Author Topic Model and Its Improvements and Applications: A Survey (Postprint)

Authors: Xu Han, Liu Xiaoping

Date: 2023-07-26T00:00:00+00:00

Abstract

[目的/意义]The Author-Topic Model, as a novel probabilistic model that has attracted considerable attention in the computer science field in recent years, has been widely applied in text mining, natural language processing, and related areas. By analyzing the improvement approaches and applications of domestic and international Author-Topic Models, this study seeks to better comprehend the current research status and provide references for researchers in computer science, library and information science, and related fields.[方法/过程]This paper selects the Web of Science Core Collection, DBLP, and CNKI (China National Knowledge Infrastructure) databases as literature sources. Through procedures including search strategy formulation, deduplication, and manual screening, a literature corpus on Author-Topic Models and their improvement methods is distilled. From the perspective of model application processes, existing research is summarized and synthesized employing literature analysis methodology.[结果/结论]Through analysis, it is revealed that existing relevant research has established a relatively complete analytical workflow, and the improvement perspectives and applicable domains of the models are becoming increasingly diversified. However, aspects such as performance optimization, standardization and refinement of model evaluation metrics, and further application within the library and information science field remain to be thoroughly explored.

Full Text

Preamble

A Review of Author-Topic Models and Their Improved Methods and Applications

Xu Han[1,2], Liu Xiaoping[1,2]
[1]National Science Library, Chinese Academy of Sciences, Beijing 100190
[2]Department of Library, Information and Archives Management, School of Eco-

nomics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract

[**Purpose/Significance**] The Author-Topic (AT) model, as a novel probabilistic model that has attracted considerable attention in computer science in recent years, has been widely applied in text mining and natural language processing. This paper analyzes the ideas and applications of AT models and their improvements both domestically and internationally to better grasp their research status, aiming to provide references for researchers in computer science, library and information science, and related fields. [**Method/Process**] This study selected the Web of Science Core Collection, DBLP, and CNKI (China Academic Journals Full-text Database) as data sources. Through the establishment of retrieval rules, deduplication, and manual screening, a literature collection on AT models and their improved methods was constructed. From the perspective of the model application process, combined with literature analysis, existing research was summarized and 归纳. [**Result/Conclusion**] The analysis reveals that existing research has formed a relatively complete analytical workflow, with increasingly diversified improvement angles and application domains for the models. However, further in-depth exploration is still needed in areas such as performance optimization, standardization and improvement of model evaluation metrics, and expanded applications in library and information science.

Classification Number: G250 TP391
Keywords: Author-Topic model, topic evolution, community detection, model evaluation

Introduction

The rapid development of computer technology and the accelerated connectivity of communities through social media, coupled with diversified data types and exponential growth in data volume, have led researchers to no longer be satisfied with using single topic information to represent dataset characteristics in large-scale data mining. Instead, they seek to uncover the relationships between topics and users to more comprehensively reveal information contained in datasets and discover underlying patterns. To this end, researchers proposed in 2004 a statistical model that could mine implicit author-topic relationships in document collections—the Author-Topic (AT) model [**?**]. The AT model extends Latent Dirichlet Allocation (LDA), inheriting LDA's ability to map high-dimensional word sets to low-dimensional topic space for dimensionality reduction while incorporating author information as extended metadata into the original model. This establishes associations between authors and topics in datasets, enabling the mining of implicit "author-topic" semantic information and helping researchers better match authors with their discussed content.

As the application scope of the model continues to expand, the diversification of tasks has made the simple AT model insufficient to meet growing demands, such as the need to consider temporal attributes in social media data analysis. Leveraging the excellent extensibility of the AT model, researchers have attempted various improvements, achieving breakthroughs that have broadened its applicability. Through preliminary literature investigation, we found that most existing review literature summarizes topic models in general, but with the development of topic models, the resulting model collection has become very broad, with few reviews focusing specifically on the "author-topic" lineage. Most reviews primarily summarize models' principles and improvements chronologically, while reviews from the perspective of specific application processes are relatively rare. Therefore, this paper, based on existing domestic and international research on AT models, employs literature analysis and 归纳 methods to deeply analyze the research status of AT models and their improved methods and applications, aiming to provide references and broaden research ideas for relevant researchers.

2 Data Sources and Overview

2.1 Data Sources

To ensure comprehensive coverage, this paper selected the Web of Science Core Collection and DBLP (Digital Bibliography & Library Project) as authoritative databases in the computer science field for foreign literature, in addition to CNKI for Chinese literature. Although there is overlap between the first two databases, they remain complementary. The specific data acquisition process was as follows: (1) In CNKI, professional retrieval was conducted using the query "Subject='author' AND ('topic model'+'LDA model') OR SU='author topic model' OR SU='Author Topic model' ", with literature types limited to journals, conferences, and dissertations. In Web of Science Core Collection, advanced retrieval was performed using the query "Topic=('Author*' AND ('Topic Model*' OR 'Latent Dirichlet Allocation'))", with literature types limited to Article, Proceedings Paper, and Review. In DBLP, retrieval was conducted using "Author Topic Model" as the subject term. All searches had no time span restrictions, with the retrieval date being May 25, 2018, yielding 174 Chinese documents and 360 foreign documents. (2) The three datasets were cleaned to remove invalid literature and duplicates. (3) Manual screening was conducted to select literature matching the topic, resulting in 139 documents (96 foreign and 43 Chinese) as the focus of this analysis.

