Hot Topic Identification in Mobile Complaint Texts Based on LDA Model (Postprint)

Authors: Fang Xiaofei, Huang Xiaoxi, Wang Rongbo, Chen Zhiqun, Wang Xiaohua

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

Abstract

[Purpose] Utilizing Chinese information processing and topic identification and tracking methods to extract valuable information from a large volume of mobile complaint texts.

[Method] Beginning with an analysis of the characteristics of complaint texts, k-means is first employed to cluster the texts. LDA is utilized to model each cluster, extract topics, and calculate word weights within each topic from three dimensions: word frequency, word span, and word length. The word with the highest weight is assigned as the label for that topic, and the mean document distribution probability is computed for each topic. For topics sharing identical labels, duplicate label topics are first removed based on the principle of maximum mean, then the document support rate is calculated for all remaining topics, and this document support rate is employed as the topic' s popularity metric to differentiate hot topics from general topics based on this popularity.

[Results] Temporal modeling is performed on the complaint texts. Through comparison between general topics and hot topics, it is found that the document support rate for hot topics is at least three times that of general topics, and the changing trend of the document support rate is also higher than that of general topics, demonstrating the effectiveness of the proposed algorithm.

[Limitations] The semantic relationships between topics are not considered.

[Conclusion] The preliminary approach of employing the LDA model for mobile complaint topic detection is relatively reasonable and effective, and offers certain reference value for future research in this domain.

Full Text

Preamble

Identifying Hot Topics from Mobile Complaint Texts Using an LDA Model

Fang Xiaofei[1], Huang Xiaoxi[1], Wang Rongbo[1], Chen Zhiqun[1], Wang Xiaohua[1],[2]
[1](Department of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China)
[2](China Jiliang University, Hangzhou 310018, China)

Corresponding Author: Huang Xiaoxi, ORCID: 0000-0003-4483-3664, E-mail: huangxx@hdu.edu.cn

---

Abstract

[**Objective**] This study aims to extract valuable information from large volumes of mobile complaint texts using Chinese information processing and topic identification techniques. [**Methods**] We begin by analyzing the characteristics of complaint texts and clustering them using the k-means algorithm. For each cluster, we apply LDA modeling to extract topics, calculating word weights from three dimensions: term frequency, term span, and word length. The highest-weighted word in each topic is selected as its label, and the mean document distribution probability is computed. For topics sharing the same label, we first eliminate duplicates by retaining only the topic with the highest mean probability, then calculate document support rates for all remaining topics. These support rates serve as indicators of topic popularity, enabling differentiation between hot topics and ordinary ones. [**Results**] Temporal modeling of complaint texts reveals that hot topics exhibit support document rates at least three times higher than ordinary topics, with more pronounced trends in support rate changes, demonstrating the effectiveness of our approach. [**Limitations**] The study does not consider semantic relationships between topics. [**Conclusions**] The proposed LDA-based method for mobile complaint topic detection proves reasonable and effective, offering valuable insights for future research in this domain.

Keywords: Mobile Complaints; k-means; Topic Detection; LDA Model

---

1. Introduction

With the growing popularity of the internet and continuous advancement of communication technologies, the number of online users—particularly mobile users—has increased dramatically. More people now use their phones for gaming,

microblogging, forum browsing, and news reading. Recognizing this opportunity, major telecom operators have introduced various promotional policies to attract customers and expand market share. However, as user bases grow, complaint volumes have surged, making efficient complaint text processing a focal concern across the industry.

Amidst numerous complaints, many reflect hot topics of public interest, such as "broadband," "data usage," and "billing charges." Identifying and tracking these topics can help operators understand business acceptance conditions, grasp user concerns, and implement targeted improvements—ultimately enhancing complaint handling efficiency. Therefore, topic mining from complaint texts is critically important.

Compared with news reports, mobile complaint texts are structurally more complex and significantly shorter, which increases the difficulty of topic extraction. This paper addresses mobile complaint texts specifically, applying topic identification techniques to recognize hot topics within them.

In topic detection and tracking research, the LDA (Latent Dirichlet Allocation) model has emerged as a popular direction in text mining, offering excellent dimensionality reduction capabilities, robust modeling for complex systems, and strong scalability. Topics discovered through LDA can help people understand the hidden semantics behind massive text collections and serve as input for other text analysis methods, enabling various mining tasks including text classification, topic detection, and automatic relevance judgment.

