Customer Segmentation Framework for Enterprise Microblogging (Postprint)

Authors: Chen Dongyi, Zhou Zicheng, Shengyi Jiang, Wang Lianxi, Wu Jialin

Date: 2017-10-11T00:00:00+00:00

Abstract

【Objective】To effectively address the representation problem of Weibo customer characteristics for improved implementation of enterprise Weibo customer segmentation. 【Method】Leveraging the personal and social relationship attributes of customers on the Weibo platform, this study employs custom tags from customers and their friends to represent customer characteristics, and proposes a customer segmentation framework for enterprise Weibo using a text clustering method based on non-negative matrix factorization. 【Results】Experimental results demonstrate that the non-negative matrix factorization-based method achieves an average ASW index of approximately 86.130%, significantly outperforming K-means and hierarchical clustering methods. 【Limitations】The approach of representing Weibo customer characteristics solely by integrating tags from customers and their followed friends cannot comprehensively capture customer features. 【Conclusion】This work can provide valuable reference and insights for the representation, segmentation, evaluation, and result visualization of customer characteristics in enterprise Weibo customer segmentation.

Full Text

Preamble

ChinaXiv Collaborative Journal, Issue 267, 2016, No. 2

A Framework for Customer Segmentation on Enterprise Microblogs

Chen Dongyi[1],[3], Zhou Zicheng[1], Jiang Shengyi[1], Wang Lianxi[2], Wu Jialin[1]
[1](School of Informatics, Guangdong University of Foreign Studies, Guangzhou 510006, China)
[2](Guangdong University of Foreign Studies Library, Guangzhou 510420, China)
[3](S.F. Express Co. Ltd., Shenzhen 518000, China)

Abstract

[**Objective**] This study aims to effectively represent the characteristics of microblog customers to improve the implementation of customer segmentation for enterprise microblogs. [**Methods**] By leveraging customers' personal and social relationship features on the microblog platform, we propose a customer segmentation framework that uses self-defined tags from customers and their friends to represent customer characteristics, employing a text clustering method based on Non-negative Matrix Factorization (NMF). [**Results**] Experimental results demonstrate that the NMF-based method achieves an average asw index of approximately 86.130%, significantly outperforming methods based on K-means and hierarchical clustering. [**Limitations**] The approach of representing microblog customer characteristics solely by integrating personal tags and those of followed friends cannot comprehensively capture customer features. [**Conclusions**] This framework provides valuable references for the representation, segmentation, evaluation, and visualization of customer characteristics in enterprise microblog customer segmentation.

Keywords: Customer segmentation; Microblog marketing; Text clustering; Non-negative matrix factorization

1. Introduction

The rapid development of social media platforms such as microblogs has profoundly transformed communication and interaction between enterprises and customers, as well as among customers themselves. Microblogs feature fast information dissemination, strong interactivity, and real-time sharing. Leveraging these characteristics for social marketing can help enterprises improve brand image, increase visibility, and expand market share, making microblog marketing a crucial component of corporate social marketing strategies. Customer segmentation serves as an important foundation for microblog marketing.

Since American scholar Wendell Smith proposed the concept of customer segmentation in the mid-1950s, research in this area has attracted widespread attention from government agencies, industry, and academia. Currently, customer segmentation plays a vital role in enterprise customer relationship management. However, traditional enterprise customer segmentation methods have limitations, and emerging marketing approaches and electronic platforms pose new challenges to conventional segmentation techniques. In the domain of social marketing, traditional methods struggle to effectively represent customer characteristics and yield unsatisfactory analysis results when faced with massive social media data.