2.2 Data Overview

Among the 139 research papers on AT models and their improvements, there were 3 review papers, 46 methodology research papers, 83 application papers, and 7 papers that used AT models as control groups in experiments. [Figure 1: see original paper] shows the chronological distribution of the 139 papers. As seen in [Figure 1: see original paper], after the AT model was proposed in 2004,

domestic research on this model with published outputs began in 2008 [?, ?]. In terms of publication trends, attention has generally increased, with two small peaks in attention occurring during 2013-2017. Based on these data, it can be predicted that research based on AT models will maintain high popularity for some time to come.

3 Research Progress Analysis

The analysis workflow for AT model improvement and application research [?] is as follows (see [Figure 2: see original paper]): First, determine the original dataset and data type to be analyzed; then perform text preprocessing; next, model and solve the preprocessing results; finally, select evaluation metrics and assess the model through comparative experiments. This paper will summarize and analyze existing research methods based on this workflow.

3.1 Dataset Selection

Through summary and 归纳, datasets selected for AT model improvement and application research both domestically and internationally are mainly divided into six categories (see). Among them, the "academic papers" dataset type ranks first (17 categories total), which can be further subdivided into five subcategories: (1) professional field databases (DBLP, PubChem, MEDLINE, PubMed Central, ACM Digital Library); (2) citation index systems (Scopus, CiNii, NIPS, CiteULike, Citeseer, Wanfang, CNKI, Web of Science); (3) preprint repositories (arXiv); (4) academic search engines (Arnetminer, Microsoft Academic Research); and (5) evaluation conferences and corpora (TREC). In addition to academic papers, text datasets also include social websites and emails, while image datasets include both open-access data and non-open-access resources such as school surveillance videos.

** Dataset Type Distribution**

Dataset Name	Type	Examples
Text Datasets	Academic Papers	DBLP[?], PubChem[?], MEDLINE[?, ?], Scopus[?], CiNii[?], NIPS[?], CiteULike, Citeseer[?], PubMed Central, Wanfang[?], arXiv[?], Web of Science[?], CNKI[?], Arnetminer, ACM Digital Library, Microsoft Academic Research[?, ?]

Dataset Name	Type	Examples
	Social Websites	perverted-justice.com[?], delicious.com, IRC logs, ProgrammableWeb, Tianya Forum[?], mooc.guokr.com[?], NLPIR, Twitter[?], digg[?], Enron[?]
Image Datasets	Open Access	Flickr[?], VIPeR[?], BrainMap, Yahoo, Wikipedia[?], Tripadvisor[?], CASIA, Weizmann, DECODA[?, ?], Quickbird, LIVEIQA, Tripadvisor, Google Earth[?]
	Non-Open Access	Hospital databases, School surveillance videos[?]

Note: This classification is based on the types of data objects extracted in AT model-related experimental studies, not all data types in the datasets. For example, Flickr itself is a social platform containing both image and text data, but in AT model experiments, its image data is typically selected for analysis, so it is classified under the image category rather than text.

After determining the dataset, target data fields need to be extracted. Academic paper datasets typically include titles, authors, abstracts, and time. Depending on the research content, researchers also extract fields such as citations, journals, and conference information to study shifts in author research interests[?], academic paper recommendation[?], expert community discovery in academic fields[?], and author-topic identification in specific disciplines[?]. For image datasets, specific resolution images and video clips are selected to achieve image and video topic recognition[?] and to mine spatial correspondence relationships of brain cognitive functions[?]. For social website datasets, tweet texts are typically extracted to mine user interests[?]. For email datasets, fields such as recipients, subject, content, and time are usually extracted to achieve associations between senders and topics/time[?].

3.2 Data Cleaning

After obtaining target data fields, data preprocessing is required. Among these steps, author disambiguation significantly impacts experimental results. Author disambiguation currently faces two major challenges: first, the name ambiguity problem, where multiple people share the same name; second, the coreference problem, where the same author has multiple name variations, which commonly occurs in English literature and is associated with name abbreviation issues. In practice, disambiguation is mostly performed by combining attribute features such as author email, affiliation, and co-authorship relationships[**?**], while external network information can further improve disambiguation accuracy[**?**]. Specific methods can be divided into unsupervised and supervised learning approaches. Unsupervised methods include graph theory and clustering: the former constructs network graphs to connect relationships among authors, calculating topological distances between nodes to determine whether authors with the same name are the same person; the latter measures similarity to cluster all possible same-name authors pointing to the same person, where the choice of clustering method and definition of similarity functions among same-name authors are key to clustering effectiveness[**?**]. Supervised learning methods mainly refer to probability model-based approaches, which typically require building complex probabilistic models such as Bayesian networks or conditional random fields to infer matching relationships among authors with the same name through statistical calculations.

3.3 Model Construction

After completing data cleaning and other operations, an appropriate model is selected or constructed based on research content and objectives, and model parameters are estimated. This section summarizes current domestic and international AT models and their improvements.

