, notably the interquartile range, mid-hinge range, mid-range, and trimean.

To calculate the novelty of the 22 topics that had strengths higher than the average. In view of these findings, topics 3, 4, 8, 19, 20, and 34 were more novel than the others were, because they remained active until 2015.

When we considered both the strength and novelty values, we found that research fronts had higher strength than the average and a publication time within the most recent 5 years of our study period. Using this definition, we identified 19 research fronts in MI.

Classification of research fronts

In this study, the corpus was divided into two time windows: T1 (2007-2011) and T2 (2012-2016). The topic growth rate was then calculated, and the results are shown in Figure 8. Each sphere in the figure represents a topic, and the size and height of the sphere represent the growth rate of the topic.

Table 1
The types of Research Fronts

Type	No	Topic label
Growing	1	Tumor image analysis
	3	Algorithm on medical data mining
	4	Medical text knowledge extraction
	8	Health care application
	19	New medical pattern based on web
	21	Disease classification method
	22	Medical system and software
	31	Health information system evaluation
	34	Disease survival model
	48	Computer-assisted diagnosis of disease
	54	Semantic analysis of clinical knowledge
	58	Medical big data platform
Stable	9	Community health service
	17	Clinical decision support
	41	Medical informatics methods and techniques
	45	Electronic medical records
	53	Machine learning algorithms in medicine
Declining	24	Medical data integration
	46	Disease risk prediction

Discussion

In the topic distribution of MI, we found that the scope of MI related research has become increasingly interdisciplinary, particular for medical data analysis. There are 16 topics concerned with medical data analysis. With the ability to deal with large volumes of both structured and unstructured data from different sources (T58), big data analytical tools (T41, T53, T3) hold the promise to study outcomes of large-scale population-based longitudinal studies, as well as to capture trends and propose predictive models (T12, T46) for data generated from electronic medical and health records (T45). A unique

opportunity lies in the integration of traditional MI with mobile health (T8) and social health (T59), addressing both acute and chronic diseases in a way that we have never seen before [31].

In the analysis of research fronts of MI, we found that the use of natural language processing and medical text knowledge extraction (T4) play an essential role for systematic analysis and indexing of the underlying semantic contents (T54). Mining electronics health records (EHRs) is a valuable tool for improving clinical knowledge and supporting clinical research (T17), for example, in discovering phenotype information.

More importantly, genetic test and analysis (T10) will help to screen out patients who are more likely to develop related disease [32]. Further studies on these high-risk patients (T46) based on tumor images analysis (T1) may provide insight into the rate of disease development. In order to assist disease diagnose (T48) and support clinical decision (T17), a combination of medical images as well as medical records (T45), demographics, and lab test results is key to characterize the structure, function, and progression of diseases (T24). This requires the implementation of effective and optimized querying systems (T22) [33] in order to reduce the computational complexity of handling these data. In addition, evaluation of health information system (T31) focuses on the quality management and evaluation of the system,such as those on the effect on health services [34, 35].

Our study has several limitations, mostly due to the complex nature of our research subject. One limitation is that we have limited ourselves by only considering sources, which are indexed by WoS. Therefore, our approach did not include a few MI journals, which are not covered by ISI. The LDA method itself imposes several limitations. The text mining method still depends on important choices of parameters; the attribution of labels to groups is also a matter of expert opinion, and needs substantial human intervention. The number of the topics can be chosen to be either smaller or larger. This depends on just how “loosely and generally” one wishes to define such a heterogeneous and complex field of study and application as MI. We set K value by perplexity of LDA model, however Chang et,al(2009) shown that surprisingly predictive likelihood (or equivalently, perplexity) and human judgment are often not correlated, and even sometimes slightly anti-correlated [36].

Conclusions

The goal of this work is to identify research fronts in the field of MI and ultimately elucidate research activities and trends in this field. Using quantitative methods to conduct in-depth mining of research fronts can provide valuable information for the interpretation of fronts. The method of categorizing fronts based on the life cycle theory emphasizes the novelty of fronts. In future work, we will integrating multi-source data to identify and analyze research topics using the LDA model and further optimize the parameters of the topic model.

