Research on Multi-Specialty Expert Identification Methods: A Case Study in the Big Data Field (Postprint)

Authors: Liu Xiaoyu, Zhu Donghua, Wang Xuefeng, Huang Ying

Date: 2023-08-26T00:00:00+00:00

Abstract

[Purpose/Significance] For national governments, large and medium-sized enterprises, and research institutions, finding suitable experts when facing technical difficulties is an urgent problem. When confronting comprehensive and complex problems that require multidisciplinary knowledge to solve, identifying experts with multiple expertise becomes particularly important. The purpose of this study is to find a suitable method to identify such multi-expertise experts. [Method/Process] This study utilizes academic paper data published by experts, extracts representative research expertise features, and employs a TFIDF-weighted overlapping K-means clustering algorithm to perform overlapping clustering of experts, thereby mining their multiple research expertise and identifying multi-expertise experts. [Results/Conclusions] The research results demonstrate that the TFIDF-weighted overlapping K-means clustering algorithm exhibits good performance in terms of precision, recall, and F-value, and can effectively identify experts with multiple expertise.

Full Text

Preamble

Multi-Expertise Researcher Identification: A Case Study of the Big Data Field

Liu Xiaoyu, Zhu Donghua, Wang Xuefeng, Huang Ying
School of Management and Economics, Beijing Institute of Technology, Beijing 100081

Abstract

[**Purpose/Significance**] When confronting technical challenges, national governments, large and medium-sized enterprises, and research institutions urgently need to identify appropriate experts. For complex interdisciplinary problems, finding multi-expertise researchers is particularly crucial. This study aims to develop a suitable method for identifying such multi-expertise researchers. [**Method/Process**] Using academic publication data, we extracted representative research expertise features from experts and applied a TF-IDF weighted overlapping K-means clustering algorithm to partition experts into overlapping clusters, thereby uncovering multiple research expertise areas and identifying multi-expertise researchers. [**Results/Conclusion**] Results demonstrate that the TF-IDF weighted overlapping K-means algorithm performs well in terms of precision, recall, and F-value, proving effective for multi-expertise researcher identification.

Keywords: expert identification; overlapping K-means; multi-expertise researcher; big data; TF-IDF
Classification Number: G316

1 Introduction

In today's increasingly competitive international and commercial environment, the ability to rapidly understand and analyze needs while providing efficient solutions is a key determinant of success. In our knowledge-driven society, urgent knowledge demands are becoming more apparent, and expert identification and recommendation—hot topics in information retrieval and knowledge management—have attracted growing attention. The goal of expert identification is to discover domain experts with rich professional knowledge, skills, and experience through systematic methods, enabling organizations to form teams, guide R&D, and solve technical problems, thereby improving work and production efficiency [1].

Currently, national governments, large and medium-sized enterprises, and research institutions face difficulties in selecting and discovering technical experts [4]. Previous studies often used an expert's most productive research area to represent their expertise, but in reality, experts frequently possess multiple research specializations [5]. Identifying and recognizing these multiple expertise areas enables better expert evaluation and recommendation. Moreover, most prior research employed non-overlapping clustering methods that uniquely assigned experts to a single category, ignoring their multiple research strengths and failing to identify multi-expertise researchers. To address this limitation, this paper adopts an overlapping clustering algorithm to cluster experts, avoiding information loss from non-overlapping clustering while better representing expert expertise and uncovering multi-expertise researchers.

Accordingly, this study uses experts' published academic papers as data, employs a vector space model to represent expert knowledge, and utilizes TF-

IDF (Term Frequency-Inverse Document Frequency) weighted overlapping K-means clustering algorithm [6] to cluster experts and identify multi-expertise researchers. We demonstrate our approach through a case study of experts in the big data field.

2 Literature Review

Expertise identification forms the foundation for expert selection and recommendation. Early methods primarily relied on experts' self-reported specializations to build databases, using traditional database query languages for identification. However, this approach suffers from subjectivity and lacks timeliness in database updates [7]. Consequently, scholars have attempted to analyze expertise using document data (papers, patents, project reports, etc.) [8-10] and behavioral data (social tags, community groups, etc.) [11]. Methodologically, current expertise identification research mainly employs ontology-based methods, topology-based community detection algorithms, and topic-based expert clustering methods [12].