3.3.1 Simple AT Model M. Steyvers et al. proposed in 2004 a method to mine author-topic associations from document collections, namely the Author-Topic model[**?**], which simultaneously models topics and authors by introducing author information into the LDA model. This model incorporates author information of analysis objects into the LDA model to achieve semantic topic associations among words, topics, authors, and documents. On the word layer, each topic can be described through a multinomial distribution, while documents can be modeled through author mixture distributions in topic space[**?**]. The generative process of the AT model is as follows[**?**]:

For each document $d \in D$ in a document collection D: 1. Draw θ (author-topic probability distribution) $\sim$ Dirichlet(α); 2. Draw β (topic-word probability distribution) $\sim$ Dirichlet(b); 3. For each word w in d: (a) Select an author x from the author set a_d of d using uniform distribution; (b) Assign a topic z according to the author-topic probability distribution $x \sim \text{Multi}(x, \theta)$; (c) Assign a word w according to the topic-word probability distribution $z \sim \text{Multi}(z, \beta)$.

From this generative process, each author in the AT model corresponds to a distribution over topics, and all authors share a common set of topics[?].

3.3.2 Extended AT Models Although the introduction of author information transforms the LDA model from unstructured to structured information, enabling simultaneous analysis of user interest distributions and document structures, several limitations remain[?]. In recent years, with the development of text mining technology, researchers have conducted a series of transformations and extensions. Below, we summarize AT model-based improvements from four aspects: time factor incorporation, supervised learning methods, metadata extension, and task-specific adaptations.

(1) Time Factor-Based Improvements. As research on AT models deepens, how to capture temporal changes in authors' topics has become a research hotspot. Time-based improvements can be further divided into two categories from a data temporal perspective: treating time as a random variable for continuous-time modeling, and discretizing time into timestamps for dynamic Bayesian network modeling.

- **Continuous-Time Modeling.** In 2010, Tang et al. introduced time as a continuous variable into the AT model, proposing the Author-Topic-Time (ATT) model[?]. This model combines AT with the Time-over-Topic (TOT) model, where word generation in documents is jointly determined by topic and time attributes, better describing author-topic distributions across different times. As a typical representative of continuous-time modeling, many subsequent studies on time-factor-based author-topic mining were developed based on this[?]. By considering dependencies between topics and the sparsity of user interest data, models such as the Hidden Markov Author-Time-Topic (HMATT) model[?] and the Author-Conference-Topic-Over-Time (ACTOT) model[?] were proposed, enriching user interest matrices and reducing model perplexity.

Although ATT and ACTOT models consider temporal information to represent the distribution intensity of author-topics across time, they still have two problems: First, the number of topics within each time window is fixed, thus only revealing changes in topic intensity while ignoring changes in topic content[?]. Second, for large document collections with frequent temporal changes, the models consume substantial computational and memory resources.

- **Discrete-Time Modeling.** To address ATT's inability to model dynamic relationships between time periods, discrete-time modeling emerged. Represented by Yang Ruyi's Dynamic Author Topic (DAT) model proposed in 2015[?] and Yu Chuanming et al.'s composite topic evolution model (Author-Topic-Time-LDA with Author Ranking, ATT-LDA) proposed in 2018[?], these models typically combine the advantages of Dynamic Topic Models (DTM) and AT models, discretizing topics into time windows. This offers significant advantages in both computational

complexity and application scenarios. Experiments have proven that such models can accurately describe latent author-topics and their dynamics.

(2) Supervised Learning-Based Improvements. Most current AT model extensions are unsupervised models that only require input of document collections and topic numbers for automatic learning. However, unsupervised learning results often have poor interpretability and are difficult to understand. To address this issue, H. Mou et al. proposed the supervised Author-Subject-Topic (AST) model in 2015[**?**], introducing a supervised "subject" layer to group documents, which helps cluster words and documents to reduce noise.

(3) Metadata Element-Based Extensions. Introducing richer and more diverse metadata elements transforms the traditional "author-topic" distribution into an "other metadata-author-topic" distribution, further expanding the scope of information that AT models can reveal. Metadata-based improvements currently fall into four categories: (1) Incorporating community elements, such as the Author-Topic-Community (ATC) model proposed by C.S. Li et al. in 2012[**?**, **?**], which combines social network analysis to achieve author community discovery based on author interests; (2) Incorporating conference elements, such as the Author-Conference-Topic (ACT) model proposed by Tang et al. in 2008[**?**] and the ACTC model proposed by Y. Ding in 2011[**?**], which dynamically set topic numbers corresponding to different conferences; (3) Incorporating user interest and other user characteristic elements[**?**, **?**, **?**], enabling better application to personalized recommendation tasks; and (4) Incorporating author citation and related elements, including citations and journal information[**?**], allowing models to fully utilize author and citation information in academic documents for better author discrimination and citation ranking.

(4) Task-Specific Improvements. To enable AT models to meet specific tasks in particular domains, researchers have proposed task-specific AT models. These models currently target mostly image data. The most representative ones incorporate geographic information (Author-related Geographic Topic Modeling, AGTM)[**?**] and color/shape features of multispectral remote sensing images (Author-Genre Topic Model, AGTM)[**?**] for topic annotation of region categories. However, how to combine geographic information with better probabilistic models to process document information and improve image category annotation accuracy still requires further exploration.