Building upon existing research, this paper focuses on enterprise microblogs as the research object and employs text clustering techniques to investigate a customer segmentation framework for enterprise microblogs, exploring the application of customer segmentation in social marketing. The social relationships and interest tags of microblog users are significant for representing customer

characteristics. In previous studies, many scholars have integrated self-defined tags from customers and their microblog friends to generate customer characteristic description texts from both personal and social perspectives, utilizing text classification techniques to identify potential customers on microblog platforms. Experimental results indicated that the accuracy of potential customer identification could reach approximately 86% [1]. Based on this foundation, this paper utilizes text clustering technology, combined with internal evaluation methods and tag cloud visualization, to propose a customer segmentation framework for enterprise microblogs. By analyzing fan data from enterprise official microblogs across different industries and comparing the effectiveness of various text clustering methods, results show that traditional algorithms such as K-means and hierarchical clustering tend to produce coarse-grained segmentations of enterprise microblog customers, leading to suboptimal clustering performance. In contrast, the text clustering based on Non-negative Matrix Factorization (NMF) adopted in this framework can effectively handle high-dimensional text data and semantic clustering, enabling the framework to discover more meaningful customer groups.

Corresponding Author: Zhou Zicheng, ORCID: 0000-0001-9164-9494, E-mail: ziceweek@gmail.com.

This work is supported by the National Natural Science Foundation of China project "Research on Reverse Social Emotion Recognition and Evolution Analysis for Microblog Public Events" (Grant No. 61572145), the Guangdong Provincial Science and Technology Plan Project "Research on Enterprise Competitive Intelligence Information Extraction and Situation Inference Mechanism in Guangdong Province—A Case Study of the Automotive Industry" (Grant No. 2015A030401093), and the Guangdong University Student Science and Technology Innovation Cultivation Special Fund Project "Microblog User-Generated Content Mining and Its Application in Microblog Marketing" (Grant No. 308-GK151019).

2. Related Research

2.1 Customer Segmentation Research

The rapid development of social media has provided enterprises and customers with new interactive platforms, making social media-based customer relationship building and marketing strategies key to long-term business success [2]. Customer segmentation has become one of the most important aspects of social media marketing and a primary concern for enterprise managers [3]. Customer segmentation for microblogs can help enterprises quickly analyze customer group characteristics and develop marketing channels, thereby reducing marketing costs and increasing profits.

Traditional enterprise customer segmentation methods mainly include clustering and classification approaches [4-6]. Since classification methods require large amounts of labeled training data and demand that enterprises have a good un-

derstanding of existing customer profiles and group characteristics, classification is not the mainstream approach in practical applications. Clustering analysis does not require labeled training data; it automatically partitions data based on similarity calculations and is currently the more commonly used segmentation method that can effectively discover enterprise customer group characteristics. Rajagopal used clustering techniques to identify high-profit, high-value, and low-risk customers in the retail industry [4]; Lefait et al. proposed a clustering-based customer segmentation framework to help enterprises segment customer groups based on customer purchase behavior information [5]; Wu et al. proposed different customer matrix models incorporating clustering techniques to discover various customer characteristics [6]. However, these studies primarily focus on customer segmentation applications in traditional industries and have limitations in terms of analytical methods and objects, making them difficult to extend to the social marketing domain. From a methodological perspective, both clustering and classification need to be performed under specific conditions. Partition-based clustering methods mostly require specifying the number of partitions, while classification requires large-scale labeled training data and involves parameter setting issues.

Traditional methods primarily analyze numerical attributes such as customer demographic information and consumption characteristics. However, these attributes often fail to effectively represent customer characteristics and cannot capture interests and hobbies, resulting in poor segmentation outcomes. Moreover, since it is difficult to ensure the authenticity of statistical features obtained from social media platforms, direct application of traditional methods yields unsatisfactory results.

Research on enterprise customer segmentation on social media platforms remains limited both domestically and internationally. Foreign scholars have explored user classification research on Twitter, including political stance classification, geographic segmentation, gender prediction, and role classification [7-10], but have not addressed customer segmentation. Domestic microblog development started later, and related research is even more scarce. To the best of our knowledge, there is currently no direct research on customer segmentation for enterprise microblogs either domestically or internationally.