Ontology-based expertise identification methods construct domain ontologies to effectively capture semantic relationships between keywords [13-14], thereby enabling expertise recognition. Hu Yuehong et al. [15] built an ontology for the information science field using Formal Concept Analysis (FCA) and association rule analysis, mapping keywords to ontology concepts to transition from keyword-based to ontology-based expertise descriptions. Liu Xinmin et al. [16] proposed a four-layer fuzzy ontology extension framework and established a fuzzy ontology for the science and technology evaluation domain to facilitate expert selection.

Topology-based methods approach the problem from network structure perspectives, treating experts as network nodes and their relationships as edges to construct models such as co-author networks, author coupling networks [17], and co-citation networks. Y. Li et al. [18] utilized Shannon entropy to calculate network information for expert community mining through citation networks. B. Dom et al. [11] applied graph-theoretic ranking algorithms for expert community analysis. Gong Jun et al. [19] used spectral partitioning algorithms and modularity metrics to divide expertise areas. Liu Ping et al. [20] employed keyword co-occurrence networks with community detection methods to cluster keywords and identify expert specializations.

Topic-based expert clustering methods [21-22] use text mining to discover research interests and scopes, grouping experts with similar interests [23]. Main algorithms include hierarchical clustering for identifying research topic hierarchies, and topic models such as LDA [25] and PLSA [26]. Zhang Xiaojuan et al. [7] applied PLSA to identify expertise in library and information science, determining expert research areas through document-topic and topic-keyword matrices.

Ontology-based identification often requires substantial time and effort to con-

struct domain ontologies, while topology-based community discovery may not adequately represent relationships between communities, and expertise identification lacks analysis of research content. Topic-based clustering better expresses semantics and offers clear advantages when processing large datasets. Therefore, this study selects topic-based expert clustering for expertise identification.

For topic-based methods, two key factors affect identification effectiveness: how to represent expertise through text and how to calculate expert membership across topics. Keywords from published papers, project proposals, and social tags have been used to represent expertise, with paper keywords being particularly effective for concrete representation. Traditional clustering algorithms often assign experts to only one category, yet most experts possess multiple research strengths, causing information loss during clustering. This study addresses this by introducing overlapping clustering concepts. Overlapping clustering algorithms [27-28] mine each object's membership degree across categories, enabling more accurate and comprehensive classification through appropriate threshold settings. This has led to the development of topic-based overlapping expert clustering methods, an area where research remains relatively scarce and where this study aims to contribute.

3 Research Methodology

Identifying expertise through topic-based clustering requires first using appropriate keywords to represent expert knowledge—the foundation of expertise identification—and second, employing suitable clustering algorithms to calculate expert membership across topics—the key to identification. This section elaborates on the methodology from two perspectives: expert-keyword matrix construction and overlapping clustering algorithm analysis.

3.1 Expert-Keyword Matrix Construction

Experts' published papers, patents, and projects contain rich knowledge. Extracting effective information to enhance knowledge discovery capabilities is a major concern in information science research. The first step in expertise identification involves screening and processing expert texts, selecting appropriate keywords to represent knowledge, and constructing an expert-keyword matrix according to specific rules.

3.1.1 Keyword Acquisition Numerous studies in information and library science have developed analysis techniques centered on keywords [29]. First, keywords are obtained from scientific literature, either from author-provided keywords or through natural language processing. Second, stopword lists and common word lists for scientific journals remove meaningless or universally occurring terms. Third, stemming and fuzzy semantic processing clean variations including noun forms, plurals, and tenses, while manually constructed abbreviation tables merge full terms with abbreviations. Finally, appropriate keywords

are selected based on frequency or TF-IDF values to represent domain expertise.

3.1.2 Expert-Keyword Matrix Construction Based on acquired keywords representing expert knowledge, we construct an expert-keyword co-occurrence matrix [30-32]. Assigning weights to features is a common method to enhance discriminative ability in text classification, with studies showing that feature weighting significantly impacts classification effectiveness [33]. In this research, we assign different weights to keywords during expert clustering to achieve better results. TF-IDF is a widely used weighting technique in information retrieval and data mining [34], based on the principle that terms appearing frequently in a document but rarely in other documents have strong category discrimination capability. TF-IDF = TF × IDF, where TF is term frequency and IDF is inverse document frequency.