The four types of extended models described above mainly address existing problems in traditional AT models from different angles. [Figure 3: see original paper] summarizes the extension patterns of various models, primarily divided into two extension modes based on LDA and AT models, where solid arrows indicate model improvements and dashed arrows indicate combinations with other models. provides a summary of the extended AT models described above, analyzing the advantages, disadvantages, and other information of each model.

3.4 Model Evaluation

Measuring whether results truly reveal data patterns and whether they are optimized compared to existing research is essential throughout the experimental process. As scientific research advances, model evaluation perspectives have become more comprehensive. Investigation reveals that common evaluation methods in AT model research are conducted from six perspectives, with specific evaluation metrics shown in [Figure 4: see original paper].

3.4.1 Model Generalization Ability. This refers to the trained model's prediction ability on unknown data, quantified through test error in practice. The most commonly used metric is perplexity. In our constructed dataset, Q.Q. Yang and L. Poddar et al. used this metric to quantify generalization ability, where perplexity represents the difficulty of topic word prediction—lower values indicate better generalization[**?**, **?**]. Additionally, Wan Lukang et al. used P@n (Precision@N) to evaluate generalization ability, detecting the accuracy of top-N prediction results[**?**]. In practice, P@5, P@10, and P@20 are commonly selected.

3.4.2 Topic Interpretability. This analyzes the semantic relevance of word distributions in topics, but experiments prove this evaluation perspective is highly subjective. To evaluate more objectively, some scholars calculate PMI (pointwise mutual information) values proposed by K.M. Schneider et al. for quantification, where PMI values measure the co-occurrence probability of event pairs[**?**].

3.4.3 Efficiency. Most researchers evaluate model efficiency through complexity[**?**], which refers to the resources required during algorithm runtime, including time complexity and space complexity. However, complexity-based evaluation cannot avoid the problem that low complexity does not guarantee good topic words at the semantic level. Therefore, H.M. Wallach et al. proposed two different evaluation dimensions including effectiveness and correctness[**?**].

3.4.4 Effectiveness. This is a quantitative representation of how well model output structures match characteristics of real-life systems. C.S. Li et al. used sensitivity to characterize the degree of response change caused by unit change in variables[**?**, **?**]. Additionally, W. Buntine improved H.M. Wallach's left-to-right algorithm, proposing a novel method for estimating document collection likelihood to make model evaluation more fair[**?**].

3.4.5 Correctness. This includes precision, recall, and related comprehensive metrics. In our dataset, Wang Yonggui et al. used precision, recall, and the F1-Measure combining both to quantitatively analyze correctness[**?**]. X. Xie et al. used the AUC (Area Under the ROC Curve) metric, where larger area under the ROC curve (with true positive rate on the y-axis and false positive rate on the x-axis) indicates higher accuracy[**?**]. J. Wang et al. used Mean Average Precision (MAP) to evaluate correctness, where average precision for a single topic is the mean of precision values after each relevant document retrieval[**?**].

3.4.6 Relevance. This evaluation perspective typically relies on specific tasks

to indirectly assess models, using quantifiable concepts like distance to reflect topic similarity between documents or authors. For instance, C.S. Li et al. used KL divergence (Kullback-Leibler Divergence)[**?**], T. Wang et al. used similarity measures[**?**], and T. Zhang et al. used Spearman rank correlation coefficients and Pearson correlation coefficients[**?**]. Spearman rank correlation assumes that if data has no repeated values and two variables are perfectly monotonically correlated, the value is +1 or -1, while Pearson correlation coefficients typically range between [-1, 1].

In addition to these six evaluation perspectives, researchers have proposed flexible methods like manual evaluation[**?**], but such methods remain highly subjective. The above analysis shows that different evaluation perspectives focus on different aspects of models, and researchers often select metrics based on task characteristics, making horizontal comparison between models difficult. Moreover, research shows that models performing well on specific metrics may have large performance gaps in other aspects. For example, with complexity—a common quantitative evaluation metric—if an improved model has lower complexity than existing models, it indicates more optimized modeling. However, as mentioned earlier, complexity does not represent the quality of topic word mining results[**?**], prompting the development of effectiveness, correctness, and other metrics to ensure fairer evaluation. Nevertheless, due to differences in application domains and time slice divisions, there is still no unified standard for judging model quality.

3.5 Applications Based on AT Models

As research on AT models deepens, they have been widely applied in text mining and other fields, with notable attention in social media analysis, academic literature, and community detection.

3.5.1 Applications in Social Media With the rapid development of the Internet, social media has emerged. Represented by Weibo and Twitter, compared with traditional document data, these platforms feature fast data updates and informal network language. How to combine social media characteristics to complete data analysis tasks is currently a research hotspot. Early applications of AT models to social media data analysis mainly modeled based on data content, ignoring the strong real-time nature of social media data and failing to combine text content with temporal information, making it impossible to observe topic changes over time. Additionally, due to the short text length on Twitter, models cannot obtain ideal topic distributions through unsupervised learning. The ATC (Author-Topic-Community) model proposed by C.S. Li et al. solved this problem while achieving simultaneous inference of author interests and community structure[**?**]. Subsequent research has included user interest mining[**?**, **?**], extraction of domain-specific Weibo accounts[**?**], and Weibo recommendations based on user interests[**?**, **?**].