Declarations

Acknowledge

Not applicable.

Authors' contributions

JF proposed the research idea, carried out the data analysis, and wrote the first draft. WW (corresponding author) designed the research program and contributed to revising the paper after review. XMM carried out the data processing and data analysis. FFL disposed the data collection and data cleaning. YS participated in the analysis of research findings, and contributed to writing the paper. YX provided expert consultation, and joined discussion of the findings. All authors read and approved the final manuscript.

Funding

This study did not receive support from any funding sources.

Availability of data and materials

The datasets generated and analyzed during the current study are available in the Web of Science.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Abbreviations

LDA: latent Dirichlet allocation; MI: Medical informatics; WC: Web of Science Category; EHRs: electronics health records;

References

1. Kulikowski CA, Shortliffe EH, Currie LM, Elkin PL, Hunter LE, Johnson TR, et al. AMIA Board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. Journal of the American Medical Informatics Association. 2012;19(6):931-8.
2. Xiong H, Cheng Y, Zhao WH, Liu JH. Analyzing scientific research topics in manufacturing field using a topic model. Computers & Industrial Engineering. 2019;135:333-47.
3. Fearon D, Hughes S, Brearley SG. Experiences of breast cancer in Arab countries. A thematic synthesis. Quality of Life Research.
4. Garfield E. Citation analysis as a tool in journal evaluation. Science (New York, NY). 1972;178(4060):471-9.
5. Hou J, Yang X, Chen C. Emerging trends and new developments in information science: a document co-citation analysis (2009-2016). Scientometrics. 2018;115(2):869-92.
6. Zhang T, Chi H, Ouyang Z. Detecting research focus and research fronts in the medical big data field using co-word and co-citation analysis[C]//2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE, 2018: 313-320.
7. Persson O. THE INTELLECTUAL BASE AND RESEARCH FRONTS OF JASIS 1986-1990. Journal of the American Society for Information Science. 1994;45(1):31-8.
8. Boyack KW, Klavans R. Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Research Front Most Accurately? Journal of the American Society for Information Science and Technology. 2010;61(12):2389-404.
9. Shibata N, Kajikawa Y, Takeda Y, Matsushima K. Detecting emerging research fronts based on topological measures in citation networks of scientific publications. Technovation. 2008;28(11):758-75.
10. Zhang L, Glanzel W, Ye FY. The Dynamic evolution of core documents: an experimental study based on h-related literature (2005-2013). Scientometrics. 2016;106(1):369-81.
11. Liu JS, Lu LYY, Lu W-M. Research fronts in data envelopment analysis. Omega-International Journal of Management Science. 2016;58:33-45.
12. Chen CM. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology. 2006;57(3):359-77.
13. Liu C, Lu X. Analyzing hidden populations online: topic, emotion, and social network of HIV-related users in the largest Chinese online community. Bmc Medical Informatics and Decision Making. 2018;18.