We construct the expert-keyword matrix with experts as row vectors and keywords as column vectors:

$$\begin{matrix} tf_{11} & \cdots & tf_{1p} \\ \vdots & \ddots & \vdots \\ tf_{n1} & \cdots & tf_{np} \end{matrix}$$

where tf_{ij} represents the frequency of keyword j in publications by expert i. Calculating TF-IDF for keywords yields vector $TFIDF = (tfidf_1, tfidf_2, ..., tfidf_p)$, where $tfidf_j$ is the TF-IDF value of keyword j. The final TF-IDF weighted expert-keyword matrix is:

$$\begin{matrix} tf_{11} \times tfidf_1 & \cdots & tf_{1p} \times tfidf_p \\ \vdots & \ddots & \vdots \\ tf_{n1} \times tfidf_1 & \cdots & tf_{np} \times tfidf_p \end{matrix}$$

3.2 Overlapping K-means Clustering Algorithm

This study employs the overlapping K-means algorithm proposed by G. Cleuziou [6] for expert clustering, with weighted improvements. Unlike traditional K-means, overlapping K-means assigns each data point to one or multiple clusters. Advantages include: (1) allowing point assignment to multiple clusters; (2) more objectively reflecting point positions through convergence conditions; (3) continuous data processing with wide applications in image recognition; and (4) low computational complexity offering time advantages for large datasets. These benefits address the limitation of previous expertise identification studies that recognized only one expertise area per expert, avoiding information loss and enabling comprehensive expertise mining.

The overlapping K-means algorithm comprises two processes: clustering and point assignment.

3.2.1 Clustering Process The clustering process iteratively updates cluster centers to minimize within-cluster differences and maximize between-cluster differences. Each expert is represented by a p-dimensional vector $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$, with the expert set denoted as $X = \{x_i\}_{i=1}^{n}$. The steps to cluster n experts into k overlapping clusters are:

1. Randomly select k initial points as cluster centers, denoted $\{m_c^{(0)}\}_{c=1}^{k}$.
2. Calculate distances from each x_i to the k centers, assign it to the nearest cluster, obtaining a cover $\{\pi_c^{(0)}\}_{c=1}^{k}$ of X, where $A_i = \{m_c | x_i \in \pi_c\}$ represents the set of clusters to which x_i belongs.
3. Set $t = 0$.
4. For each cluster $\{\pi_c^{(t)}\}_{c=1}^{k}$, compute new centers $m_h^{(t+1)} = PROTOTYPE(\pi_h^{(t)})$.
5. Perform new cluster assignment by computing assignment function $A_i^{(t+1)} = ASSIGN(x_i, \{m_c^{(t+1)}\}_{c=1}^{k})$ to obtain new cover $\{\pi_c^{(t+1)}\}_{c=1}^{k}$.
6. Compute objective function value $I(\{\pi_{t+1}\}) = \sum_{x_i \in X} dist(x_i, \phi(x_i))$. If $t_{max} > t$ or $I(\{\pi_t\}) - I(\{\pi_{t+1}\}) > \epsilon$, set $t = t + 1$ and return to step 4; otherwise, end the loop and output $\{\pi_c^{(t+1)}\}_{c=1}^{k}$.

The prototype calculation method is: $m_h^{(t+1)} = PROTOTYPE(\pi_c^{(t)}) = \frac{\sum_{x_i \in \pi_h} \alpha_i m_t}{\sum_{x_i \in \pi_h} \alpha_i}$, where $\alpha_i = \frac{1}{|A_i|^2}$, $A_i = \{m_c | x_i \in \pi_c\}$ represents the cluster set to which x_i belongs, $|A_i|$ is the number of clusters containing x_i, and the mapping of data point x_i in cluster h is $m_i = |A_i| x_i - \sum_{m_c \in A_i \{m_h\}} m_c$.

3.2.2 Point Assignment Process The point assignment process computes the assignment function by determining each expert's membership degree across categories during each iteration to find the optimal assignment.

