Hadoop-based Weibo Public Opinion Monitoring System Model Research (Postprint)

Authors: Yang Aidong, Liu Dongsu

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

Abstract

【目的】Objective: To address the current big data environment, a Weibo public opinion monitoring system model based on Hadoop is proposed for the collection, mining, and monitoring analysis of massive Weibo information.

【方法】Methods: This study analyzes public opinion monitoring technologies, constructs a public opinion monitoring system model, improves relevant algorithms, utilizes Hadoop to build a big data platform, conducts simulation experiments, and validates the feasibility of the model.

【结果】Results: Experimental results demonstrate that the model can effectively monitor and analyze massive Weibo data, achieving the objective of public opinion monitoring.

【局限】Limitations: The Hadoop cluster scale is relatively small; multiple clustering algorithms were not compared, and the performance advantages and disadvantages of the improved algorithm relative to other algorithms were not obtained.

【结论】Conclusion: This model can conduct public opinion monitoring and analysis on massive Weibo data, providing scientific information support for decision-makers to respond to public opinion crises.

Full Text

Research on a Hadoop-Based Public Opinion Monitoring System Model for Micro-blogs

Yang Aidong, Liu Dongsu
(School of Economics and Management, Xidian University, Xi'an 710126, China)

Abstract

[**Objective**] This paper proposes a Hadoop-based public opinion monitoring system model for the current big data environment, enabling the collection, mining, and monitoring analysis of massive micro-blog information. [**Methods**] We analyze public opinion monitoring technologies, construct a public opinion monitoring system model, improve relevant algorithms, build a big data platform using Hadoop, and conduct simulation experiments to verify the model' s usability. [**Results**] Experimental results demonstrate that the model can effectively monitor and analyze massive micro-blog data, achieving the goal of public opinion monitoring. [**Limitations**] The Hadoop cluster scale was relatively small; we did not compare multiple clustering algorithms, nor did we evaluate the advantages and disadvantages of the improved algorithm against others. [**Conclusions**] The model can perform public opinion monitoring and analysis on massive micro-blog data, providing scientific information support for decision-makers to respond to public opinion crises.

Keywords: Public opinion monitoring; Hadoop; Micro-blog; Big data

1. Introduction

With the rapid development of the Internet, it has become the primary channel for information dissemination and an important pathway for public opinion propagation. According to the 36th "China Internet Development Statistics Report" by the China Internet Network Information Center (CNNIC), as of June 2015, China's internet user base reached 668 million, with an internet penetration rate of 48.8%. Mobile internet users continued to grow, reaching 594 million. The rapid increase in internet penetration has accelerated the development of online social networks, with social platforms such as micro-blogs becoming the backbone of information dissemination. While bringing convenience to information propagation, these platforms also pose challenges for public opinion management in China. According to incomplete statistics, the combined registered users of Tencent Weibo and Sina Weibo have reached the billion level, with daily data increments reaching the terabyte scale. The emergence of massive data and the challenge of mining and analyzing such vast quantities to obtain important information, detect sensitive content, and track hot topics have become critical research directions and significant challenges for public opinion workers in China.

In the era of big data, information is growing explosively. However, most traditional public opinion monitoring systems are based on workstations or servers, resulting in high operational costs. Conventional database solutions for massive data processing often prove expensive, poorly scalable, and vulnerable to single-point communication failures. Utilizing Hadoop big data technology to process massive data has become a popular solution. Therefore, this paper constructs a Hadoop-based micro-blog public opinion monitoring system model that can efficiently mine and analyze massive micro-blog data to achieve public opinion

monitoring, which holds practical significance.

