Postprint of Research on Blog Post Clustering Based on Participant Co-occurrence Analysis

Authors: Gong Kaile, Cheng Ying, Sun Jianjun

Date: 2017-11-08T00:00:00+00:00

Abstract

【Purpose】 To utilize participant co-occurrence in blog posts as a feature and investigate its value in blog post clustering. 【Method】 A two-step clustering approach: construct a co-occurrence matrix of different blog post participants and transform it into a correlation matrix, employ the Affinity Propagation (AP) algorithm for the first-step clustering; utilize the centroids of the AP clustering results as initial cluster centers, apply position weighting to terms, and employ the K-means algorithm for the second-step clustering of blog post content. 【Results】 The clustering algorithm integrating blog post participant co-occurrence and term position weighting achieved average precision and purity of 0.66 and 0.57, respectively, significantly outperforming comparative experiments. 【Limitations】 The primary contribution of this study is introducing participant co-occurrence as a feature to improve blog post clustering effectiveness; for blog posts with scarce such features, its value for clustering is limited. 【Conclusion】 Blog post clustering that integrates term and participant features significantly improves clustering quality, and the two-step clustering approach also provides a feasible solution for selecting initial cluster centers for the K-means algorithm.

Full Text

Clustering Blog Posts with Co-occurrence Analysis of Participants

Gong Kaile, Cheng Ying, Sun Jianjun
School of Information Management, Nanjing University, Nanjing 210023, China

Abstract

[**Objective**] This study investigates the value of participant co-occurrence as a feature for blog post clustering. [**Methods**] We propose a two-step clustering approach: first, constructing a co-occurrence matrix of blog participants from

different posts and transforming it into a correlation matrix, followed by clustering using the Affinity Propagation (AP) algorithm; second, using the AP cluster centroids as initial centers for K-means clustering of blog content with term position weighting. [**Results**] The proposed algorithm achieved average precision and purity of 0.66 and 0.57, respectively, significantly outperforming comparative experiments. [**Limitations**] The primary contribution lies in introducing participant co-occurrence to improve clustering effectiveness, though its value is limited for blog posts with sparse participant features. [**Conclusions**] Integrating both term features and blog participant features significantly enhances clustering quality, while the two-step approach provides a viable solution for selecting initial cluster centers in K-means algorithms.

Keywords: Co-occurrence analysis; Text clustering; Blog participants; Initial cluster centers

1. Introduction

In 1973, Small proposed co-citation theory, which posits that co-citation reflects content similarity between two documents and that co-citation relationships can be used to reveal scientific structures [**?**]. Subsequent scholars extended this co-occurrence concept to author co-citation analysis [**?**] and co-word analysis [**?**]. With the popularization of the Internet, Larson applied this idea to the Web through co-link analysis to reveal thematic relationships between websites [**?**]. Currently, co-occurrence analysis—abstracted from specific applications such as co-citation, co-word, and co-link analysis—has become an important research method in bibliometrics, informetrics, and scientometrics, enabling the discovery of relationships between research objects, mining of latent knowledge, and revelation of structural patterns and changes [**?**].

In terms of applications, researchers have applied co-occurrence features such as words [**?**], citations [**?**], and links [**?**] to text clustering, thereby improving clustering quality. Author co-citation analysis can delineate author groups and identify core authors in a field [**?**]. Co-recommendation relationships in social media reflect homogeneous characteristics such as shared interests and similar backgrounds, providing a basis for community network construction [**?**]. Related studies [**?**] have found that user interest stability makes it likely for users to participate in discussions on the same topic within a certain period, while relationships such as following and friendship established through long-term interaction and recommendations from users with similar backgrounds enable more like-minded users to join topic discussions.

As social media users, we observed an interesting phenomenon in blogs: a considerable number of common participants frequently appear in blog posts with similar or identical themes. The co-occurrence concept and research findings from [**?**] suggest that participant co-occurrence in blog posts could serve as a clustering feature. Studies [**?**] provide valuable empirical foundations, indicating a correlation between participant co-occurrence and blog themes. Accordingly,

this paper employs participant co-occurrence intensity to measure blog post similarity and uses co-occurrence analysis to uncover implicit thematic associations.