2.2 Text Clustering and Non-negative Matrix Factorization

Text clustering is the process of grouping different documents into different categories by calculating similarity between documents based on the assumption that "documents within the same class have greater similarity, while documents from different classes have smaller similarity." Existing methods address text vectorization through the Vector Space Model (VSM) and Term Frequency-Inverse Document Frequency (TF-IDF) weight calculation, and then implement clustering using partition-based or hierarchical clustering algorithms to calculate text similarity. Common clustering algorithms include partition-based, hierarchical, density-based, and grid-based methods, with K-means and hierarchical clus-

tering being the most classic and widely used [11-13]. Although text clustering can be implemented through traditional clustering algorithms, problems remain: text features are typically high-dimensional and sparse, affecting clustering effectiveness; traditional clustering methods seldom consider text semantics, making it difficult to intuitively present clustering results.

Non-negative Matrix Factorization (NMF) attracted academic attention through a series of studies published by Lee et al. in Nature in 1999 [14]. NMF can be briefly described as: for any non-negative matrix A, find non-negative matrices W and H such that A = W × H, thereby decomposing non-negative matrix A into the product of two non-negative matrices W and H. Thus, a column vector in matrix A can be interpreted as a weighted sum of all column vectors in the left matrix W (called basis vectors), with the weight coefficients being the elements in the corresponding column vector of the right matrix H. This representation has an intuitive semantic interpretation, reflecting human thinking of "parts forming the whole." NMF' s characteristics of seeking dimensionality reduction representation and matrix factorization with non-negative elements have led to its wide application in various fields such as text mining, image processing, and bioinformatics. Drawing on the successful application of NMF in text analysis [15-19], this paper introduces NMF into the processing of enterprise microblog customer tag texts, leveraging its advantages in handling high-dimensional data and capturing text semantic information to achieve text clustering.

3. Customer Segmentation for Enterprise Microblogs

Microblogs contain rich social relationship information. The follow relationships between microblog users indicate real social connections, shared interests, or interest in the information shared by followed users. When a user follows an enterprise microblog, it typically indicates that the user is either an existing customer seeking to learn more about the enterprise' s products or services, or a potential customer interested in the products or services but not yet making a purchase. A small portion may be enterprise employees or competitors who, as industry insiders, also reflect enterprise characteristics and customer commonalities to some extent. Therefore, this paper assumes that followers of an enterprise microblog account are existing or potential customers of the enterprise with similar product or service needs, whose characteristics in lifestyle, profession, and interests can be described from different perspectives, making clustering technology suitable for discovering these underlying patterns. In other words, customer segmentation for enterprise microblogs can be viewed as the process of segmenting followers of an enterprise' s official microblog, which can be formalized as an unsupervised clustering problem.

3.1 Customer Segmentation Framework for Enterprise Microblogs

This paper combines tag text information from followers of enterprise official microblogs and their microblog friends, applies text clustering technology to cus-

tomer segmentation research on microblog platforms, and proposes a customer segmentation framework for enterprise microblogs, as shown in [Figure 1: see original paper].

The framework involves collecting tag data of followers and their microblog friends for enterprises in specific domains, constructing customer characteristic description texts using the method proposed in [1], converting customer characteristic description texts into document-term matrices through the Vector Space Model (VSM) and TF-IDF formula, applying different text clustering algorithms for segmentation to obtain different clusters, evaluating clustering results to identify meaningful clusters, and presenting results through tag cloud visualization. Finally, combining domain knowledge and expert analysis, segmentation strategies beneficial for microblog marketing can be identified.

3.2 Customer Characteristic Representation

Microblog users can freely define tags to describe their interests, which reflect users' characteristics in lifestyle and profession, thus representing personal characteristics to some extent. Since tags are user-defined, they provide more concise and accurate descriptions of personal interests. Users' interests are influenced by close friends and classmates, and conversely, friends' interests also reflect the user' s interests to a certain degree. Analogously, on microblog platforms, users' social relationships are reflected in follow relationships, forming mutual follow relationships with friends and classmates, and one-way follow relationships with media, celebrities, and public service accounts of interest. Therefore, integrating tag information from users and their microblog friends can describe customer characteristics from both personal and social relationship perspectives, leading to the proposed method of generating customer characteristic description texts based on tags from enterprise microblog followers (customers) and their followed friends [1].