As of December 10, 2015, a search of the China National Knowledge Infrastructure (CNKI) academic literature database, including the China Academic Journal Network Publishing Database, China Doctoral Dissertation Full-text Database, China Master' s Dissertation Full-text Database, China Important Conference Papers Full-text Database, and China Important Newspapers Full-text Database, using the keywords "micro-blog + public opinion" yielded 9,397 relevant documents. However, when adjusting the search query to "micro-blog + big data + public opinion," only 22 relevant documents were found. The search results indicate that most scholars' research focuses on the dissemination direction of micro-blog information, discussing characteristics and influence mechanisms of micro-blog information propagation. Among them, Lan Yuexin et al. [**?**] studied the information interaction between micro-blogs and other online media under the big data background through mathematical modeling. Some scholars have analyzed micro-blog information propagation characteristics based on complex network theory, such as Tian Zhanwei et al. [**?**], who used complex network theory methods to analyze the constructed micro-blog information propagation network based on degree and path statistical indicators, ultimately demonstrating that information propagation efficiency in micro-blog networks is higher than in other online social networks. Additionally, some scholars have studied micro-blog opinion leaders from a user perspective, such as Liu Zhiming et al. [**?**], who constructed a micro-blog opinion leader indicator system from the perspectives of user influence and user activity, proposing a theoretical framework using analytic hierarchy process and rough set decision analysis theory to identify and analyze opinion leader characteristics. In the area of micro-blog public opinion monitoring, Gao Chengshi et al. [**?**] proposed utilizing Sina Weibo' s existing ranking functions for audience monitoring, analyzing audience geographical distribution combined with audience emotion assessment to monitor the stability of event occurrence areas in real-time. Ma Yan [**?**] designed a micro-blog public opinion hotspot mining system structure model by analyzing the development characteristics of micro-blog public opinion and specific requirements for automatic monitoring in big data environments, describing the main functions and implementation methods of each layer. Some scholars have also conducted public opinion research based on neural networks, such as Pan Fang et al. [**?**], who constructed a BP neural network-based early warning monitoring model to address dynamic and volatile micro-blog network community emergencies. Currently, few scholars have constructed micro-blog public opinion monitoring system models based on the current big data environment to process and analyze massive data for early warning and monitoring purposes.

In China, Sina Weibo is the largest online social network and the most influential in the micro-blog domain, representing the micro-blog field. Therefore, this paper uses Sina Weibo as the data source. Based on this, we introduce the framework and structure of the micro-blog public opinion monitoring system model and conduct system model simulation.

2. System Framework

Hadoop [**?**] is an open-source project under the Apache Software Foundation. Designed by the Apache Software Foundation in 2005 as an early cloud computing platform, its open-source nature has enabled rapid development. Hadoop now has a mature community and relatively mature technology, demonstrating excellent performance in data processing efficiency, stability, and fault tolerance. Based on these characteristics, Hadoop platform users can freely develop and run applications based on massive data, significantly reducing development costs while improving performance. Hadoop is currently considered the standard for big data processing. The entire platform includes the Hadoop kernel, HDFS [**?**] (Hadoop Distributed File System), the MapReduce [**?**] parallel computing framework, and related open-source projects such as Hive data warehouse infrastructure and HBase [**?**] non-relational distributed database.

The micro-blog public opinion monitoring system model proposed in this paper is based on the Hadoop platform, using HBase as the massive data storage database. The entire model includes five components: Hadoop infrastructure, micro-blog data collection module, data preprocessing module, micro-blog public opinion monitoring analysis module, and visualization interaction, as shown in [Figure 1: see original paper].

3. System Architecture

3.1 Functional Module Analysis The micro-blog public opinion monitoring system consists of five modules: Hadoop infrastructure, micro-blog data collection module, data preprocessing module, micro-blog public opinion monitoring analysis module, and visualization interaction module. [Figure 2: see original paper] illustrates the interaction between these modules.

The functional modules are: 1. **Hadoop Infrastructure**: Provides operation interfaces for Hadoop distributed data (index library, HBase library, analysis library) and the MapReduce parallel computing framework. 2. **Micro-blog Data Collection Module**: Collects micro-blog author information, content, like counts, repost counts, original link information, etc. 3. **Data Preprocessing Module**: Performs deduplication, noise removal, Chinese word segmentation, feature extraction, and other related tasks to prepare data for monitoring and analysis. 4. **Micro-blog Public Opinion Monitoring Analysis Module**: Implements public opinion monitoring and analysis functions including text vectorization, clustering analysis of preprocessed data, and text similarity calculation. 5. **Visualization Interaction Module**: Provides user interaction based on J2EE architecture.

3.2 Key Technologies in the Micro-blog Data Collection Module The data collection module is primarily responsible for collecting Sina Weibo data, including author information, micro-blog content, and follow relationships. The main method for obtaining Sina Weibo data is through the API interface pro-

vided by Sina Weibo. Using the API interface is convenient and efficient. However, during experiments, we found that Sina Weibo' s service provider does not expose all interfaces to ordinary users, and restricts the frequency and query scope for different API interfaces. Sina Weibo limits the number of results returned per server request and the hourly API access frequency for ordinary authorized users, and rejects high-frequency API calls within short time periods. Therefore, during collection, we optimized the process by using queues and rotating multiple micro-blog accounts to solve this problem. After obtaining JSON data, we parse it to remove noise and perform deduplication before storing it in HBase. [Figure 3: see original paper] shows the data collection process through the Sina Weibo API.