We define participants as the set of users involved in blog post writing, commenting, and recommending. The clustering process employs a two-step method: (1) constructing a participant co-occurrence matrix for different blog posts and transforming it into a correlation matrix, followed by the first-step clustering using the Affinity Propagation (AP) algorithm [**?**]; (2) using the AP clustering centroids as initial centers and applying K-means algorithm with term position weighting for the second-step blog content clustering.

2. Literature Review

2.1 Text Clustering Based on Co-word Analysis

Co-word analysis fundamentally involves counting the frequency of word pairs co-occurring in the same document and performing clustering based on this feature. Liu et al. [**?**] used high-frequency co-occurring terms in a document collection as features to achieve vector space dimensionality reduction. Chang et al. [**?**] extracted word pairs from the same context using association rule algorithms and used them as vector elements for document representation. Zhang et al. [**?**] proposed the CoHC algorithm, which discovers frequent 2-grams in document collections based on term co-occurrence relationships, extends them to n-grams, removes redundancy, sorts them to obtain candidate cluster labels, and builds base classes (documents sharing the same phrase constitute a base class) to complete clustering through base class merging. Building on Zhang et al.' s work, Xiao et al. [**?**] constructed an important term set from terms frequently co-occurring with user keywords in search engine results, merged documents where each term appeared into corresponding base classes, performed base class merging based on document overlap, and optimized clustering results by calculating inter-class similarity using HowNet' s semantic ontology. Li and He [**?**] extracted document keywords based on features such as term frequency and position, constructed a co-word matrix based on keyword co-occurrence frequency between documents, performed label clustering using hierarchical clustering, and then achieved document clustering based on label combinations.

2.2 Text Clustering Based on Co-citation and Co-link Analysis

Beyond content-based co-word features, scholars have also applied external features such as co-citation and co-link to text clustering research. The main algorithmic approaches include three categories: direct application of co-citation and co-link features to clustering [**?**]; fusion of similarity based on co-citation/co-link relationships with text similarity for clustering [**?**, **?**]; and two-step clustering, where co-citation/co-link features are used for the first step and content features for the second step [**?**]. Table 1 lists the main research findings for each approach.

Table 1 Research on Text Clustering Based on Co-citation and Co-link Analysis

Research Approach	Representative Studies	Key Contributions
Direct clustering based on citation/link co-occurrence relationships	Wang Y, Kitsuregawa M [?]	Applied co-link analysis of Web search results to measure text similarity.
	Mukhopadhyay D, Sing S R [?]	Extended link analysis units from single web pages to thematic document sets, where links citing a thematic document set are equivalent to those citing a single page, enabling co-link analysis for text clustering.
Fusion of citation/link co-occurrence similarity with content similarity	He X, Zha H, Ding C H Q, et al. [?]	Proposed a new similarity calculation method integrating Web text link relationships, co-citation patterns, and text content.
	Modha D S, Spangler W S [?]	Fused text similarity and co-link similarity by constructing text vectors and link relationship vectors.
	Gu Jun, Zheng Xiaodong, Zhang Lianming [?]	Proposed a clustering algorithm fusing text content and citation information, where citation similarity comprehensively considers citing, cited, and co-citation relationships.

Two-step clustering based on citation and text	Wu Suhui, Cheng Ying, Zheng Yanning, et al. [?]	Constructed a co-citation matrix for academic literature and used hierarchical clustering to obtain the K value and initial centroids required by K-means algorithm.