3.5.2 Applications in Academic Literature In academic literature, the most representative application of AT models is revealing the distribution correspondence between authors and topics across different time periods by considering temporal factors, deeply understanding the topic evolution processes and development trends of leading researchers in cutting-edge interdisciplinary fields, and revealing inter-topic influences to some extent, with certain reference value for predicting research trends. Additionally, since academic literature typically includes references, journals, and conference information, incorporating these elements has gradually attracted researchers' attention. For example, Y. Tu et al. proposed the Citation-Author-Topic (CAT) model in 2010[**?**], which simultaneously models paper authors and cited authors. Subsequent research has extended to author influence issues[**?**].

3.5.3 Applications in Data Communities Currently, AT model applications in data communities mainly include community detection and recommendation systems. Community detection can be divided into social media and scientific collaboration teams based on data type. Community detection methods can effectively discover data structures and evolution processes, helping researchers analyze network structures and properties in datasets to understand overall network trends, providing services for resource search, recommendation, and network structure optimization[**?**]. Similarly, recommendation systems can also be divided into social media and research domain data, with recommendations based on user/author interests.

3.5.4 Other Applications In addition to the main application domains mentioned above, there are specific applications based on researchers' particular experimental content. For example, in image domain automatic category annotation, specific resolution images and video clips are selected to achieve image and video topic recognition, with typical experiments including mining spatial correspondence relationships of brain cognitive functions[**?**]. Additionally, email data analysis extracts fields such as recipients, subject, content, and time to achieve associations between senders and topics/time[**?**].

4 Research Gaps and Development Trends

The above summarizes specific applications of AT models. Despite continuous improvements, due to demand updates and technological iterations, some shortcomings remain, specifically in four aspects:

1. **Technical Level:** Social media data is increasingly massive, with characteristics such as short text length and rapid update speed, generating large amounts of network language and symbolic language in short periods, increasing data noise and insufficient context information. This significantly increases the difficulty for AT models to process data, preventing topic identification from meeting expectations.

2. **Domain-Specific Method Selection Level:** The AT model was originally designed to analyze scientific literature data to achieve "author-topic" associations and reveal development evolution patterns in specific disciplines, with relatively ideal effects. However, later applications found AT models widely used in social media and even image domains, while applications in library and information science remain immature. Most research focuses on algorithm improvement and performance enhancement, but relatively few studies fully apply AT models and their improvements for topic discovery in library and information science, with some domestic research still 停留在 the level of applying bibliometric analysis software.

3. **Dataset Selection Level:** In research applying AT models to scientific literature data, most experimental data comes from corpora or public datasets, focusing on verifying model effectiveness. However, research that comprehensively determines representative datasets in fields through JCR, Chinese Academy of Sciences journal partitions, and authoritative papers/conference evaluation systems to reveal evolution in specific disciplines remains insufficient.

4. **Lack of Model Quality Evaluation System:** Researchers typically select evaluation metrics based on experience or task characteristics, but whether selected metrics can objectively measure model quality remains a major problem in the AT model field. How to improve or propose new evaluation metrics will become a future research focus.

In summary, future AT model research can develop in three directions: performance optimization (efficient training algorithms, reduced time/space complexity), standardization of model evaluation systems (divided according to task nature, combining task characteristics with metric principles to achieve consensus on evaluation methods for different tasks within domains), and deeper applications in library and information science (such as mining user interest characteristics for scholar profiling, personalized and precise services for libraries or scholars, or studying research collaboration and topic evolution in cutting-edge interdisciplinary fields to support management decision-making).

References

[?] Steyvers M, Smyth P, Rosen-Zvi M, et al. Probabilistic author-topic models for information discovery[C]//The tenth ACM SIGKDD international conference on knowledge discovery and data mining. Seattle, Washington: ACM, 2004: 306-315.

[?] Luo Guojing. Research on modular network and community mining based on topic models[D]. Hangzhou: Zhejiang University, 2008.

[?] Tang J, Zhang J, Yao L, et al. ArnetMiner: extraction and mining of academic social networks[C]//ACM SIGKDD international conference on knowledge discovery and data mining. Henderson: ACM, 2008: 990-998.

[?] Wang Yanpeng. Research progress on topic discovery and evolution of scientific literature based on topic models in China[J]. Library and Information Service, 2016, 60(3): 130-137.

[?] Wu Liang, Huang Weijing, Chen Wei, et al. ACT-LDA: integrated topic, community and influence analysis probabilistic model[J]. Computer Science and Exploration, 2013, 7(8): 718-727.

[?] Wang T, Huang Z, Gan C. On mining latent topics from healthcare chatlogs[J]. Journal of biomedical informatics, 2016, 61(C): 247-259.

[?] Morchid M, Bouaziz M, Kheder WB, et al. Spoken language understanding in a latent topic-based subspace[C]//International symposium on computer architecture. San Francisco: Springer, 2016: 710-714.

[?] Lee M, Huang R, Tong W. Discovery of transcriptional targets regulated by nuclear receptors using a probabilistic graphical model[J]. Toxicological sciences, 2016, 150(1): 64-73.