The LDA topic model's superior dimensionality reduction and solid probabilistic theoretical foundation make it particularly promising for short text mining. In recent years, numerous improvements to LDA have been proposed to enhance efficiency and accuracy, which can be categorized into vertical process extensions and horizontal model extensions. To address the brevity of microblog texts, process extension methods aggregate short texts into more suitable long documents for mining. Weng et al. aggregated all microblogs from the same user into a single long document for LDA modeling, while Hong et al. proposed user pattern modeling and term pattern modeling approaches based on training data. To adapt LDA for short text mining and mitigate data noise, various extended models have been developed, including ATM, Twitter-LDA, Labeled-LDA, MB-LDA, HLDA, and MA-LDA. Zhao et al. introduced Twitter-LDA to mine representative topics from Twitter data, while Ramage et al. proposed Labeled-LDA for incorporating label information. Zhang et al. developed MB-LDA, which considers both text and contact relationships to aid topic mining. Comprehensive comparisons of these vertical and horizontal extensions are summarized in Table 1 .

Given LDA' s inherent advantages and its effectiveness for short text topic identification, and considering that complaint texts differ from microblogs—where microblogs typically center on a single topic with comments and retweets, while complaint texts lack a clear topic and represent simple customer feedback with

brief yet complex content—this paper proposes an LDA-based approach for identifying hot topics in mobile complaint texts. We first cluster complaint texts, then apply Gibbs sampling to extract topics from each cluster, process these topics systematically, and finally identify hot topics by calculating document support rates. Experimental validation demonstrates the method' s effectiveness.

---

3. Methodology

3.1 Text Clustering

Unlike news reports, complaint texts are brief and contain limited information per individual text. To improve topic extraction, we first cluster the texts. This ensures that complaint texts within each cluster share commonalities while providing more substantial content, enabling LDA to extract topics more effectively and with greater specificity.

We employ the classic k-means clustering algorithm, which is simple and fast. Before clustering documents, we must determine the number of clusters k. The algorithm randomly selects k texts as initial cluster centers from a collection of n documents, calculates distances between each remaining text and all centers, assigns documents to their nearest cluster, and iteratively repeats this process until the criterion function is satisfied or cluster centers stabilize. This iterative process increases intra-cluster compactness while reducing inter-cluster similarity. Figure 1 [Figure 1: see original paper] illustrates the clustering workflow. After clustering with k-means, complaint texts from each category are stored in separate text files.

3.2 LDA Model Topic Extraction

In the LDA model, a topic is defined as a set of semantically related words and their probability distribution within that topic. Since direct solution of LDA' s unknown parameters is infeasible, we use Gibbs Sampling for approximate inference. Gibbs Sampling achieves convergence to true results through iterative sampling, with the key challenge being the calculation of sampling probability for the current word, as shown in Equation (1) [1].

In the equation, w represents vocabulary size; K denotes the number of topics; C_{ij} is the entry in count matrix V×K representing occurrences of word i in topic j; and C_{dj} is the entry in count matrix D×K representing the number of words from topic j in document d. Through Gibbs Sampling, we obtain posterior estimates for and ϕ, as shown in Equations (2) [1] and (3) [1].

Before parameter inference, we must preset the number of topics K. Larger K values produce more topics with finer granularity, while smaller K yields fewer, coarser topics. K significantly impacts LDA' s extraction and fitting

performance, with optimal values determinable through either word selection probability p(w|T) or perplexity. We use perplexity, where lower values indicate better topic fit, calculated as shown in Equation (4) [17].

Here, M represents the number of documents, N_i is the length of document d_i (word count), and p(d_i) is the probability of document d_i under the LDA model.

3.3 Hot Topic Identification

Gibbs Sampling yields two probability distributions: "topic-word" and "document-topic." For the "topic-word" distribution, each topic z contains words w with their probability p(w|z). For the "document-topic" distribution, each document d contains k topic probability distributions p(z|d). Gibbs Sampling typically produces numerous topics, some semantically similar and others failing to represent document content meaningfully, necessitating topic selection.

Topic Tag Word Selection. After clustering texts into H categories, Gibbs Sampling extracts several latent topics from each category. Each topic contains n topic-related words. We calculate each word' s weight within its topic based on term frequency (count), term span (cover), and word length (length) within the cluster' s texts, using Equation (5).