14. Rombach I, Rivero-Arias O, Gray AM, Jenkinson C, Burke O. The current practice of handling and reporting missing outcome data in eight widely used PROMs in RCT publications: a review of the current literature. Quality of Life Research. 2016;25(7):1613-23.
15. Khanh-Ly N, Shin B-J, Yoo SJ, Ieee. Hot Topic Detection and Technology Trend Tracking for Patents utilizing Term Frequency and Proportional Document Frequency and Semantic Information. 2016 International Conference on Big Data and Smart Computing. International Conference on Big Data and Smart Computing2016. p. 223-30.
16. Ngamaba KH, Panagioti M, Armitage CJ. Income inequality and subjective well-being: a systematic review and meta-analysis. Quality of Life Research. 2018;27(3):577-96.
17. Wei W-J, Shi B, Guan X, Ma J-Y, Wang Y-C, Liu J. Mapping theme trends and knowledge structures for human neural stem cells: a quantitative and co-word biclustering analysis for the 2013-2018 period. Neural Regeneration Research. 2019;14(10):1823-32.
18. Wang F, Jia X, Wang X, Zhao Y, Hao W. Particulate matter and atherosclerosis: a bibliometric analysis of original research articles published in 1973-2014. BMC Public Health. 2016;16.
19. Price-Haywood EG, Harden-Barrios J, Carr C, Reddy L, Bazzano LA, van Driel ML. Patient-reported outcomes in stroke clinical trials 2002-2016: a systematic review. Quality of Life Research. 2019;28(5):1119-28.
20. Liu H, Yu Z, Chen C, Hong R, Jin K, Yang C. Visualization and Bibliometric Analysis of Research Trends on Human Fatigue Assessment. Journal of Medical Systems. 2018;42(10).
21. Li Y, Rapkin B, Atkinson TM, Schofield E, Bochner BH. Leveraging Latent Dirichlet Allocation in processing free-text personal goals among patients undergoing bladder cancer surgery. Quality of Life Research. 2019;28(6):1441-55.
22. Boonyasopon P, Riel A, Uys W, Louw L, Tichkiewitch S, du Preez N. Automatic knowledge extraction from manufacturing research publications. Cirp Annals-Manufacturing Technology. 2011;60(1):477-80.
23. Park K, Kremer GEO. Text mining-based categorization and user perspective analysis of environmental sustainability indicators for manufacturing and service systems. Ecological Indicators. 2017;72:803-20.
24. Sun LJ, Yin YF. Discovering themes and trends in transportation research using topic modeling. Transportation Research Part C-Emerging Technologies. 2017;77:49-66.
25. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research. 2003;3(Jan):993-1022.
26. Grun B, Hornik K. Topicmodels: An R Package for Fitting Topic Models. Journal of Statistical Software. 2011;40(13):1-30.
27. Lu Y, Mei Q, Zhai C X. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA[J]. Information Retrieval, 2011, 14(2): 178-203.
28. Griffiths TL, Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America. 2004;101:5228-35.

29. Lee L. Foundations of statistical natural language processing. Computational Linguistics. 2000;26(2):277-9.
30. Braun T, Schubert AP, Kostoff RN. Growth and trends of fullerene research as reflected in its journal literature. Chemical Reviews. 2000;100(1):23-37.
31. Andreu-Perez J, Poon CCY, Merrifield RD, Wong STC, Yang G-Z. Big Data for Health. Ieee Journal of Biomedical and Health Informatics. 2015;19(4):1193-208.
32. den Hartog AW, Franken R, Zwinderman AH, Timmermans J, Scholte AJ, van den Berg MP, et al. The Risk for Type B Aortic Dissection in Marfan Syndrome. Journal of the American College of Cardiology. 2015;65(3):246-54.
33. Aji A, Wang F, Saltz J H. Towards building a high performance spatial query system for large scale medical imaging data[C]//Proceedings of the 20th International Conference on Advances in Geographic Information Systems. 2012: 309-318.
34. Qin Y, Zhou R, Wu Q, Huang X, Chen X, Wang W, et al. The effect or nursing participation in the design of a critical care information system: a case study in a Chinese hospital. Bmc Medical Informatics and Decision Making. 2017;17.
35. Stutzel MC, Filippo MP, Sztajnberg A, da Costa RMEM, Brites AdS, da Motta LB, et al. Multi-part quality evaluation of a customized mobile application for monitoring elderly patients with functional loss and helping caregivers. Bmc Medical Informatics and Decision Making. 2019;19.
36. Chang, J., J. Boyd-Graber, C. Wang, S. Gerrish, and D. M. Blei (2009). Reading tea leaves:How humans interpret topic models. InNIPS.