Specifically, each customer characteristic description text is generated from the total frequency of tags appearing in a microblog user' s own tags and their friends' tags, calculated as follows: the user' s tag vector represents user i' s tags, and the friend' s tag vector represents friend j' s tags (user i has n friends in total), with the resulting user profile representing customer i' s characteristic description text. A user' s tag vector refers to a vector with tags as dimensions and tag frequency as dimension values. Consequently, each customer characteristic description text can be viewed as a document vector, with term weights calculated according to the TF-IDF method. Since the text vectors representing customer characteristic description texts are high-dimensional sparse data, effective dimensionality reduction of high-dimensional text data is necessary.

4. Experimental Analysis

4.1 Data Collection and Preprocessing

Experimental data was collected from the Sina Weibo platform, using followers and their friends' tags from three enterprise official microblogs across different domains. Basic information is shown in .

To prevent noise from machine-registered users ("zombie fans"), this paper removes them based on their characteristics. Typically, "zombie fans" form large numbers of one-way follow relationships by continuously following different users while having extremely few followers themselves. Additionally, normal users tend to use personalized domains (e.g., http://weibo.com/username), while "zombie fans" generally do not set their domains. Based on this analysis, we filter out "zombie fan" users using the criteria of having mutual follow counts of no less than 10 and having defined personalized user domains, thereby selecting high-quality user data. The basic information of the processed data is shown in .

4.2 Experimental Analysis Process

(1) Text Preprocessing

Before implementing text clustering, text data must be preprocessed through tokenization, stop word removal, term frequency and document frequency statistics, and text vectorization. In this experiment, text preprocessing mainly includes three aspects: First, constructing user characteristic representation texts according to the method described in Section 3.2. Second, considering the applicability and complexity of traditional dimensionality reduction methods, this paper achieves simple dimensionality reduction through two processing steps: converting tags with different cases or traditional/simplified Chinese characters, filtering stop words, and removing terms with document frequencies higher than 90% or lower than 10% in the document collection. Third, through these steps, a document collection composed of user characteristic representation texts is obtained, and based on the TF-IDF weight calculation formula, user characteristic representation texts are vectorized to produce document-term matrices as input for text clustering algorithms.

(2) Text Clustering Process

The text clustering process involves clustering algorithm selection, evaluation metrics, and parameter settings.

Clustering Method Selection

This paper selects K-means and hierarchical clustering algorithms and NMF-based clustering algorithms. K-means pre-specifies the number of clusters K and partitions data into K clusters. Hierarchical clustering algorithms organize data into groups forming corresponding tree structures for clustering. This paper adopts K-means and Ward' s hierarchical agglomerative clustering algorithm

based on within-cluster sum of squares, using cosine similarity suitable for text data.

The NMF-based clustering algorithm consists of three main steps: constructing the target matrix to be decomposed (the document-term matrix in this paper), performing non-negative matrix factorization on the target matrix to obtain the basis vector matrix W and weight coefficient matrix H, and extracting meaningful semantic clusters from the decomposed matrices.

Clustering Evaluation and Parameter K Selection

Clustering evaluation methods include external and internal evaluation. External evaluation applies to data with labeled categories, using metrics such as accuracy, recall, and F-score. Internal evaluation applies to unlabeled data, measuring intra-cluster cohesion through the sum of squared errors of samples to cluster centers and inter-cluster separation through the sum of distances between clusters, primarily using the Calinski-Harabasz Index (ch) [20] and Average Silhouette Width (asw) [21]. The ch index evaluates clustering effectiveness through the ratio of between-cluster distance sum of squares to within-cluster error sum of squares, with larger values indicating better clustering (larger between-cluster distances and smaller within-cluster distances). The asw index measures intra-cluster cohesion and inter-cluster separation by calculating the dissimilarity between data points and other points in their own cluster versus points in other clusters, with a range of [-1, 1]; values closer to 1 indicate better clustering. Since experimental data lacks category labels and considering common evaluation metrics across the three clustering methods, this paper adopts the internal metric asw for clustering evaluation.