2.3 Blog Post Clustering Research

Existing blog post clustering research primarily focuses on applying textual features including titles, keywords, tags, and main content. Brooks et al. [?] used TF-IDF to extract a small number of terms from blog posts as tags for clustering. He and He [?] utilized Web 2.0' s social tagging system, using user-added tags as features for clustering. Zhang et al. [?] expanded the social tagging system using various information such as blog websites, users, posts, categories, and tags to improve upon the limitation of using only tags as clustering features. Chen et al. [?] proposed a clustering algorithm based on blog search query keywords, collecting user query terms when searching blogs and using formal concept analysis to extract concepts characterizing blog content for clustering. Li et al. [?] proposed a multi-feature fusion clustering method using title, content, and comments as blog features, with comments given higher weight. Regarding link feature applications, Kopel et al. [?] calculated blog similarity using content relationships, blogger social network relationships based on XFN links, and blog topic relationships based on WordNet to achieve clustering. Blog link features have also been applied to community discovery, blogger recommendation, and blog information propagation [?].

These studies demonstrate that clustering algorithms employing co-word, co-citation, and co-link features achieve superior results, thereby validating Small' s co-citation analysis concept that frequently co-occurring feature pairs in the same context form stable combinations expressing latent thematic information. Therefore, exploratory clustering research using co-occurrence information of different blog participants as features is worthwhile.

Drawing on Wu et al.' s research approach [?], this paper proposes a two-step clustering algorithm for blog posts, as shown in Figure 1 [Figure 1: see original paper]. The algorithm first constructs a participant co-occurrence matrix, transforms it into a correlation matrix using Jaccard coefficients, and performs co-occurrence analysis using the AP clustering algorithm [?]. The second step uses the AP clustering centroids as initial centers and employs K-means algorithm based on position-weighted term features to complete blog post clustering.

3. Methodology

3.1 Blog Participant Co-occurrence Analysis

The participant co-occurrence matrix is constructed as shown in Equation (1), where N_{ij} represents the number of participants appearing in both blog post i and j, and when $i = j$, it represents the number of participants in blog post i, abbreviated as N_i.

$$\begin{pmatrix} N_{11} & N_{12} & \cdots & N_{1n} \\ N_{21} & N_{22} & \cdots & N_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ N_{n1} & N_{n2} & \cdots & N_{nn} \end{pmatrix} \tag{1}$$

The Jaccard coefficient (Equation (2)) transforms the co-occurrence matrix into a correlation matrix (Equation (3)), where J_{ij} represents the similarity between blog post i and j based on participant co-occurrence.

$$J_{ij} = \frac{N_{ij}}{N_i + N_j - N_{ij}} \tag{2}$$

$$\begin{pmatrix} J_{11} & J_{12} & \cdots & J_{1n} \\ J_{21} & J_{22} & \cdots & J_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ J_{n1} & J_{n2} & \cdots & J_{nn} \end{pmatrix} \tag{3}$$

(1) AP Algorithm

This study employs the AP clustering algorithm [**?**] for co-occurrence analysis. AP determines the optimal representative point for each sample through iterative updating of similarity relationships in the sample space, thereby forming multiple clusters. The algorithm typically uses an $n \times n$ similarity matrix as input, with diagonal elements assigned the same value p, called the preference parameter. The magnitude of p determines the initial likelihood of each sample being selected as a cluster representative, positively correlating with the number of clusters. Thus, the number of clusters can be set by adjusting p [**?**]. The algorithm achieves clustering by iteratively updating responsibility $r(i,k)$ and availability $a(i,k)$. The AP algorithm flow using Jaccard similarity coefficients is as follows:

1. Define similarity matrix element $s(i,k)$ as Equation (4). Initialize $a(i,k) = 0$.

$$s(i,k) = J_{ik}, \quad i \neq k \tag{4}$$

2. Update $r(i,k)$ and $a(i,k)$ according to Equations (5) and (6).

$$r(i,k) = s(i,k) - \max_{k' \neq k}\{a(i,k') + s(i,k')\} \tag{5}$$

$$a(i,k) = \min\{0, r(k,k) + \sum_{i' \neq i,k} \max\{0, r(i',k)\}\}, \quad i \neq k \tag{6}$$

3. Introduce damping coefficient λ to eliminate potential oscillations using Equation (7).

$$r_{\text{new}}(i,k) = \lambda \times r_{\text{old}}(i,k) + (1-\lambda) \times r(i,k) \tag{7}$$

$$a_{\text{new}}(i,k) = \lambda \times a_{\text{old}}(i,k) + (1-\lambda) \times a(i,k)$$

where $\lambda (0 < \lambda < 1)$; larger values yield better oscillation elimination effects, typically set to 0.9 [**?**].