Figures

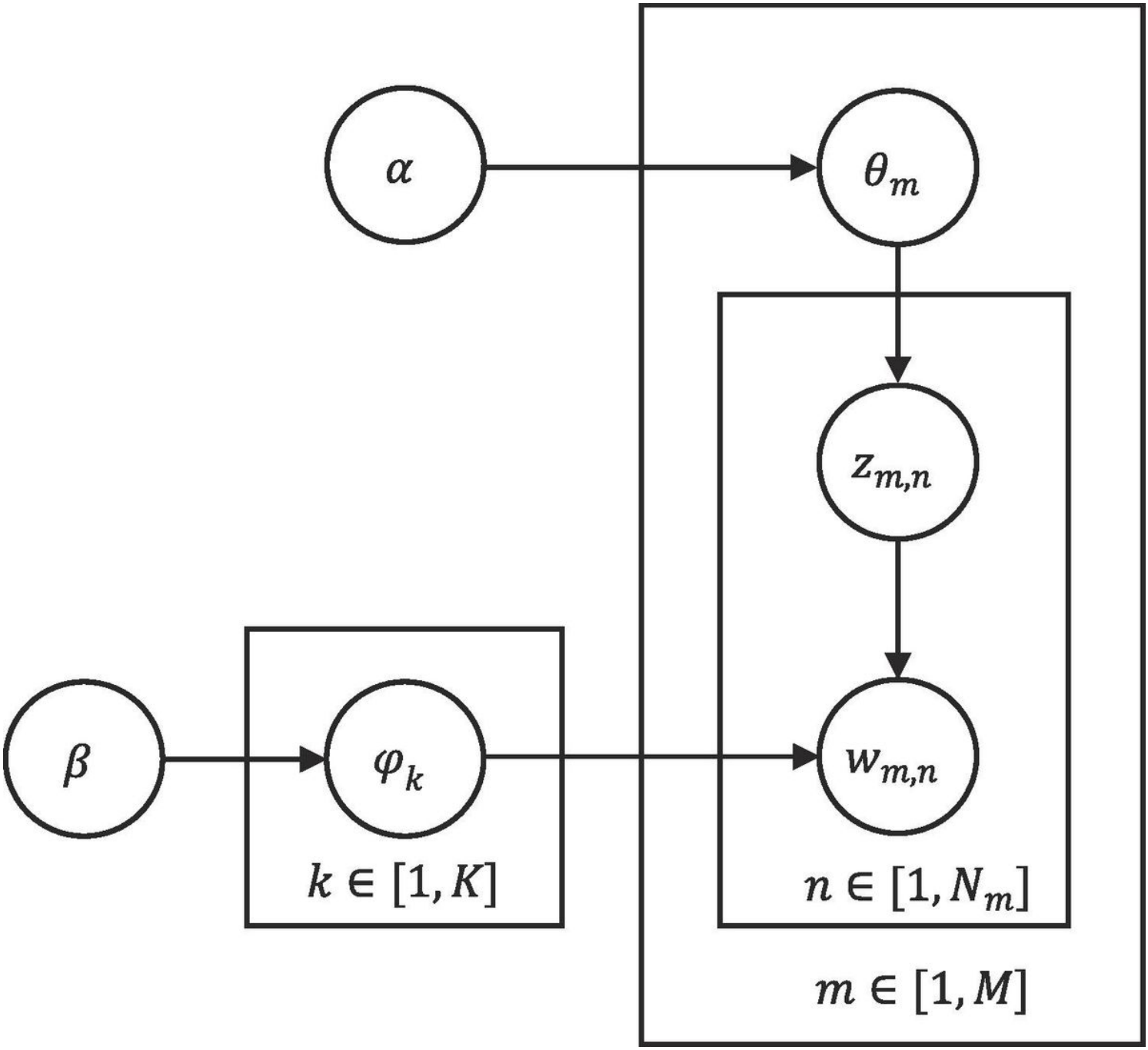


Figure 1

A graphical model of the LDA