[?] Xuan J, Lu J, Zhang G, et al. Infinite author topic model based on mixed gamma-negative binomial process[C]//IEEE international conference on data mining. Atlantic City: IEEE, 2016: 489-498.

[?] Chen Xiaodong. Research on expert finding in biomedical domain[D]. Shanghai: Fudan University, 2013.

[?] Newman D, Karimi S, Cavedon L. Using topic models to interpret medline's medical subject headings[C]//Australasian joint conference on advances in artificial intelligence. Berlin, Heidelberg: Springer, 2009: 270-279.

[?] Chaiwanarom P, Ichise R, Lursinsap C. Finding potential research collaborators in four degrees of separation[C]//Advanced data mining and applications, international conference, ADMA2010. Chongqing: DBLP, 2010: 399-410.

[?] Ichise R, Fujita S, Muraki T, et al. Research mining using the relationships among authors, topics and papers[C]//International conference information visualization. Zurich: IEEE Computer Society, 2007: 425-430.

[?] Mou H, Geng Q, Jin J, et al. An author subject topic model for expert recommendation[M]//Information retrieval technology. Cham: Springer International Publishing, 2015: 83-95.

[?] Xue Wei. Author-topic model with asymmetric priors[D]. Hangzhou: Zhejiang University, 2011.

[?] Kim J, Kim D, Oh A. Joint modeling of topics, citations, and topical authority in academic corpora[J]. Transactions of the Association for Computational Linguistics, 2017, 5(8): 191-204.

[?] Mao J, Cao Y, Lu K, et al. Topic scientific community in science: a combined perspective of scientific collaboration and topics[J]. Scientometrics, 2017, 112(2): 851-875.

[?] Wu Zhonggang, Lü Zhao. A community detection algorithm based on local similarity[J]. Computer Engineering, 2016, 42(12): 196-203.

[?] Guan Peng, Wang Yuefen. Analysis of author research interest evolution in disciplinary domain life cycle[J]. Library and Information Service, 2016, 60(19): 116-124.

[?] Jeong YS, Lee SH, Gweon G. Discovery of research interests of authors over time using topic model[C]//International conference on big data and smart computing. New York: IEEE Computer Society, 2016: 24-31.

[?] Feng S, Cao J, Chen Y, et al. A model for discovering unpopular research interests[C]//International conference on knowledge science, engineering and management. Cham: Springer, 2015: 382-393.

[?] Kuznetsova A, Kyprianou AE, Pardo JC. Analyzing topics and authors in chatlogs for crime investigation[J]. Knowledge-based systems, 2017, 126(C): 1-13.

[?] Chen C, Ren J. Forum Latent Dirichlet Allocation for user interest mining in course reviews based on author topic model[J]. International journal of distance education technologies, 2017, 15(3): 1-14.

[?] Liu SN, Liu C, Peng Z, et al. Mining individual learning topics on mooc.guokr.com[C]//International conference on advanced data mining and applications. Berlin: Springer, 2015: 75-87.

[?] Yang T, Comar PM, Xu L. Community detection by popularity-based models for authored networked data[C]//IEEE/ACM international conference on advances in social networks analysis and mining. New York: IEEE, 2013: 74-81.

[?] Morchid M, Portilla Y, Josselin D, et al. Author-topic based representation of call-center conversations[C]//Spoken language technology workshop. New York: IEEE, 2014: 218-223.

[?] Mukherjee S, Basu G, Joshi S. Joint author sentiment topic model[C]//Computer security applications conference. New York: IEEE, 2014: 90-98.

[?] Li Chunshan. Research on clustering algorithms for social media content[D]. Harbin: Harbin Institute of Technology, 2014.

[?] Fan Changjun. Research on individual role identification methods based on information interaction networks[D]. Changsha: National University of Defense Technology, 2015.

[?] Rogers. Image topic modeling combining user and geographic information[D]. Hangzhou: Zhejiang University, 2012.

[?] Cheng K, Zhang Y, Qi M. AL-DDCNN: a distributed crossing semantic gap learning for person re-identification[J]. Concurrency & computation practice & experience, 2017, 29(3): 1-16.

[?] Morchid M, Dufour R, Bouallegue M, et al. Author-topic based representation of call-center conversations[C]//Spoken language technology workshop. New York: IEEE, 2014: 218-223.

[?] Morchid M, Dufour R, Linarès G, et al. Latent topic model based representations for robust theme identification of highly imperfect automatic transcriptions[J]. Lecture notes in computer science, 2015, 9042(2): 596-605.

[?] Luo W, Li H, Liu G, et al. Semantic annotation of satellite images using author-genre-topic model[J]. IEEE transactions on geoscience & remote sensing, 2013, 52(2): 1356-1368.

[?] Zhu XD, Yao Y, Liu ZJ, et al. Activity clustering for online anomaly detection[J]. Journal of computers, 2011, 6(6): 441-448.

[?] Shi Qingwei, Qiao Xiaodong, Xu Shuo, et al. Author topic evolution model and its application in research interest evolution analysis[J]. Journal of the China Society for Scientific and Technical Information, 2013, 32(9): 912-919.