1. Let $A_i = \{m^*\}$, where $m^* = \arg\min_{\{m_c\}_{c=1}^{k}}(dist(x_i, m_c))$, and compute $\phi(x_i) = \frac{\sum_{m_c \in A_i} m_c}{|A_i|}$.
2. Find the nearest center point m' not already assigned: $m' = \arg\min_{c=1\ A_i}(dist(x_i, m_c))$, and compute $\phi'(x_i)$ under new assignment $A_i \cup \{m'\}$.
3. If $\|x_i - \phi'(x_i)\| < \|x_i - \phi(x_i)\|$, update $A_i \leftarrow \{m'\}$, set $\phi(x_i) = \phi'(x_i)$, and return to step 2; otherwise, compute $\phi_{old}(x_i)$ under original assignment. If $dist(x_i, \phi(x_i)) < dist(x_i, \phi_{old}(x_i))$, output A_i; otherwise, output A_{old}.

Through overlapping K-means clustering, we obtain each expert's cluster memberships. Each cluster represents a research expertise area, and an expert's cluster assignments indicate their research strengths. Experts belonging to multiple clusters are identified as multi-expertise researchers.

4 Case Study: Multi-Expertise Researchers in Big Data

4.1 Empirical Data Selection

This study selected papers indexed in SCIE/SSCI from the Web of Science Core Collection in the big data field. Retrieval strategy quality directly affects result quality and final analysis accuracy. After reviewing extensive literature, we adopted a rigorous search strategy [36]. The search query was: TS=((“Big Data” or Bigdata) OR (((Big NEAR/1 Data or Huge NEAR/1 Data) OR “Massive Data” OR “Huge Information” OR “Big Information” OR “Large-scale Data” OR “Semi-Structured Data” OR “Unstructured Data”) AND (“analytic*” *OR ”analyz*” OR ”analys*”*))), with the time range set from 2008 to 2016, yielding 17,381 papers.

Keywords can represent article themes to some extent, but require processing due to unstandardized vocabulary, synonyms, near-synonyms, and meaningless terms. Processing steps included: (1) merging author keywords and Keywords Plus fields, obtaining 39,394 keywords; (2) using VantagePoint’s fuzzy matching module [37] to eliminate singular/plural forms and morphological variations (e.g., merging “networks” and “network” into “network”), reducing to 35,426 keywords; (3) creating manual abbreviation tables to merge abbreviations with full terms (e.g., merging “HDF” and “Hierarchical Data Format”), resulting in 35,299 final keywords.

Small sample sizes yield poor cluster interpretability, while excessively large samples make manual labeling for accuracy assessment impractical. Therefore, we selected 137 experts with 10+ publications (from 47,489 total authors) for clustering analysis. Keywords should be both representative and comprehensive; the top 251 keywords with frequency >40 covered 71.7% of articles, so we used these 251 keywords to classify experts.

Based on the constructed 137$\times$251 author-keyword matrix, we performed clustering calculations. Overlapping K-means requires setting cluster numbers and selecting initial centers. We determined these based on big data domain literature and keyword principal component analysis (PCA). Reading domain reviews and reports [38], big data research divides into three aspects: (1) basic theoretical research (origins, concepts, characteristics, architectures, significance); (2) storage and analysis technologies (cloud computing, Hadoop, MapReduce algorithms, data mining, clustering, and other techniques); (3) application research (gene sequencing in biomedicine, social network mining, etc.). The 2014 Big Data White Paper [39] identifies five key stages from data source to value: preparation, storage/management, processing, analysis, and knowledge presentation, with storage and analysis being critical. Application development remains in early stages requiring support.

PCA of the top 251 keywords yielded 12 categories: Classification, Lasso, RecommenderSystem, Hadoop, Hadoop(2), City, Gene, Managers, Massspectrometry, Risk, Thing, and Twitter [Figure 1: see original paper]. Classification and

Lasso showed correlation; RecommenderSystem, Hadoop, Hadoop(2), Gene, and Massspectrometry were correlated; City, Managers, Thing, and Twitter were correlated. These 12 components also divided into three main groups, consistent with literature and industry reports.

Thus, we divided the big data field into three categories: (1) Basic theoretical research (represented by Classification, Lasso, etc.); (2) Storage and analysis processing technologies (represented by CloudCompute, Hadoop, MapReduce, RecommenderSystem, etc.); (3) Application research (represented by InternetofThing, SmartCity, Twitter, Gene, Manager, etc.).

K-means algorithms are sensitive to initial centers, with results varying based on different inputs [41]. We selected representative experts after PCA. A.J. Jara published 12 papers, 11 focusing on big data applications, making him the application research representative. Similarly, S. Fong represents basic research, and X. Zhang represents storage and analysis processing technologies.