3.3 Data Preprocessing Module The data preprocessing module 主要包括文本的去重去噪、中文分词生成倒排索引文件和文本特征提取三个部分. The overall process is shown in [Figure 4: see original paper].

(1) Distributed Preprocessing
Chinese lexical analysis is the key and foundation of Chinese information processing. Distributed preprocessing primarily uses the Chinese lexical analysis system ICTCLAS developed by the Institute of Computing Technology, Chinese Academy of Sciences, for text segmentation, ultimately generating inverted index files. ICTCLAS' s main functions include Chinese word segmentation, part-of-speech tagging, named entity recognition, and new word identification, while supporting user dictionaries [**?**]. Additionally, it demonstrates very high performance and accuracy for Chinese information segmentation, producing excellent results with optimistic performance when combined with Hadoop.

In this stage, we combine MapReduce with the ICTCLAS segmentation system to implement Chinese word segmentation and other functions. The Map phase is primarily responsible for mapping a line of text into several key-value pairs. In the parallel Reduce phase, it ensures that all mapped key-value pairs are grouped together based on different keys. [Figure 5: see original paper] shows the overall process flow.

The process for generating inverted index files through Chinese word segmentation is: 1. Deduplicated and denoised text is processed using the MapReduce framework. In the Map phase, parallel segmentation of the text collection is implemented. Compared with traditional ICTCLAS segmentation, this stage introduces <key, value> pairs, where key represents the 'word' after Map function processing, and value represents the 'term frequency' after Map function segmentation. 2. The Reduce function aggregates identical words, where value represents the total term frequency after aggregation of word groups. 3. The <key, value> pairs are swapped, placing term frequency before the word to enable descending order arrangement by frequency.

(2) Feature Extraction Module
The feature selection task performs dimensionality reduction on the inverted

index file obtained from text preprocessing, calculating the weight of feature words in each text to ultimately obtain a text vector collection. This paper selects the TF-IDF algorithm and implements it under MapReduce. TF-IDF is a vector space model-based classification algorithm used to evaluate the importance of a word to a document in a document collection or corpus. The importance of a word increases proportionally with its frequency in the document but decreases inversely with its frequency in the corpus [**?**]. Its advantages include a simple and convenient vector model structure, significantly improved classification accuracy with increasing data scale, excellent performance, and easy parallelization, which aligns well with Hadoop' s core philosophy of task division and parallel execution [**?**]. Under the Hadoop distributed platform, text classification can be effectively performed.

The TF-IDF algorithm processing flow under MapReduce is: 1. In the Map phase, each Mapper reads text blocks from the index file. 2. Document counts and feature word occurrence frequencies in each document are counted and output in key-value pair form. 3. Key-value pairs are locally sorted by key and sent to Reducers, which normalize TF-IDF values of all feature words with the same document ID. 4. TF-IDF values of each feature word are used as items in the text vector to construct new text vectors.

The feature extraction module process is shown in [Figure 6: see original paper].

3.4 Micro-blog Public Opinion Monitoring Analysis Module The public opinion monitoring analysis module is the core of the system, including latest news, hot topic discovery, sensitive topic detection, topic tracking, sentiment tendency analysis, public opinion trend analysis, and active author tracking. The following sections elaborate on the main functions.

(1) Hot Topic Discovery
Hot topic discovery is the focus of network public opinion analysis. It performs text clustering based on the constructed feature matrix, uses text clustering algorithms to calculate similar content, stores each cluster center and its subitems, and visualizes the clustering results. During text clustering, this paper optimizes the K-means clustering algorithm based on actual conditions: 1. Since Chinese text is being processed, to address polysemy and synonyms in Chinese, the HowNet [**?**] is combined to calculate text similarity when computing vector products during clustering to improve clustering accuracy. 2. As the K-means clustering algorithm is sensitive to changes in the K value, this paper uses the Canopy algorithm to determine the number of clusters K and cluster centers. 3. Running on the Hadoop framework accelerates text processing speed and achieves parallelization of the K-means clustering algorithm.