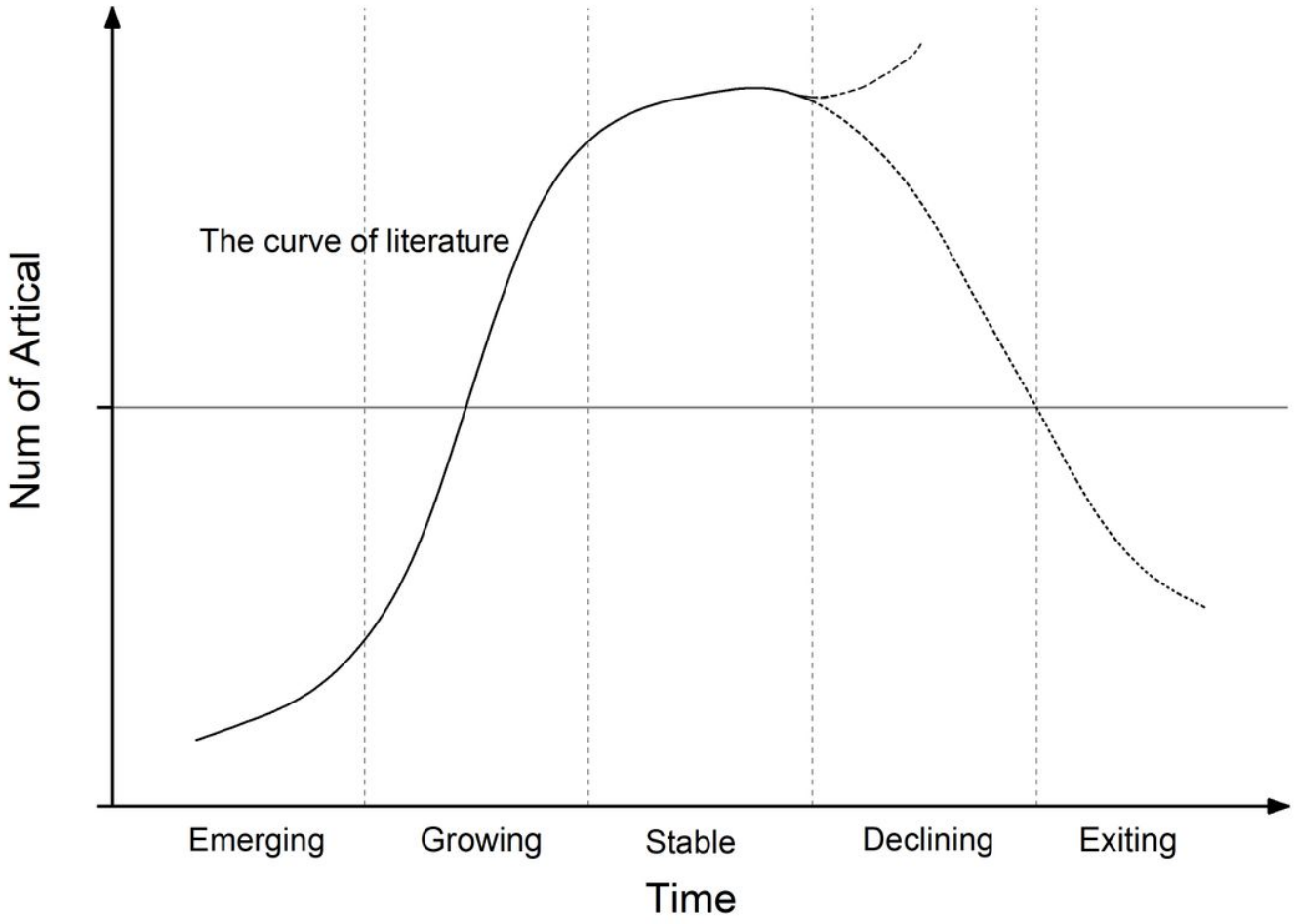


Figure 2

Research fronts classification based on the life cycle theory.