4. Determine the cluster representative c_i for each sample i using Equation (8).

$$c_i = \arg\max_k \{a(i,k) + r(i,k)\} \tag{8}$$

5. The algorithm terminates when cluster representatives remain unchanged after 10 iterations or when the maximum preset iteration count is reached; otherwise, it returns to step 2.

(2) Selection Rationale

Zhou et al. [**?**] argue that commonly used algorithms such as K-means and hierarchical clustering have limitations in co-occurrence analysis because they interpret the "sample-sample" correlation matrix as a "sample-variable" two-dimensional table for vector space modeling. Since the correlation matrix already embodies sample similarity based on co-occurrence relationships, calculating vector similarity in "sample-variable" form during clustering means the basis for classification is no longer direct co-occurrence relationships, representing an essential conceptual difference. Experimental results demonstrate that hierarchical clustering using Euclidean distance based on Pearson coefficients yields poor co-occurrence analysis effects [**?**]. In contrast, the AP algorithm employed in this study uses the correlation matrix as input and bases clustering on sample co-occurrence similarity relationships, effectively addressing these limitations. Additionally, AP treats all samples as potential centers, considering each sample' s similarity relationships with all others, and does not converge to local optima. Related research also indicates that AP outperforms most existing algorithms in quality and stability [**?**]. Its advantage of intuitive centroids (cluster centers are actual existing samples) also meets the requirements for selecting initial centers for the K-means algorithm in this study.

3.2 K-means Algorithm with Term Position Weighting

The second-step clustering uses title, tags, main content, and comments as features to construct a vector space model, employing K-means algorithm to complete clustering. Since these four content features vary in their ability to express themes, different weights are assigned to features at different positions. Position weighting strategies lack unified standards in academia, with considerable variation across studies. Han and Wang [**?**] suggest that terms in academic literature titles are 3-5 times more important than those in abstracts and 10-15 times more important than those in main content. Li et al. [**?**] demonstrate that incorporating comments improves blog text clustering quality, but Miao et al. [**?**] find through statistical analysis that blog comments contain substantial noise and exhibit topic drift. Guo et al. [**?**] experimentally determine that satisfactory clustering results are achieved when weights for blog titles, tags, main content, and comments are set to 2, 2, 3, and 2, respectively. Through observation and experimentation, this study sets title, tag, main content, and comment weights to 2:3:2:1, using position-weighted TF-IDF for term weighting as shown in Equation (9).

$$\text{TF-IDF}_{t,d} = \text{TF}_{t,d} \times \text{IDF}_t \tag{9}$$

where $\text{TF}_{t,d}$ is defined as the position-weighted term frequency of term t in document d, and IDF_t is the inverse document frequency.

The position-weighted term frequency is calculated as:

$$\text{TF}_{t,d} = \lambda_t \times tf_{tt} + \lambda_k \times tf_{tk} + \lambda_c \times tf_{tc} + \lambda_r \times tf_{tr} \tag{10}$$

In Equation (10), λ_t, λ_k, λ_c, and λ_r represent weight coefficients for titles, tags, main content, and comments, respectively, while tf_{tt}, tf_{tk}, tf_{tc}, and tf_{tr} denote the occurrence counts of term t in the title, tags, main content, and comments of document d. In Equation (11), N refers to the total number of documents in the dataset, and DF_t is the document frequency representing the number of documents containing term t.

$$\text{IDF}_t = \log \frac{N}{\text{DF}_t} \tag{11}$$

The K centroids determined in Section 3.1 serve as initial centers for the K-means text clustering algorithm. The K-means algorithm details are provided in [**?**]. Vector similarity is calculated using Euclidean distance as shown in Equation (12).