Through the parallelization of the optimized K-means clustering algorithm, clustering effectiveness is improved and accuracy is enhanced. The workflow is: 1. Read the feature matrix obtained from the feature extraction module. 2. Obtain cluster centers through the MapReduce-based Canopy algorithm [**?**]. 3. Calcu-

late the similarity between data objects and cluster centers using the optimized K-means algorithm. 4. Write each cluster center and its contained sub-items in the clustering results to the analysis library and perform visualization output.

(2) Sentiment Tendency Analysis
Micro-blog sentiment tendency analysis involves analyzing the speaker's attitude (or opinion, sentiment), that is, analyzing subjective information in the text [**?**]. Sentiment tendency analysis essentially uses computers to automatically judge the sentiment tendency expressed in text based on the content posted by information publishers, classifying text emotional color into three categories: positive/praiseworthy, neutral, and negative/derogatory.

This paper adopts the micro-blog sentiment tendency algorithm proposed in [**?**] to implement text sentiment analysis. This algorithm primarily improves upon the MBEWC (Micro-blog Emotion Weight Calculator) proposed by Shen [**?**], addressing the particularities of micro-blog text information by adding multiple dictionaries and constructing a new sentiment dictionary scheme, which significantly improves sentiment discrimination accuracy. The main implementation process is shown in [Figure 7: see original paper].

This process uses n Map stages and one Reduce stage, storing calculation results in the analysis library and displaying them through the user interaction module after visualization.

4. Experiments

4.1 Experimental Environment The hardware consists of four homogeneous ordinary PCs connected through a switch to build a small Hadoop cluster. Hadoop 2.2.0 was deployed on the Ubuntu system to complete cluster construction. One PC serves as the master node named HostMaster to run JobTracker and NameNode processes, while the remaining three machines named slave1, slave2, and slave3 serve as slave nodes to run TaskTracker and DataNode processes. The IP addresses of the four PCs are 172.30.78.1-172.30.78.4. The specific hardware and software configurations and node topology structure are shown in , , and [Figure 8: see original paper].

TABLE:1 Hardware Environment Configuration
- CPU: Intel(R) Core (TM) i3-3240 3.40GHz
- Hard Disk: 500GB
- Network Card: Realtek PCIe GBE Family Controller

TABLE:2 Software Environment Configuration
- Operating System: Ubuntu 14.04
- JDK: jdk1.7.0_{51}
- Hadoop: Hadoop-2.2.0
- HBase: HBase0.96
- IDE: eclipse-jee-kepler-SR1-linux-gtk-x86_{64}

Speedup is used to measure the performance and effectiveness of parallel systems

or program parallelization. It is defined as the ratio of time consumed by a task running in a single-processor system versus a parallel processor system. The speedup calculation formula is:

$$Speedup = \frac{T_1}{T_p}$$

where T_1 is the running time on a single processor (single node), and T_p is the running time on a parallel system with P processors (multiple nodes).

4.2 Data Collection The micro-blog data collection module is the data source for the entire system, and its success affects every subsequent 环节. It is impossible to collect all Sina Weibo data. The main strategy adopts breadth-first traversal of user lists: starting from a highly-followed seed user, obtaining their follow list to form the first layer of users, then obtaining the first layer users' follow lists to form the second layer, and continuously expanding to followed micro-blog users until the layer number or number of users in the current layer reaches the set value. The data collection module algorithm then obtains relevant data. In this collection, we used "Today' s Headlines" as the seed user, with data collection time set from June 1, 2015, to November 30, 2015, ultimately collecting nearly 150,000 micro-blog data items, mainly including micro-blog links, content, author personal information, fan information, repost counts, and comment counts.

4.3 Text Preprocessing After data collection, to facilitate verification, we selected 300MB of data for text preprocessing experiments, primarily to evaluate system processing efficiency for Chinese word segmentation and feature extraction under the Hadoop platform. Speedup is generally used to measure system performance under different node configurations.

(1) Chinese Word Segmentation for Inverted Index Files
This experiment primarily evaluates system processing efficiency for text preprocessing tasks such as word segmentation and inverted index construction in a distributed environment. Experiments were conducted with 1, 2, and 3 nodes to obtain time consumption for single-machine and cluster systems with different node counts. The results are shown in .

TABLE:3 Chinese Word Segmentation Processing Time and Speedup for Inverted Index Files

(2) Text Vectorization
Using feature words obtained after feature selection, text vectorization processing was performed, and speedup was calculated for different numbers of compute nodes. The experimental results are shown in .