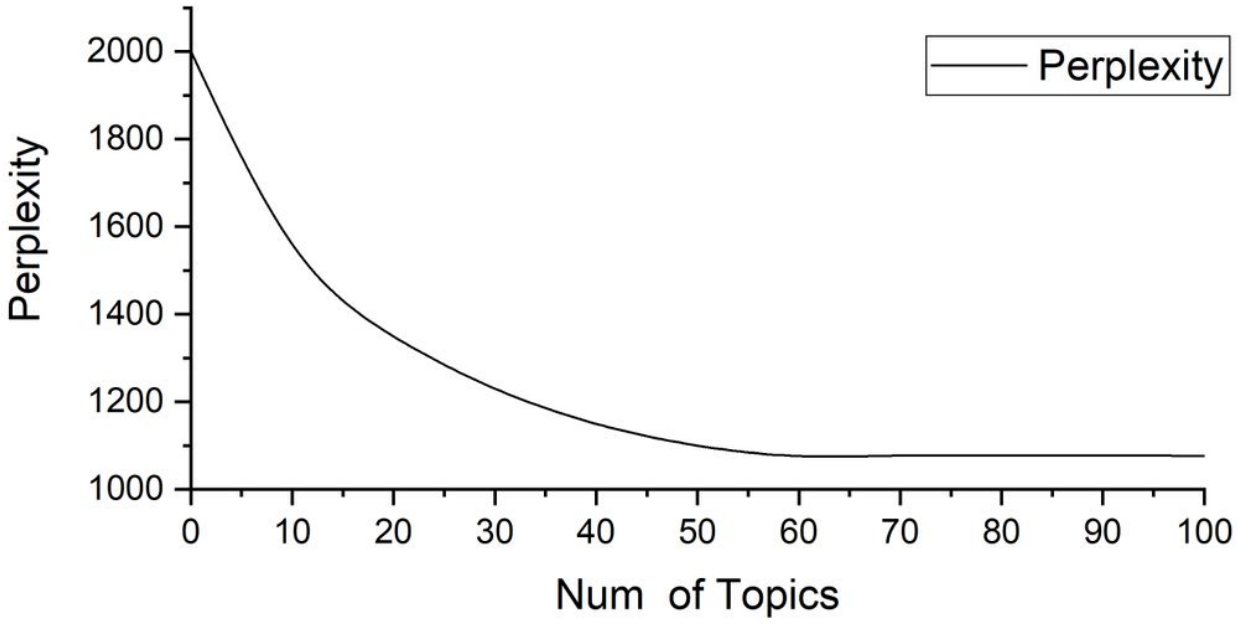


Figure 3

Perplexity versus number of topic

Figure 4

Word clouds for Topic 1-30.

T31: Health information system evaluation

T32: Biomedical informatics

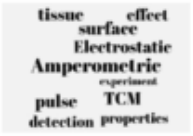
T33: Neural signal analysis

T34: Disease survival model

T35: Sensor wireless communication

T36: Care for patients

T37:Healthcare facilities



T38:Electrochemical information analysis

T39:Traditional Chinese medicine analysis

T40: Efficacy evaluation

T41:Medical informatics methods and techniques

T42: Privacy protection for healthy users

T43: Maternal and child health care

T44: Online (electronic) medical education

T45: Electronic health records

T46:Disease risk prediction

T47: Blood pressure monitoring

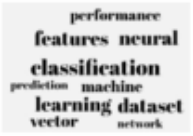
T48: Computer-assisted diagnosis of disease

T49:Electrophysiological information analysis

T50:Clinical experimental design

T51: Motion detection and analysis

T52: Gene expression and disease



T53:Machine learning algorithms in medicine

T54: Semantic analysis of clinical knowledge

T55: Medical ontology research

T56: Robotics in surgery

T57: Medical statistics

T58: Medical big data platform

T59: Behaviors in the context of social media

T60:Electroencephalo -gram analysis

T61: Medical diagnostic system

T62: Epidemiological surveillance analysis

Figure 5

Word clouds for Topic 31-62.