$$\text{Sim}(d_i, d_j) = \|V_i - V_j\| \tag{12}$$

4. Experiments

4.1 Experimental Data

ScienceNet Blog is the most active academic blog community in China. We collected 600 "popular blog posts within half a year" and their participant information as of December 19, 2015, using a web crawler. For text preprocessing, we employed the NLPIR Chinese word segmentation system for tokenization and part-of-speech tagging. Han et al. [**?**] demonstrate that using a combination of nouns, verbs, adjectives, and adjectives as features yields better clustering results than other part-of-speech combinations; therefore, we filtered out content not belonging to these part-of-speech categories. The preprocessed blog posts and participant information constitute the experimental dataset. Two researchers independently performed manual classification of blog content, and discrepancies were resolved by consulting a third researcher. The classification distribution is shown in Table 2 .

Table 2 Manual Classification of Academic Blog Posts

Category	Count
Research Insights	140
Daily Life & Others	102
Education & Teaching	101

4.2 Evaluation Metrics

We employ Average Accuracy [**?**] and Purity [**?**] to evaluate blog post clustering quality.

Average Accuracy examines the consistency between manual classification and automatic clustering results for any two blog posts. Relevant definitions are shown in Table 3 , where Positive Accuracy (PA) is given by Equation (13), Negative Accuracy (NA) by Equation (14), and Average Accuracy (AA) by Equation (15). Higher average accuracy indicates better clustering quality.

Table 3 Consistency Relationships Between Classification and Clustering

	Same Cluster (Auto)	Different Clusters (Auto)
Same Class (Manual)	True Positives (TP)	False Negatives (FN)
Different Classes (Manual)	False Positives (FP)	True Negatives (TN)

$$\mathrm{PA} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{13}$$

$$\mathrm{NA} = \frac{\mathrm{TN}}{\mathrm{FN} + \mathrm{TN}} \tag{14}$$

$$\mathrm{AA} = \frac{\mathrm{PA} + \mathrm{NA}}{2} \tag{15}$$

Purity is another commonly used clustering quality metric. It labels each cluster with the class label of the majority class in that cluster, then calculates the ratio of correctly labeled documents to the total number of documents in the collection using Equation (16).

$$\mathrm{Purity} = \frac{1}{N} \sum_{i} \max_{j} |\omega_i \cap c_j| \tag{16}$$

where $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$ is the manual classification set and $C = \{c_1, c_2, \ldots, c_K\}$ is the blog post cluster set.

4.3 Results Analysis

The primary research objective is to explore the impact of different blog features on clustering effectiveness. Specific experiments include:

- **Experiment 1:** Participant co-occurrence analysis (AP algorithm) + term position weighting for titles, tags, main content, and comments (K-means algorithm)
- **Experiment 2:** Participant co-occurrence analysis (AP algorithm) + TF-IDF weighting for titles and main content (K-means algorithm)
- **Experiment 3:** Term position weighting for titles, tags, main content, and comments (AP algorithm)
- **Experiment 4:** Term position weighting for titles, tags, main content, and comments (K-means algorithm), where classic K-means runs independently three times and clustering quality is averaged

(1) Impact of Different Text Features on Clustering Effectiveness

As shown in Table 4 , all four evaluation metrics demonstrate that using term position weighting for blog posts yields superior quality compared to using only titles and main content with TF-IDF weighting. This confirms that when clustering texts from social media like blogs, introducing tags, comments, and other content with appropriate feature weights provides clear value. Tags serve as concise summaries of blog content by authors, similar to keywords in academic literature, and have clear significance for expressing and distinguishing text themes. Comments primarily represent interactions and discussions between readers and authors about blog content, containing numerous topic-relevant terms. This guidance becomes particularly important for blogs containing multimedia such as images, music, and videos.

(2) Value of Participant Co-occurrence Information for Blog Clustering

The value of participant co-occurrence information can be assessed by comparing Experiments 1, 3, and 4, with results shown in Table 5 .