TABLE:4 Text Vectorization Processing Time and Speedup

Speedup curves for both the Chinese word segmentation for inverted index files and text vectorization stages are plotted in [Figure 9: see original paper].

(3) Experimental Results Analysis
1. In single-node processing, speedup is slightly less than 1. This is because communication overhead between the node' s TaskTracker and DataNode processes affects performance, making it slower than a single-machine system, but the impact is not significant. This characteristic is reflected in all experimental stages. 2. As the number of nodes increases, Hadoop' s performance advantages become apparent: speedup increases and becomes larger, with various system stages running in parallel. Chinese word segmentation and text vectorization processing speeds improve significantly, indicating that more nodes lead to finer data block segmentation granularity and higher task concurrency. 3. Speedup does not increase proportionally with node count; it is slightly lower than proportional growth due to increased inter-node communication time, which affects parallel efficiency. With the small number of nodes, this characteristic would become clearer if more nodes were added.

4.4 Hot Topic Discovery and Visualization In the hot topic discovery stage, we extracted relevant fields such as author ID, posting time, micro-blog content, and collection time from the collected data. After Chinese word segmentation, feature extraction, and text vectorization, we performed clustering on micro-blog data using cosine similarity measurement. The clustering results for data from August 15, 2015, were visualized as shown in [Figure 10: see original paper]. The "+" at the center represents the day' s topics, while surrounding symbols represent authors participating in these topics.

4.5 Sentiment Tendency Analysis In the sentiment tendency analysis stage, to facilitate statistics and calculation, we randomly selected 10,000 micro-blog data items for sentiment tendency judgment. To verify algorithm accuracy, we manually annotated and judged the data. During annotation, the following process was adopted: five people completed the labeling, with each person judging the tendency of 2,000 micro-blog data items to accelerate the process. After individual annotation, micro-blog content with the same text but labeled with different sentiment tendencies was discussed to establish unified judgment criteria. Disagreements were resolved through discussion following the principle of majority rule until all data was labeled. The final statistics for the three types of micro-blog tendencies are shown in .

TABLE:5 Manual Annotation Statistics of Micro-blog Sentiment Tendencies

Using the algorithm proposed in [**?**], we calculated precision and recall. Precision evaluates algorithm accuracy, while recall evaluates the probability that the algorithm successfully identifies micro-blog texts with a certain tendency. The formulas are:

$$Precision = \frac{Correct}{Propose}$$

$$Recall = \frac{Correct}{Gold}$$

where *Correct* refers to the number of correctly classified items, *Propose* refers to the number of items submitted as belonging to that classification, and *Gold* refers to the number of manually labeled items of that classification in the sample.

and show the statistical results for precision and recall, respectively. Overall, the algorithm demonstrates high accuracy, successfully completing automatic judgment of micro-blog sentiment tendencies, which provides guidance for practical public opinion work.

TABLE:6 Precision Statistical Results
TABLE:7 Recall Statistical Results

5. Conclusion

This paper proposes a Hadoop-based micro-blog public opinion monitoring system model in response to the rapid development of micro-blog social networks. It investigates the application of Hadoop distributed storage and MapReduce parallel computing models to massive micro-blog public opinion monitoring and analysis in big data environments, and provides detailed designs of the workflow and implementation methods for each module. The main contributions are:

1. Research on key technologies for network public opinion analysis, including in-depth analysis of information collection, information preprocessing, and text clustering modules, completing the construction of the entire model framework.
2. Construction of a Hadoop cluster using ordinary PCs to conduct comparative analysis of the proposed model' s system performance under different node configurations.
3. Completion of data crawling work and successful sentiment tendency judgment of crawled micro-blogs using the algorithm proposed in [**?**].
4. Validation of the proposed Hadoop-based micro-blog public opinion monitoring system model.

Through experimental simulation, the Hadoop-based micro-blog public opinion monitoring system can effectively monitor and analyze large-scale micro-blog data. However, the following issues require further research:

1. Improving experimental conditions by expanding the Hadoop cluster and testing the model' s efficiency with more nodes.

2. Trying other clustering algorithms for comparative analysis to complete optimization research on hot topic acquisition for the Hadoop-based micro-blog public opinion monitoring system.
3. This paper' s research on the micro-blog public opinion monitoring system mainly focuses on processing micro-blog text. Future work should pay more attention to multimedia data processing to obtain greater practical value.