For selecting the number of clusters K in K-means and hierarchical clustering algorithms, this paper performs clustering with different K values and selects the K value with better clustering performance based on evaluation metrics ch and asw. For NMF, it is necessary to specify the number of semantic clusters K and the matrix initialization algorithm. For K selection, we adopt the method proposed by Brunet et al. [22], performing multiple decompositions with different K values to construct a consensus matrix, and using visualized reordered consensus matrices and cophenetic correlation coefficient curves to find appropriate K values. For matrix initialization, we use random initialization and run multiple iterations to reduce decomposition instability.

Using enterprise A's dataset as an example, appropriate parameters are selected through the above methods. Based on marketing knowledge, the number of customer segments typically does not exceed 10, so K values are selected within the range [2, 10]. Since ch and asw have different value ranges ([-1, 1] and [0, $+\infty$] respectively), normalization is applied (subtracting the mean and dividing by the standard deviation) to facilitate observation of maximum values. The K values and evaluation metric curves for K-means and hierarchical clustering algorithms are shown in [Figure 2: see original paper] and [Figure 3: see original paper], indicating that both K-means and hierarchical clustering algorithms

tend to partition data into two clusters.

For K value selection in NMF, we perform 50 NMF decompositions for each K value, accumulate the connectivity matrices obtained from each decomposition to calculate the consensus matrix, reorder and plot matrix heatmaps as shown in [Figure 4: see original paper] to observe K values, and calculate cophenetic correlation coefficients from the consensus matrix to plot curves, where the coefficient measures the stability of clusters after NMF decomposition, with larger coefficients indicating better consensus matrices.

From [Figure 4: see original paper], when K = 2 or 3, data aggregates into larger dark blocks, particularly when K = 3, where data forms three dark blocks, indicating that dividing data into three semantic clusters is reasonable from the NMF perspective. When K ranges from 4 to 10, data begins to aggregate into more than three dark blocks of varying degrees, but with impure color distribution, indicating overlapping semantic clusters that can be further segmented. Notably, when K = 8, 9, 10, there are basically eight relatively obvious blocks on the diagonal, suggesting that increasing K tends to form eight blocks, making K = 8 another viable option. Combined with the cophenetic correlation coefficient curve in [Figure 5: see original paper], the coefficient is maximized when K = 3. In summary, the document-term matrix can be decomposed into 3 or 8 semantic clusters. (Note: In matrix heatmaps, color values range from 0 to 1, where 0 (light) indicates data samples are not in the same cluster, and 1 (dark) indicates data samples are in the same cluster. The color and structure of diagonal blocks can be used to observe appropriate K values.)

Similarly, K values are selected for enterprises B and C datasets using the above methods, with results shown in .

(3) Clustering Result Analysis and Visualization

By comparing the effectiveness of different clustering algorithms on enterprise microblog datasets across different industries and evaluating them using the common metric asw, the better-performing clustering algorithm is selected for different datasets, with clustering results visualized through tag clouds. Since K-means, hierarchical clustering, and NMF have different K value selections, this paper performs clustering with different K values for each algorithm.

As shown in , the NMF-based clustering algorithm far outperforms K-means and hierarchical clustering algorithms in terms of evaluation metrics. On average, the NMF-based clustering asw evaluation metric is 86.130%, significantly exceeding methods based on K-means and hierarchical clustering. Notably, when K = 2 or 3, the NMF-based clustering method tends to roughly partition text data into 2-3 clusters, which from a practical domain knowledge perspective remains coarse and can be further segmented to discover more meaningful clusters. Therefore, this paper considers alternative K value selections for NMF such as K = 8, 5, or 6. Although clustering evaluation metrics may decrease slightly, this approach facilitates the discovery of more valuable customer group information.

Based on the above analysis, this paper determines to use the NMF-based clustering algorithm for clustering text data across different industries. Using enterprise A as an example, segmentation results corresponding to different K values are extracted, as shown in and .