[?] Liu Zhichao, Lu Meilian. Academic paper recommendation method based on hybrid model[EB/OL]. [2019-01-18]. http://www.paper.edu.cn/releasepaper/content/201411-282.

[?] Wang J, Hu X, Tu X, et al. Author-conference-topic-connection model for academic network search[C]//ACM international conference on information and knowledge management. New York: ACM, 2012: 2179-2183.

[?] Ha JK, An JY, Yoo KJ, et al. Exploring the leading authors and journals in major topics by citation sentences and topic modeling[C]//BIRNDL2016 joint workshop on bibliometric-enhanced information retrieval and NLP for digital libraries. New York: ACM, 2016: 42-50.

[?] Li Jie, Wang Xiaowei. Remote sensing image automatic category annotation method based on author topic model[J]. Computer Applications and Software, 2013, 26(10): 263-265.

[?] Luo W, Li H, Liu G, et al. Global salient information maximization for saliency detection[J]. Signal processing image communication, 2012, 27(3): 238-248.

[?] Luo W, Li H, Liu G. Automatic annotation of multispectral satellite images using author-topic model[J]. IEEE geoscience & remote sensing letters, 2012, 9(4): 634-638.

[?] Bertolero MA, Yeo BT, D'Esposito M. The modular and integrative functional architecture of the human brain[J]. Proceedings of the National Academy of Sciences of the United States of America, 2015, 112(49): 798-807.

[?] Wang Yonggui, Zhang Xu, Ren Junyang, et al. UF_{AT} model combining Weibo following characteristics for user interest mining[J]. Application Research of Computers, 2015, 32(7): 1982-1985.

[?] Wang Yonggui, Zhang Fengtian, Liu Yushi, et al. User interest topic mining method combining forwarding characteristics in Weibo[J]. Application Research of Computers, 2017, 34(7): 2068-2071.

[?] Li Jing, Yin Jian, Liu Shaopeng, et al. Weibo topic mining based on hashtag labels[J]. Computer Engineering, 2015, 41(4): 30-35.

[?] Wang Ping. Domain knowledge mining in network environment[D]. Shanghai: East China Normal University, 2010.

[?] Zhong Y, Fan Y, Tan W, et al. Web service recommendation based on topic modeling[C]//Asian conference on intelligent information and database systems. Cham: Springer International Publishing, 2015: 105-115.

[?] Tang J, Zhang J. Modeling the evolution of associated data[J]. Data & knowledge engineering, 2010, 69(9): 965-978.

[?] Yang M, Mei J, Xu F, et al. Discovering author interest evolution in topic modeling[J]. 2016, 21(7): 801-804.

[?] Zhou X, Shunxiang WU, Zhou X, et al. The biterm author topic in the sentences model for e-mail analysis[J]. IEICE transactions on information & systems, 2017, 100(8): 1852-1859.

[?] McCallum A, Wang X, Corrada-Emmanuel A. Topic and role discovery in social networks with experiments on enron and academic email[J]. Journal of artificial intelligence research, 2010, 30(2): 249-272.

[?] Kang IS, Na SH, Lee S, et al. On co-authorship for author disambiguation[J]. Information processing & management, 2009, 45(1): 84-97.

[?] Zhang R, Shen D, Kou Y, et al. Author name disambiguation for citations on the deep web[J]. Lecture notes in computer science, 2010, 6185(10): 198-209.

[?] Zheng Weijie. Research on author disambiguation methods for scientific literature[D]. Hangzhou: Hangzhou Dianzi University, 2017.

[?] Chen Yongheng, Zuo Wanli, Lin Yaojin. Application of author label topic model in scientific literature[J]. Computer Applications, 2015, 35(4): 1001-1005.

[?] Li CS, Ye YM, Zhang XF. TPS: an unsupervised webpage segmentation algorithm based on DOM tree structure mining[J]. Information, 2012, 1(15): 387-394.

[?] Gui Xiaoqing, Zhang Jun, Zhang Xiaomin, et al. Review of temporal topic model methods and applications[J]. Computer Science, 2017, 44(2): 46-55.

[?] Dufour R, Morchid M, Parcollet T. Tracking dialog states using an author-topic based representation[C]//Spoken language technology workshop. New York: IEEE, 2017: 544-551.

[?] Li DC, Okamoto J, Leischow S, et al. An author topic analysis of tobacco regulation investigators[C]//Pacific-Asia conference on knowledge discovery and

data mining. Cham: Springer, 2014: 616-627.

[?] Qi Xiaoqing, Jing Xiaojun. Improved LDA model applied to Weibo[EB/OL]. [2019-01-18]. http://www.paper.edu.cn/releasepaper/content/201212-118.

[?] Sun Guochao, Xu Shuo, Qiao Xiaodong. Development of ATOT model visualization tool[J]. Information Engineering, 2016, 2(4): 20-29.

[?] Xu S, Shi Q, Qiao X, et al. A dynamic users' interest discovery model with distributed inference algorithm[J]. International journal of distributed sensor networks, 2014, 2014(1): 1-11.

[?] Ho T, Do P. Analyzing users' interests with the temporal factor in social networks[C]//International conference on information technology & applications. New York: IEEE, 2013: 63-65.

[?] Wang X, McCallum A. Topics over time: a non-Markov continuous-time model of topical trends[C]//ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2006: 424-433.