4.2 Empirical Analysis Results

Table 1 shows expertise identification results from overlapping K-means and TF-IDF weighted overlapping K-means (listing top 20 experts by publication count). Results are sorted by membership degree across categories. For example, top-ranked L. Wang's clustering result "2,1,3" indicates primary expertise in storage and analysis processing, with involvement in basic theory and application research. The top two experts in the table have research across all three areas, qualifying as multi-expertise researchers.

Statistical analysis of clustering results (Figures 2 [Figure 2: see original paper] and 3 [Figure 3: see original paper]) shows that overlapping K-means identified 65 multi-expertise researchers (47.4% of all experts). The TF-IDF weighted version identified 40.9% with overlapping expertise in storage/analysis and application research, indicating many experts conduct technical research while simultaneously focusing on applied studies—consistent with big data technology transitioning from basic research to practical applications in smart cities, IoT, healthcare, e-commerce, transportation, security, and communications.

4.3 Empirical Results Evaluation

We invited five big data domain experts to manually label the 137 researchers' expertise as evaluation criteria. First, we introduced our classification standards and boundaries. Domain experts then categorized each author's papers; if all five agreed, the label was confirmed. Disagreements were resolved through discussion after re-reading disputed papers.

This yielded manual labels for all 137 authors (Table 1 shows labels sorted by article count per category). Some experts received identical results from both methods (e.g., top-ranked L. Wang published 35 papers covering basic theory like algorithm improvements, storage/analysis like G-Hadoop, and appli-

cations like IoT). Differences also emerged: J. Wang (ranked 16th) was assigned to basic theory and storage/analysis by overlapping K-means, but only to storage/analysis by the TF-IDF weighted version—consistent with manual labeling. TF-IDF enhances category discrimination, giving greater influence to discriminative keywords during clustering, optimizing results and preventing the original algorithm from overestimating multi-expertise researchers.

Comparing both methods against manual labels, we calculated precision, recall, and F-values (Table 2).

Table 2 Clustering Results Evaluation

Method	Category 1 Precision	Category 1 Recall	Category 1 F-value	Category 2 Precision	Category 2 Recall	Category 2 F-value	Category 3 Precision	Category 3 Recall	Category 3 F-value	Avg Precision	Avg Recall	Avg F-value
Overlapping K-means	42.86%	69.23%	52.94%	76.92%	83.33%	80.00%	63.46%	73.08%	73.55%	97.80%	83.96%	73.74%
TF-IDF weighted Overlapping K-means	83.33%	76.92%	80.00%	88.00%	92.31%	77.06%	84.00%	88.37%	80.00%	-	-	-

While overlapping K-means alone produced suboptimal results, the TF-IDF weighted version showed strong performance: average recall reached 81.73%, average precision 83.11%, and average F-value 81.75%. This confirms that the proposed method accurately and efficiently identifies expert expertise.

Comparative analysis shows the TF-IDF weighted version significantly outperformed the baseline: Category 1 precision improved by 40.48% (from 42.86% to 83.33%), recall by 7.69%; Category 2 precision improved by 0.37%, recall by 9.62%; Category 3 precision improved by 3.51% while recall decreased by 5.49%, but the overall average improved by 3.94%. F-values increased by 27.06%, 6.26%, and 0.04% respectively, with an average improvement of 11.12%. Notably, Category 1's original recall, precision, and F-value were all below 70%, likely due to fewer experts in basic theory research and low inter-expert discrimination. After TF-IDF weighting, Category 1 precision rose to 83.33% and recall to 76.92%, demonstrating that TF-IDF weighting significantly improves identification effectiveness when data is limited.

5 Conclusion and Future Work

As research subjects, experts often possess multiple research interests, giving them irreplaceable advantages in interdisciplinary and fusion research. Traditional classification methods uniquely assign experts to single domains, neglecting multi-expertise researcher identification. To address this, we applied overlapping K-means clustering and innovatively proposed a TF-IDF weighted overlapping K-means algorithm for expert analysis. Our big data case study of 137 SCI/SSCI authors with 10+ publications revealed most experts cover multiple research directions, with substantial overlap between storage/analysis technology and application research. The TF-IDF weighted overlapping K-means algorithm demonstrated strong performance in precision, recall, and F-value, enabling accurate and efficient expertise identification.

This method addresses limitations in traditional expertise identification research, with experimental results confirming its effectiveness for multi-expertise researcher identification.