Table 5 Clustering Results of Experiments 1, 3, and 4

Metric	PA	NA	AA	Purity
Experiment 1	0.386162	-	-	-

The following conclusions can be drawn:

1. **Participant co-occurrence features significantly improve clustering effectiveness.** The PA, NA, AA, and purity results in Table 5 all show that with identical text features and weighting methods, the proposed participant co-occurrence-based clustering algorithm significantly outperforms both the text content-based AP clustering algorithm and the classic K-means algorithm, confirming that co-occurrence analysis effectively extracts implicit thematic associations from participant co-occurrence information.
2. **The two-step clustering algorithm improves K-means initial center determination.** The first-step clustering produces optimized initial centers, enhancing clustering stability. Classic K-means randomly selects initial centers, which significantly impacts results and introduces considerable randomness. In contrast, the AP algorithm maintains stable clustering results across multiple runs on the same dataset, ensuring the stability of our proposed algorithm. AA and purity data also reveal that AP' s accuracy surpasses classic K-means. Further examination of PA and NA shows that while AP' s PA is slightly lower, its NA is significantly higher than classic K-means, indicating that AP better distinguishes objects from different classes and forms more uniform clusters on datasets with small inter-class distinctions.

5. Conclusion

This study improves blog post clustering effectiveness through participant co-occurrence analysis. Experimental results demonstrate that implicit thematic association information in participant co-occurrence improves initial center selection for classic K-means algorithms, effectively enhancing clustering quality and stability. Additionally, comprehensive application of term position weighting and text features also contributes to improved clustering quality.

Currently, blog post clustering represents a research hotspot, and this work provides valuable insights for the field. However, this study primarily relies on participant co-occurrence features to improve clustering effectiveness, which may

result in overly sparse correlation matrices for blog posts with few participants, limiting the first-step clustering. Therefore, exploring blog post clustering under sparse participant matrices constitutes a future research direction.

References

[1] Small H. Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents[J]. Journal of the American Society for Information Science, 1973, 24(4): 265-269.

[2] White H D, Griffith B C. Author Cocitation: A Literature Measure of Intellectual Structure[J]. Journal of the American Society for Information Science, 1981, 32(3): 163-171.

[3] Callon M, Law J, Rip A. Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World[M]. Macmillan Press Ltd., 1986.

[4] Larson R R. Bibliometrics of the World Wide Web: An Exploratory Analysis of the Intellectual Structure of Cyberspace[C]. In: Proceedings of the Annual Meeting-American Society for Information Science. 1996, 33: 71-78.

[5] Wang Yuefen, Song Shuang, Lu Ning, et al. Applications of Co-occurrence Analysis in Text Knowledge Mining[J]. Journal of Library Science in China, 2007, 33(2): 59-64.

[6] Zhang Shuliang, Leng Fuhai. Study on the Applicational Development of Literature-based Knowledge Discovery[J]. Journal of the China Society for Scientific and Technical Information, 2006, 25(6): 700-712.

[7] Wang Yuefen, Song Shuang, Miao Lu. Application Study of Co-occurrence Analysis in Knowledge Service[J]. New Technology of Library and Information Service, 2006(4): 29-34.

[8] Sun Jianjun, Li Jiang. On Webometrics Theories, Tools and Applications[M]. Beijing: Science Press, 2009.

[9] Liu Y C, Wang X L, Liu B Q. A Feature Selection Algorithm for Document Clustering Based on Word Co-occurrence Frequency[C]. In: Proceedings of 2004 International Conference on Machine Learning and Cybernetics. IEEE, 2004, 5: 2963-2968.

[10] Zhang Y, Feng B Q. A Co-occurrence Based Hierarchical Method for Clustering Web Search Results[C]. Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, WI 2008, Sydney, Australia. 2008: 217-223.

[11] Wu Suhui, Cheng Ying, Zheng Yanning, et al. Improvement of K-means Algorithm Based on Co-citation Analysis[J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(1): 82-94.

[12] He X, Zha H, Ding C H Q, et al. Web Document Clustering Using Hyperlink Structures[J]. Computational Statistics & Data Analysis, 2002, 41(1): 19-45.

[13] Deng Sanhong, Gu Tingting. A Visual ACA Analysis of Core Journals in the Field of LIS[J]. Information Science, 2010, 28(11): 1728-1732.