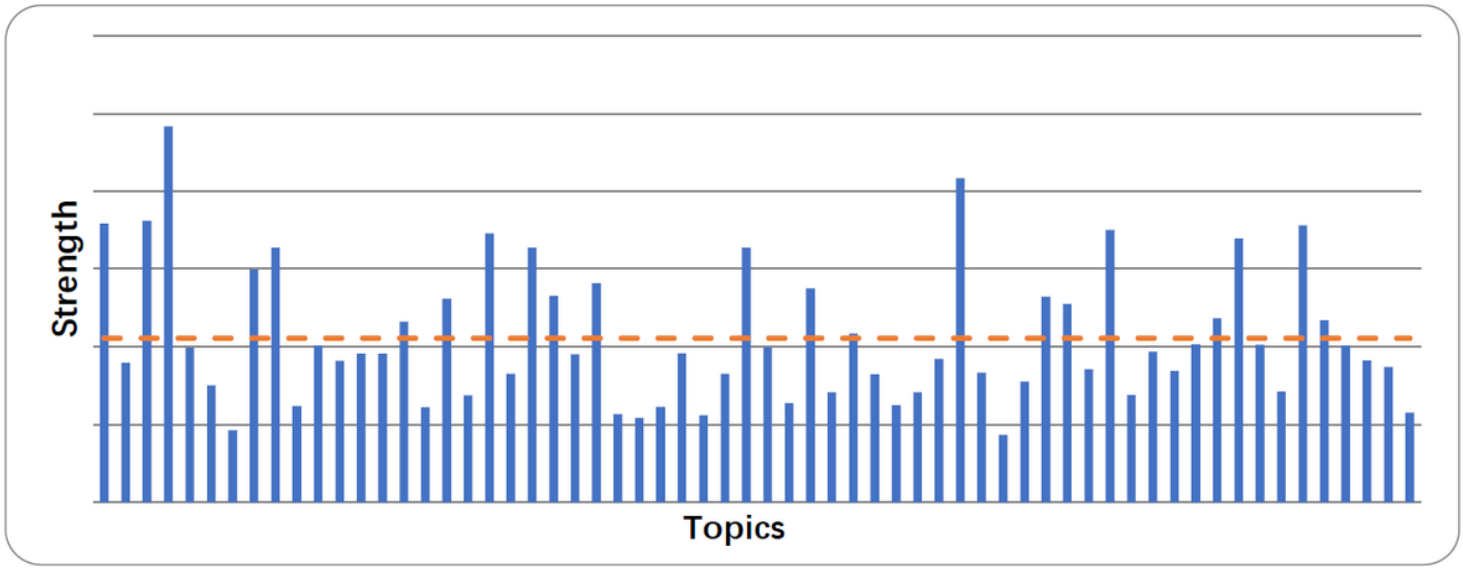


Figure 6

Strength of topics.