However, limitations remain. K-means requires predetermined cluster numbers and initial centers—a constraint overlapping K-means cannot overcome. We defined expertise categories and cluster numbers based on domain reports and literature, but our classification granularity is coarse without deep technical details, yielding broad expertise identification. Future research will examine fine-grained classification. Additionally, manual selection of initial center representatives may impact results—a factor worth investigating. This study used Euclidean distance; future work will analyze alternative distance metrics like cosine distance to better characterize inter-expert distances and optimize clustering.

References

[1] Long Xin. Research on Community Mining for Expert Retrieval [D]. Kunming: Yunnan University, 2010.
[2] BEDARD J. Expertise and its relation to audit decision quality [J]. Contemporary accounting research, 2010, 8(1): 198-222.
[3] GLASER R. The nature of expertise. Occasional paper No.107. [EB/OL]. [2017-05-20]. https://eric.ed.gov/?id=ED261190
[4] YIMAM SEID D, KOBSA A. Expert-finding systems for organizations: problem and domain analysis and the DEMOIR approach [J]. Journal of organizational computing & electronic commerce, 2003, 13(1): 1-24.
[5] Lu Wei, Liu Jie, Qin Xiyan. Expert Retrieval and Evaluation in Library and Information Science Based on Expertise Vocabulary [J]. Journal of Library Science in China, 2010, 36(2): 70-76.
[6] CLEUZIOU G. An extended version of the k-means method for overlapping clustering [C]//IEEE. Proceedings of 19th international conference on pattern recognition. Tampa: IEEE Press, 2008: 563-566.
[7] Zhang Xiaojuan, Lu Wei, Cheng Qikai. Application of PLSA in Expertise

Identification in Library and Information Science [J]. New Technology of Library and Information Service, 2012, 28(2): 76-81.
[8] APPIO F P, CESARONI F, MININ A D. Visualizing the structure and bridges of the intellectual property management and strategy literature: a document co-citation analysis [J]. Scientometrics, 2014, 101(1): 623-661.
[9] SONG X, TSENG B L, LIN C Y, et al. ExpertiseNet: relational and evolutionary expert modeling [C]//ARDISSONO L, BRNA P, MITROVIC A. Proceedings of user modeling 2005. Edinburgh: Springer, 2005: 99-108.
[10] YANG K W, HUH S Y. Automatic expert identification using a text categorization technique in knowledge management systems [J]. Expert systems with applications, 2008, 34(2): 1445-1455.
[11] DOM B, EIRON I, COZZI A, et al. Graph-based ranking algorithms for e-mail expertise analysis [C]//KIM W, KOHAVI R, GEHRS J, et al. Proceedings of the 8th ACM SIGMOD workshop on research issues in data mining and knowledge discovery. San Diego: IEEE Press, 2003: 42-48.
[12] Mao Jin, Li Gang. A Method for Constructing Researcher Expertise Graphs Based on OKM [J]. Library and Information Service, 2012, 56(4): 17-21.
[13] Wei Yuanyuan, Qian Ping, Wang Rujing, et al. Knowledge Base, Ontology and Expert System in Knowledge Engineering [J]. Computer Systems & Applications, 2012, 21(10): 220-223.
[14] Wu Chunyin, Chen Zhuangguang, Wang Haojie, et al. Review of Ontology-Based Expert Systems [J]. Agricultural Network Information, 2013(4): 5-8.
[15] Hu Yuehong, Liu Ping. Research on Expertise Representation Based on Ontology Concepts [J]. Library and Information Service, 2014, 58(14): 34-40.
[16] Liu Xinmin, Gui Weihua, Yang Liu, et al. Research on Expert Selection Service Based on Fuzzy Domain Ontology [J]. Transactions of Beijing Institute of Technology, 2013, 33(5): 484-489.
[17] LIU R. A new bibliographic coupling measure with descriptive capability [J]. Scientometrics, 2016, 110(2): 1-21.
[18] LI Y, ZHANG G, FENG Y, et al. An entropy-based social network community detecting method and its application to scientometrics [J]. Scientometrics, 2015, 102(1): 1003-1017.
[19] Gong Jun, Liu Lu. Representation and Measurement of Expert Knowledge Based on Knowledge Networks [J]. Scientific Research, 2010, 28(10): 1521-1529.
[20] Liu Ping, Zhou Menghuan. Expertise Mining Based on Co-word Networks [J]. Information Science, 2012, 30(12): 1815-1819.
[21] STEYVERS M, SMYTH P, ROSEN-ZVI M, et al. Probabilistic author-topic models for information discovery [C]//KOHAVI R. Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2004: 306-315.
[22] TEH Y W, JORDAN M I, BEAL M J, et al. Hierarchical Dirichlet processes [J]. Journal of the American Statistical Association, 2006, 101(476): 1566-1581.
[23] YAU C K, PORTER A L, NEWMAN N, et al. Clustering scientific documents with topic modeling [J]. Scientometrics, 2014, 100(3): 767-786.