References

[1] Zhang Kesheng. National Decision-making: Mechanism and Public Opinion [M]. Tianjin: Tianjin Academy of Social Sciences Press, 2004: 17.

[2] China Internet Network Information Center. The 36th China Internet Development Statistics Report [R/OL]. [2015-07-23]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201507/P0

[3] Wasserman S, Faust K. Social Network Analysis: Methods and Applications [M]. Cambridge, NY: Cambridge University Press, 1994.

[4] Lan Yuexin, Dong Xilin, Su Guoqiang, et al. Research on Micro-blog Public Opinion Information Interaction Model Under the Background of Big Data [J]. New Technology of Library and Information Service, 2015(5): 24-33.

[5] Tian Zhanwei, Sui Yang. The Empirical Analysis of Micro-blog Information Flow Based on Complex Network Theory [J]. Library and Information Service, 2012, 56(8): 42-46.

[6] Liu Zhiming, Liu Lu. Recognition and Analysis of Opinion Leaders in Micro-blog Public Opinions [J]. Systems Engineering, 2011, 29(6): 8-16.

[7] Gao Chengshi, Rong Xing, Chen Yue. Research on Public Opinion Monitoring Index-system in Micro-blogging [J]. Journal of Information, 2011, 30(9): 66-70.

[8] Ma Yan. Study on the Method of Micro-blogging Public Opinion Hotspots Mining in Big Data [J]. Modern Information, 2014, 34(11): 29-33.

[9] Pan Fang, Zhang Xia, Zhong Weijun. Precautionary Monitoring of the Sudden Burst of Public Opinion in Weibo Community on Internet Based on BP Neural Network [J]. Journal of Information, 2014, 33(5): 125-128.

[10] Hadoop [EB/OL]. [2016-01-12]. http://hadoop.apache.org/.

[11] HDFS User Guide [EB/OL]. [2016-01-12]. http://hadoop.apache.org/docs/r1.2.1/hdfs_{{{user}}}_{{guide

[12] Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters [J]. Communications of the ACM, 2004, 51(1): 107-113.

[13] George L. HBase: The Definitive Guide [M]. O' Reilly Media, 2011.

[14] Song Y, Cai D F, Zhang G P, et al. Approach to Chinese Word Segmentation Based on Character-Word Joint Decoding [J]. Journal of Software, 2009, 20(9):

2366-2375.

[15] TF-IDF [EB/OL]. [2016-01-12]. http://baike.baidu.com/view/1228847.htm.

[16] HowNet Knowledge Database [EB/OL]. [2016-01-12]. http://www.keenage.com/.

[17] Li Ying' an. Research on Parallelization of Clustering Algorithm Based on MapReduce [D]. Guangzhou: Sun Yat-Sen University, 2010.

[18] Feng Xiying, Wang Laihua. Discussion of the Concept of Public Opinion and Sentiments [J]. Journal of Social Work, 2011(10): 83-87.

[19] Zhang Weishu, Lv Yunxiang. The Improvement and Implementation of the Micro-blog Sentiment Orientation Algorithm [J]. Knowledge Management Forum, 2013(9): 21-27.

[20] Shen Y. Emotion Mining Research on Micro-blog [C]. In: Proceedings of the 1st IEEE Symposium on Web Society. Lanzhou: Lanzhou University, 2009.

Conflict of Interest Statement

All authors declare no conflict of interest.

Support Data

Support data is self-archived by the authors, E-mail: yangaidongcumt@163.com.

1. Yang Aidong, Liu Dongsu. blog_{data}.rar. Crawled Sina Weibo data from June 1, 2015, to November 30, 2015.
2. Yang Aidong, Liu Dongsu. order_{index}.rar. Inverted index files obtained after Chinese word segmentation data preprocessing.
3. Yang Aidong, Liu Dongsu. attr_{reduce}.rar. Text vector sets obtained after feature selection.
4. Yang Aidong, Liu Dongsu. topic.rar. Preprocessed micro-blog data used in the hot topic discovery stage.
5. Yang Aidong, Liu Dongsu. data_{label}.rar. Manually labeled data for 10,000 micro-blog items.

Author Contributions

Liu Dongsu: Proposed research ideas, designed research methodology, revised final version of the paper.
Yang Aidong: Designed experiments, performed experimental data collection, preprocessing and analysis, wrote the paper.

Received: December 11, 2015
Revised: January 29, 2016

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