[14] Tan Min, Xu Xin, Zhao Xing. Exploring the Co-recommendation Relationship and Its Core Structure Features of Academic Blogs—Taking ScienceNet.cn Blog as an Example[J]. New Technology of Library and Information Service, 2015(7): 24-30.

[15] Xia F, Yang Q, Li J, et al. Data Dissemination Using Interest-tree in Socially Aware Networking[J]. Computer Networks, 2015, 91: 495-507.

[16] McPherson M, Smith-Lovin L, Cook J M. Birds of a Feather: Homophily in Social Networks[J]. Annual Review of Sociology, 2001, 27(1): 415-444.

[17] Katsaros D, Dimokas N, Tassiulas L. Social Network Analysis Concepts in the Design of Wireless Ad Hoc Network Protocols[J]. IEEE Network, 2010, 24(6): 23-29.

[18] Frey B J, Dueck D. Clustering by Passing Messages Between Data Points[J]. Science, 2007, 315(5814): 972-976.

[19] Wu Suhui, Cheng Ying, Zheng Yanning, et al. Survey on K-means Algorithm[J]. New Technology of Library and Information Service, 2011(5): 28-35.

[20] Chang Peng, Feng Nan, Ma Hui. Document Clustering Algorithm Based on Word Co-occurrence[J]. Computer Engineering, 2012, 38(2): 213-214.

[21] Xiao Xinyan, Zhang Dongzhan, Gao Junjie, et al. A New Method for Web Search Results Clustering[J]. Journal of Computer Research and Development, 2007, 44(S2): 79-83.

[22] Li Fenglin, He Zhoufang. Study on Clustering of Retrieval Results Based on Co-occurrence Analysis of Keywords[J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(8): 819-825.

[23] Wang Y, Kitsuregawa M. Link Based Clustering of Web Search Results[C]. In: Proceedings of International Conference on Advances in Web-Age Information Management. Springer-Verlag, 2001: 225-236.

[24] Mukhopadhyay D, Sing S R. An Algorithm for Automatic Web-page Clustering Using Link Structures[C]. Proceedings of the IEEE INDICON Annual Conference 2004. IEEE, 2004: 472-477.

[25] Modha D S, Spangler W S. Clustering Hypertext with Applications to Web Searching[C]. In: Proceedings of the 11th ACM Conference on Hypertext and Hypermedia. ACM, 2000: 143-152.

[26] Gu Jun, Zheng Xiaodong, Zhang Lianming. Research on Bio-medical Document Clustering with Citation Information Incorporated[J]. Computer Applications and Software, 2012(10): 5-7.

[27] Brooks C H, Montanez N. Improved Annotation of the Blogosphere via Autotagging and Hierarchical Clustering[C]. In: Proceedings of the 15th International Conference on World Wide Web. ACM, 2006: 625-632.

[28] He Wenjing, He Lin. Research on Text Clustering Based on Social Tagging[J]. New Technology of Library and Information Service, 2013(7-8): 49-54.

[29] Zhang Y, Gao K, Zhang B, et al. Clustering Blog Posts Using Tags and Relations in the Blogosphere[C]. In: Proceedings of the 1st International Conference on Information Science and Engineering (ICISE). IEEE, 2010: 817-820.

[30] Chen Y H, Lu J L, Wu T Y. A Blog Clustering Approach Based on Queried Keywords[C]. In: Proceedings of the 2013 International Symposium on Biometrics and Security Technologies (ISBAST). IEEE, 2013: 1-9.

[31] Li B, Xu S, Zhang J. Enhancing Clustering Blog Documents by Utilizing Author/Reader Comments[C]. In: Proceedings of the 45th Annual Southeast Regional Conference. ACM, 2007: 94-99.

[32] Kopel M, Zgrzywa A. Search Result Clustering Using Semantic Web Data[C]. In: Proceedings of the 3rd International Conference on Intelligent Information and Database Systems. Springer Berlin Heidelberg, 2011: 395-404.