[?] Yang Ruyi. Text semantic mining based on topic models[D]. Xi'an: Xidian University, 2015.

[?] Yang Ruyi, Liu Dongsu, Li Hui. An improved topic model incorporating external features[J]. New Technology of Library and Information Service, 2016, 32(1): 48-54.

[?] Yu Chuanming, Zuo Yuheng, Guo Yajing, et al. Dynamic discovery of author research interests based on composite topic evolution model[EB/OL]. [2018-05-26]. http://kns.cnki.net/kcms/detail/37.1389.N.20180419.1330.002.html.

[?] Li CS, Cheung WK, Ye Y, et al. The Author-Topic-Community model: a generative model relating authors' interests and their community structure[C]//International conference on advanced data mining and applications. Berlin: Springer, 2012: 753-765.

[?] Li C, Cheung WK, Ye Y, et al. The Author-Topic-Community model for author interest profiling and community discovery[J]. Knowledge & information systems, 2015, 44(2): 359-383.

[?] Yan E, Ding Y, Milojevic S, et al. Topics in dynamic research communities: an exploratory study for the field of information retrieval[J]. Journal of informetrics, 2012, 6(1): 140-153.

[?] Ding Y. Topic-based PageRank on author cocitation networks[J]. Journal of the American Society for Information Science and Technology, 2011, 62(3): 449-466.

[?] Yang J, Zeng J, Cheung WK. Multiplex Topic Models[C]//Pacific-Asia conference on knowledge discovery and data mining. Berlin Heidelberg: Springer, 2013: 568-582.

[?] Jiang Yuyan, Li Ping, Wang Qing, et al. Topic model combining DSTM and USTM methods[J]. Computer Science and Exploration, 2014, 8(5): 630-639.

[?] Yang Z, Hong L, Davison BD. Academic network analysis: a joint topic modeling approach[C]//IEEE/ACM international conference on advances in social networks analysis and mining. New York: IEEE, 2014: 324-333.

[?] McCallum A, Corrada-Emmanuel A, Wang X. Topic and role discovery in social networks[J]. IJCAI, 2005, 30(2): 786-791.

[?] Naveed N, Sizov S, Staab S. ATT: analyzing temporal dynamics of topics and authors in social media[C]//Proceedings of the 3rd international web science conference. New York: ACM, 2011: 1-7.

[?] Yang QQ, Li WJ. The LDA Topic Model Extension Study[J]. Logistics engineering, management and computer science, 2015, 169(15): 857-860.

[?] Poddar L, Hsu W, Lee ML. Author-aware aspect topic model to retrieve supporting opinions from reviews[C]//International conference on web intelligence and intelligent agent technology. New York: IEEE, 2011: 422-429.

[?] Wan Lukang, Zhang Qianwen, Xie Jinkui. Conformity model and analysis based on topic probability in social networks[J]. Small Microcomputer Systems, 2017, 38(2): 277-281.

[?] Schneider KM. Weighted average pointwise mutual information for feature selection in text categorization[M]//Knowledge discovery in databases: PKDD 2005. Berlin Heidelberg: Springer, 2005: 252-263.

[?] Wallach HM, Murray I, Salakhutdinov R, et al. Evaluation methods for topic models[C]//International conference on machine learning. New York: ACM, 2009: 1105-1112.

[?] Buntine W. Estimating likelihoods for topic models[M]. Advances in machine learning. Berlin Heidelberg: Springer, 2009: 51-64.

[?] Xie X, Li L, Zhang Z, et al. Back-buy prediction based on TrifG[C]//ACM SIGKDD workshop on mining data semantics. New York: ACM, 2012: 1-8.

[?] Zhang T, Luo W. Image quality assessment using author topic model[C]//International conference on information technology & applications. New York: IEEE, 2013: 63-65.

[?] Xu Z, Ru L, Xiang L, et al. Discovering user interest on twitter with a modified author-topic model[C]//IEEE/WIC/ACM international conference on web intelligence. New York: IEEE, 2013: 422-429.

[?] Yu Pan. Extraction of educational Weibo accounts based on topic models[D]. Wuhan: Central China Normal University, 2017.

[?] He Li, Jia Yan, Han Weihong, et al. Weibo user interest mining based on user topic model[J]. China Communications, 2014, 11(8): 131-144.

[?] Yan Liang. Research and implementation of enterprise Weibo recommendation method based on topic model[D]. Hefei: Anhui University, 2016.

[?] Tu Y, Johri N, Dan R, et al. Citation author topic model in expert search[C]//International conference on computational linguistics: Posters. New York: ACM, 2010: 1265-1273.

[?] Kataria S, Mitra P, Caragea C, et al. Context sensitive topic models for author influence in document networks[C]//International joint conference on artificial intelligence. California: AAAI, 2011: 2274-2280.

[?] Yan L, Niculescu-Mizil A, Gryc W. Topic-link LDA: joint models of topic and author community[C]//International conference on machine learning. New York: ACM, 2009: 665-672.

Author Contributions

Xu Han: Determined research direction, collected/processed/analyzed data, wrote the paper.
Liu Xiaoping: Provided revision suggestions.

Note: Figure translations are in progress. See original paper for figures.