As shown in and , the clustering implemented through NMF yields meaningful semantic clusters. reveals that the enterprise' s microblog customer groups mainly consist of three segments: students, business professionals, and fashion enthusiasts. However, these clusters can be further segmented; for instance, the second cluster can be subdivided into travel and automotive customer groups. shows more specific and meaningful customer group features, corresponding to students, fashion enthusiasts, business professionals, travel enthusiasts, performing arts professionals, young mothers, internet industry practitioners, and creative arts enthusiasts.

Similarly, clustering is performed on datasets for enterprises B and C, with customer group keywords visualized through tag clouds as shown in [Figure 6: see original paper] and [Figure 7: see original paper]. The results reveal distinct customer group characteristics for enterprises B and C, corresponding to maternal/infant care and overseas study consulting, respectively.

4.3 Analysis and Discussion

Experimental results show that the evaluation metrics of the NMF-based text clustering method substantially exceed those of traditional clustering methods. From actual clustering effectiveness, the NMF-based method can indeed better identify different enterprise customer group characteristics. Two reasons can be hypothesized: First, traditional text clustering methods rely on the assumption of "independence between words in text"and lack semantic consideration. When confronted with high-dimensional sparse text data, these algorithms struggle to effectively calculate similarity between data objects, resulting in poor clustering effectiveness. Second, since NMF possesses characteristics of seeking dimensionality reduction representation and extracting latent semantics, applying NMF to text clustering can uncover latent semantics in document collections. Documents are represented through weighted combinations of column vectors from the decomposed basis vector matrix and weight coefficient matrix, enabling intuitive semantic interpretation. Different semantic clusters are extracted based on NMF, indirectly achieving text clustering. Through multiple iterative decompositions, precise results can be obtained, leading to better NMF-based clustering performance.

Additionally, [Figure 6: see original paper] and [Figure 7: see original paper] reveal that segmentation results may include a few similar semantic clusters. From a practical operational perspective, customers can be segmented into 5-8 groups, with appropriate merging based on granularity to obtain more reasonable results.

In the context of prevalent social marketing, this paper proposes a customer seg-

mentation framework for enterprise microblogs based on customer characteristic representation. By leveraging personal and social characteristics of microblog customers, the framework uses tags from microblog customers and their friends to represent customer characteristics and applies text clustering technology to cluster microblog customer tag texts. After evaluating clustering results, customer segmentation outcomes are presented through tag cloud visualization. Experimental results demonstrate that the NMF-based clustering algorithm significantly outperforms traditional K-means and hierarchical clustering algorithms, yielding more meaningful customer segmentation results.

However, the proposed framework remains relatively simple and requires improvement in several aspects: First, representing microblog customer characteristics solely by integrating personal tags and those of followed friends cannot comprehensively capture customer features. Future work can combine microblog customers' registration background information or microblog texts to explore better representation methods. Second, given the complexity of other clustering algorithms, this paper only considers traditional K-means, hierarchical clustering, and NMF-based methods; the effectiveness of other algorithms can be further investigated. Third, only internal evaluation methods are used to assess clustering; domain knowledge should be incorporated for comprehensive evaluation. Finally, how to apply the proposed method and framework to actual microblog marketing practice represents future research directions.

References

[1] Pang G S, Jiang S Y, Chen D Y. A Simple Integration of Social Relationship and Text Data for Identifying Potential Customers in Microblogging [A]. //Advanced Data Mining and Applications [M]. Springer Berlin Heidelberg, 2013: 45-56.

[2] Hennig-Thurau T, Malthouse E C, Friege C, et al. The Impact of New Media on Customer Relationships [J]. Journal of Service Research, 2010, 13(3): 311-330.

[3] Stelzner M A. Social Media Marketing Industry Report [EB/OL]. [2014-06-15]. http://www.socialmediaexaminer.com/SocialMediaMarketingReport2011.pdf.

[4] Rajagopal S. Customer Data Clustering Using Data Mining Technique [J]. International Journal of Database Management Systems, 2011, 3(4): 1-11.