[33] Chin A, Chignell M. A Social Hypertext Model for Finding Community in Blogs[C]. In: Proceedings of the 17th Conference on Hypertext and Hypermedia. ACM, 2006: 95-104.

[34] Lin Y R, Sundaram H, Chi Y, et al. Discovery of Blog Communities Based on Mutual Awareness[C]. Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystem. 2006.

[35] Lu L, Zhu F. Blogger Clustering by Utilizing Link Information[C]. In: Proceedings of the 2010 International Conference on Intelligent Computing and Intelligent Systems (ICIS). IEEE, 2010, 2: 267-270.

[36] Bruns A, Burgess J, Highfield T, et al. Mapping the Australian Networked Public Sphere[J]. Social Science Computer Review, 2011, 29(3): 277-287.

[37] Xiao Yu, Yu Jian. Semi-Supervised Clustering Based on Affinity Propagation Algorithm[J]. Journal of Software, 2008, 19(11): 2803-2813.

[38] Zhou Lei, Yang Wei, Zhang Yufeng. Issues and Re-consideration on Cluster Analysis in Co-occurrence Matrix[J]. Journal of Intelligence, 2014, 33(6): 32-36.

[39] Hang Wenlong, Jiang Yizhang, Liu Jiefang, et al. Transfer Affinity Propagation Clustering Algorithm[J/OL]. Journal of Software, (2015-11-26).[2016-04-01]. http://www.cnki.net/kcms/detail/11.2560.TP.20151126.1606.001.html.

[40] Han Kesong, Wang Yongcheng. Methods of Keyword and Subject Concept Indexing to Chinese Full-text[J]. Journal of the China Society for Scientific and Technical Information, 2001, 20(2): 212-216.

[41] Miao Jia, Ma Jun, Chen Zhumin. A New HITS-Based Summarization Approach for Blog[J]. Journal of Chinese Information Processing, 2011, 25(1): 104-109.

[42] Guo Pengwei, Gao Kening, Zhang Bin. Public Blog Clustering Algorithm Based on Revision by Comments[J]. Journal of Northeastern University: Natural Science, 2010, 31(6): 782-785.

[43] MacQueen J. Some Methods for Classification and Analysis of Multivariate Observations[C]. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967, 1: 281-297.

[44] Han Pu, Wang Dongbo, Liu Yanyun, et al. Influence of Part-of-Speech on Chinese and English Document Clustering[J]. Journal of Chinese Information Processing, 2013, 27(2): 65-73.

[45] Wang Juan, Fan Shaoping, Zheng Chunhou. Penalized Matrix Decomposition Method for Text Clustering[J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(9): 998-1008.

[46] Manning C D, Raghavan P, Schütze H. Introduction to Information Retrieval[M]. Cambridge: Cambridge University Press, 2008.

Author Contributions

Gong Kaile: Designed research, collected data, conducted experiments, drafted manuscript; Cheng Ying: Proposed research idea, optimized research design, revised manuscript; Sun Jianjun: Proposed writing framework.

Conflict of Interest

All authors declare no conflict of interest.

Supporting Data

Supporting data is available in the online version of the journal at http://www.infotech.ac.cn: [1] Gong Kaile, Cheng Ying, Sun Jianjun. Participants.zip. Blog participant data. [2] Gong Kaile, Cheng Ying, Sun Jianjun. ParticipantCorrelationMatrix.txt. Blog participant correlation matrix. [3] Gong Kaile, Cheng Ying, Sun Jianjun. BlogText.zip. Preprocessed blog content. [4] Gong Kaile, Cheng Ying, Sun Jianjun. PositionWeightedVSM.txt. Vector space matrix with position weighting for titles, tags, main content, and comments. [5] Gong Kaile, Cheng Ying, Sun Jianjun. TitleBodyTFIDFVSM.txt. Vector space matrix with TF-IDF weighting for titles and main content. [6] Gong Kaile, Cheng Ying, Sun Jianjun. ExperimentalData.xlsx. Manual classification and clustering results.

Received: 2016-05-04
Revised: 2016-06-13