[24] BLEI D M, JORDAN M, GRIFFITHS T L, et al. Hierarchical topic models and the nested Chinese restaurant process [C]//THRUN S, SAUL L K. Proceedings of the 16th international conference on neural information processing systems. Cambridge: MIT Press, 2003: 17-24.
[25] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of machine learning research, 2003, 3(4/5): 993-1022.
[26] HOFMANN T. Probabilistic latent semantic indexing [C]//GEY F, HEARST M, TONG R. Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. Berkeley: ACM, 1999: 50-57.
[27] MULDER W D. Optimal clustering in the context of overlapping cluster analysis [J]. Information sciences, 2013, 223(4): 56-68.
[28] N'CIR C B, BEN C, CLEUZIOU G, et al. Overview of overlapping partitioning clustering methods [M]. New York: Springer International Publishing, 2015: 245-275.
[29] ZHANG Y, PORTER A L, HU Z, et al. "Term clumping" for technical intelligence: a case study on dye-sensitized solar cells [J]. Technological forecasting & social change, 2014, 85(4): 26-39.
[30] LEE K, JUNG H, SONG M. Subject-method topic network analysis in communication studies [J]. Scientometrics, 2016, 109(3): 1761-1787.
[31] AHLGREN P, JARNERING B. Bibliographic coupling, common abstract stems and clustering: a comparison of two document-document similarity approaches in the context of science mapping [J]. Scientometrics, 2008, 76(2): 273-290.
[32] Liu Kan, Liu Ping. Research on Expert Domain Analysis and Visualization Based on VSM [J]. Library and Information Service, 2011, 55(10): 74-77.
[33] Shi Congying, Xu Chaojun, Yang Xiaojiang. Review of TF-IDF Algorithm Research [J]. Computer Applications, 2009, 29(6): 167-170.
[34] Zhang Yufang, Peng Shiming, Lü Jia. Improvement and Application of TF-IDF Method in Text Classification [J]. Computer Engineering, 2006, 32(19): 76-78.
[35] N'CIR C B, BEN C, ESSOUSSI N. On the extension of K-means for overlapping clustering—average or sum of clusters' representatives? [C]//FRED A, DIETZ J, LIU K. IC3K. KDIR/KMIS 2013-Proceedings of the international conference on knowledge discovery and information retrieval and the international conference on knowledge management and information sharing. Vilamoura: Springer, 2013: 208-213.
[36] HUANG Y, SCHUEHLE J, PORTER A L, et al. A systematic method to create search strategies for emerging technologies based on the Web of Science: illustrated for 'big data' [J]. Scientometrics, 2015, 105(3): 2005-2022.
[37] VantagePoint [EB/OL]. [2017-05-20]. http://www.thevantagepoint.com/.
[38] Li He, Yuan Cuimin, Li Yafeng. Bibliometric Review of Big Data Research [J]. Information Science, 2014, 32(6): 148-155.
[39] China Academy of Telecommunication Research of MIIT. Big Data White Paper [R]. Beijing: China Academy of Telecommunication Research of MIIT, 2014.

[40] CHURYk N T, JANVRIN D, WATSON M. Special issue on big data [J]. Journal of accounting education, 2017, 38(1): 1-2.
[41] STEINLEY D. Local optima in K-means clustering: what you don't know may hurt you [J]. Psychological methods, 2003, 8(3): 294-304.

Author Contributions

Liu Xiaoyu: Conceived research idea, designed methodology, collected and analyzed data, wrote manuscript;
Zhu Donghua: Participated in methodology design;
Wang Xuefeng: Participated in conceptualization and manuscript revision;
Huang Ying: Participated in manuscript revision.

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