[5] Lefait G, Kechadi T. Customer Segmentation Architecture Based on Clustering Techniques [C]. In: Proceedings of the 4th International Conference on Digital Society. IEEE, 2010: 381-386.

[6] Wu J, Lin Z. Research on Customer Segmentation Model by Clustering [C]. In: Proceedings of the 7th International Conference on Electronic Commerce. ACM, 2005: 316-318.

[7] Pennacchiotti M, Popescu A M. Democrats, Republicans and Starbucks Af-

ficionados: User Classification in Twitter [C]. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2011: 430-438.

[8] Tinati R, Carr L, Hall W, et al. Identifying Communicator Roles in Twitter[C]. In: Proceedings of the 21st International Conference Companion on World Wide Web. ACM, 2012: 1161-1168.

[9] Fink C, Kopecky J, Morawskib M. Inferring Gender from the Content of Tweets: A Region Specific Example [C]. In: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media, Dublin, Ireland. AAAI, 2012: 86-89.

[10] Steinbach M, Karypis G, Kumar V. A Comparison of Document Clustering Techniques [C]. In: Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2000: 1-20.

[11] Jain A K, Murty M N, Flynn P J. Data Clustering: A Review [J]. ACM Computing Surveys, 1999, 31(3): 264-323.

[12] Willett P. Recent Trends in Hierarchic Document Clustering: A Critical Review [J]. Information Processing and Management, 1988, 24(5): 577-597.

[13] Rao D, Yarowsky D, Shreevats A, et al. Classifying Latent User Attributes in Twitter [C]. In: Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents. ACM, 2010: 37-44.

[14] Lee D D, Seung H S. Learning the Parts of Objects by Non-negative Matrix Factorization [J]. Nature, 1999, 401(6755): 788-791.

[15] Shahnaz F, Berry M W, Pauca V P, et al. Document Clustering Using Nonnegative Matrix Factorization [J]. Information Processing & Management, 2006, 42(2): 373-386.

[16] Wang X, Tang J, Liu H. Document Clustering via Matrix Representation [C]. In: Proceedings of the 11th International Conference on Data Mining. IEEE, 2011: 804-813.

[17] Gautam B P, Shrestha D. Document Clustering Through Non-Negative Matrix Factorization: A Case Study of Hadoop for Computational Time Reduction of Large Scale [J]. Journal of Wakkanai Hokusei Gakuen University, 2010, 10(3): 15-25.

[18] Huang G S, Lu J J, Zhang Y F. Text Clustering Method Based on Non-negative Matrix Factorization [J]. Computer Engineering, 2004, 30(11): 113-114.

[19] Zhang L, Feng X S, Xiang X Z. Topic Classification of Chinese Document Based on NMF [J]. Computer Engineering, 2009, 35(13): 26-27.

[20] Calinski T, Harabasz J. A Dendrite Method for Cluster Analysis [J]. Communications in Statistics, 1974, 3(1): 1-27.

[21] Rousseeuw P J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis [J]. Journal of Computational and Applied Mathematics, 1987, 20(1): 53-65.

[22] Brunet J P, Tamayo P, Golub T, et al. Metagenes and Molecular Pattern Discovery Using Matrix Factorization [J]. Proceedings of the National Academy of Sciences (PNAS), 2004, 101(12): 4164-4169.

Supporting Data

Supporting data is self-archived by the authors and can be obtained by emailing the authors at E-mail: ziceweek@126.com.

[1] Zhou Zicheng. data_V.rar. Microblog Influencer Follow and Tag Data.

Received Date: 2015-07-27
Revised Date: 2015-09-07

Author Contributions: Chen Dongyi: Proposed research ideas and experimental design; Zhou Zicheng: Conducted experimental analysis and verification; Zhou Zicheng, Wu Jialin: Collected and processed experimental data; Chen Dongyi, Zhou Zicheng: Conducted literature review and paper writing; Jiang Shengyi, Wang Lianxi: Revised and finalized the paper.

Conflict of Interest Statement: All authors declare no conflict of interest.

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