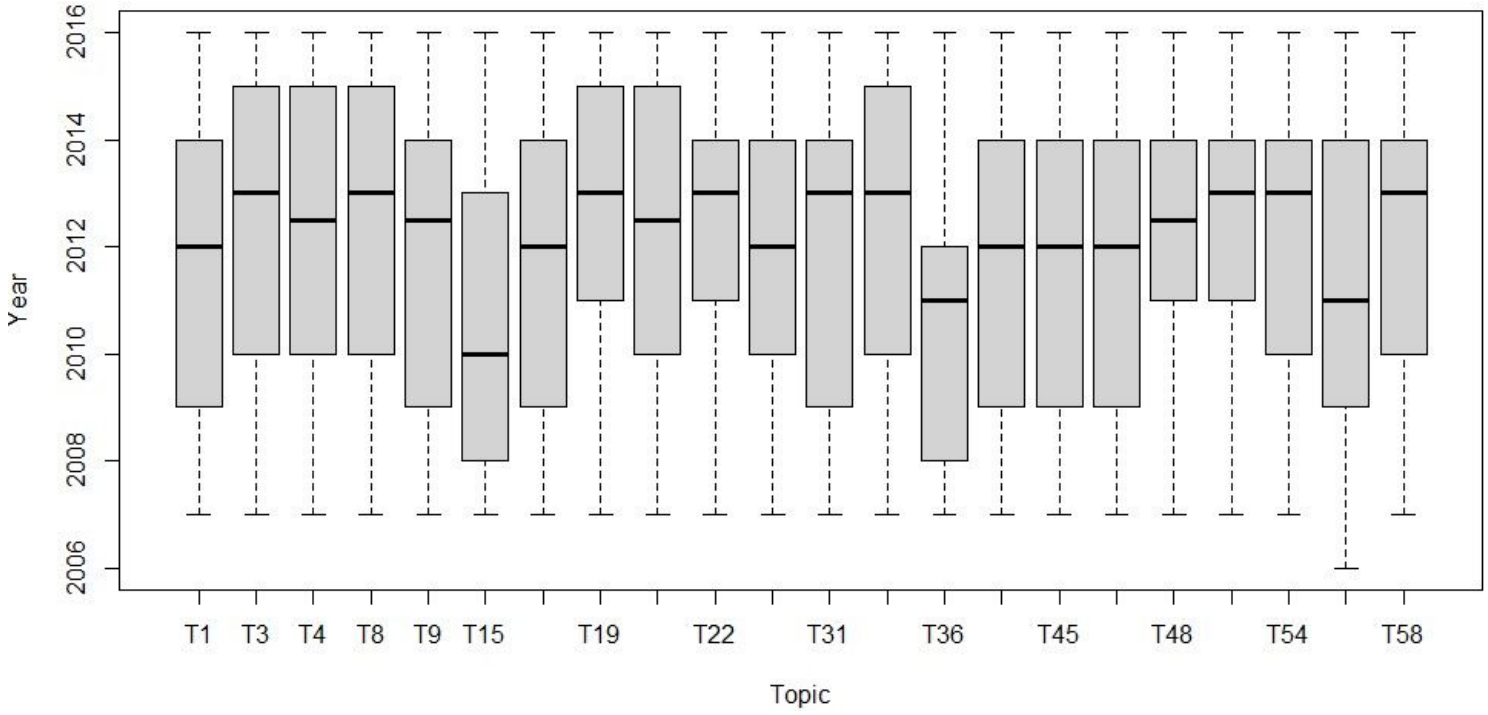


Figure 7

Boxplot of novelty.

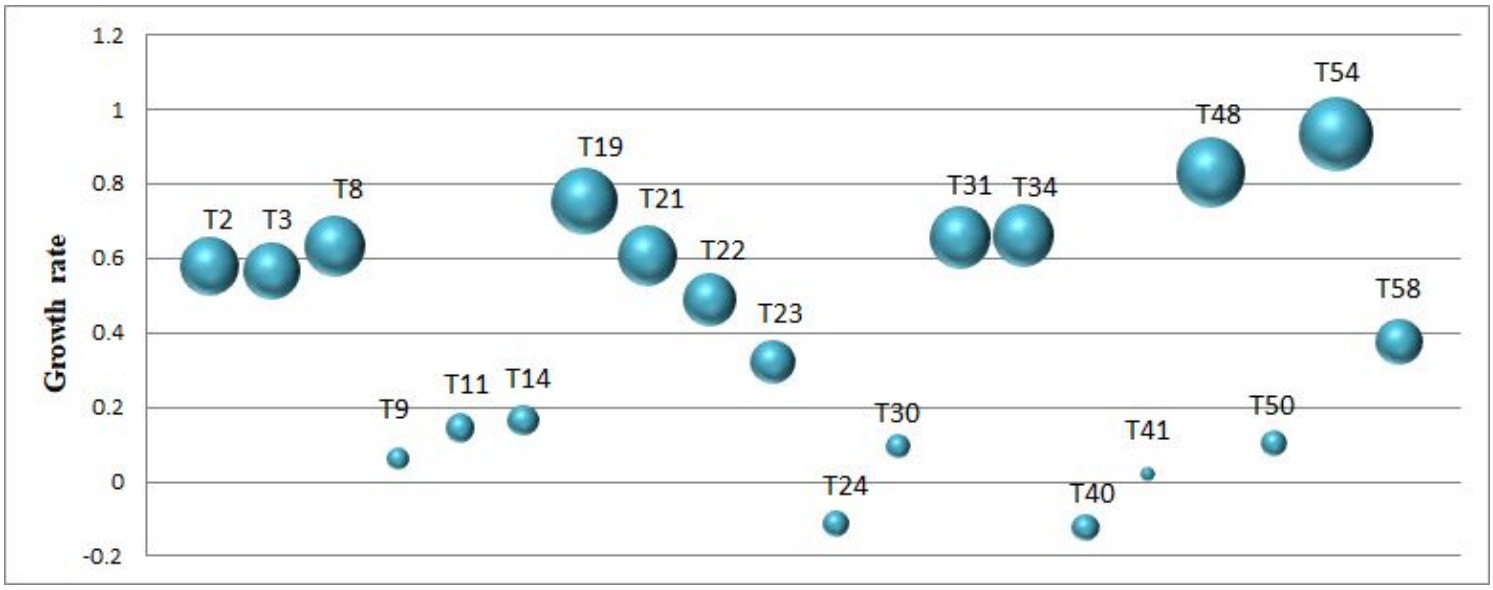


Figure 8

The growth rate of research fronts.