To prevent disproportionate influence from any single factor, we normalize these values as shown in Equations (6)-(8).

Here, count(i) is the word' s occurrence frequency; length(i) is the word length; max(length(i)) is the maximum word length in the document; last(i) and first(i) are the word' s final and initial positions; and c_{total} is the last word position in the document. After calculating word weights, we select the highest-weighted word as the topic' s tag word and store it in a database with fields for tag word, topic, and category (H).

Document Probability Distribution Mean Calculation. Gibbs Sampling produces a "document-topic" probability distribution matrix for each category, expressed as Equation (9).

The matrix contains n topics, where topic_i_{tag} is the tag word for topic_i, and avg(topic_i) is the mean document probability distribution for topic_j, with H_I, H_J, and H_K belonging to the text category set H. Since identical tag words may occur, we first group topics by tag word, considering topics within the same group as semantically similar. For groups with multiple topics, we retain only the topic with the highest distribution probability mean. We then sort topics by their means and remove those with extremely small values, as they poorly represent document content. The remaining topics constitute the global topics.

Hot Topic Identification. According to LDA' s principle, each document is

generated from several topics in certain proportions. We assume that if a preprocessed complaint text contains no fewer than a certain percentage of words from topic z, it qualifies as a supporting document for that topic. Following Xu et al.' s method [18], we calculate the document topic support rate as shown in Equation (12) [18]. If a topic' s number of supporting documents or document support rate exceeds a predetermined threshold within a time period, it is identified as a hot topic.

Here, z represents the topic, t denotes the time period, |D_i| is the number of supporting documents for topic z in period t, and |D_t| is the total number of documents in period t.

The matrix contains k topics and m documents, with each row showing k topics' distribution probabilities in one document and each column showing one topic' s distribution across m documents. From this matrix, we calculate each topic' s distribution probability mean using Equation (10).

We determine thresholds for supporting document counts or support rates through boxplot analysis [21], as illustrated in Figure 2 [Figure 2: see original paper].

Boxplots analyze data distribution and identify outliers. As shown in Figure 2, values beyond the upper and lower edges are considered outliers and disregarded. We set the upper quartile as the threshold for supporting document counts—topics exceeding this value are deemed hot topics. Boxplots rely on actual data without assuming specific distributions, authentically representing data characteristics. Based on quartiles and interquartile range, this method is robust: up to 25% of data can become arbitrarily distant without significantly affecting quartiles, making outlier identification objective and reliable.

---

4. Experiments

4.1 Data Source

Our data was provided by a telecom company' s complaint department. The experiments use complaint texts from March-April 2015, comprising over 20,000 entries for March and 50,000 for April. The March data serves for training and topic extraction, while April data validates hot topic identification effectiveness. We use Jieba for word segmentation [20] and the Harbin Institute of Technology stopword dictionary [21].

4.2 Corpus Preprocessing

(1) Custom Dictionary Construction. Existing dictionaries cannot fully recognize professional terminology and business terms in complaint data. To improve segmentation quality, we manually constructed a custom dictionary with assistance from telecom business staff. The dictionary contains 1,600 key

business terms, each represented as a triple (word, frequency, part-of-speech) using the Chinese Academy of Sciences part-of-speech tagset. Triples are space-separated, with each on a separate line in a text file. Table 2 shows dictionary examples.

(2) Noise Removal. Since complaint texts are entered by service personnel using template-based software, they contain many repetitive phrases that fail to reflect semantic information, such as "appeal," "user called to report," "customer asset number," "please process," "thank you," etc. The preprocessing workflow is shown in Figure 3 .

(3) Segmentation and Filtering. We apply regular expressions to remove complaint-specific phrases like phone numbers and ticket numbers composed of letters and digits. Using Jieba with the custom dictionary, we segment texts and retain only nouns and verbs while removing stopwords.

(4) High-Frequency Noise Removal. Additional irrelevant high-frequency terms are eliminated. Table 3 demonstrates preprocessing results.

4.3 Experimental Setup

Using fuzzy k-means clustering with k=200, we analyzed the distribution of text counts across clusters, which ranged from 45 to 362 entries per cluster, as detailed in Table 4 .

We performed parameter inference via Gibbs Sampling using the Java-based JGibbLDA tool (v.1.0) [22], with default parameters α=50/k and β=0.1, and 10 words per topic. For topic count k, we used Equation (4) with manual evaluation. Based on cluster sizes, we set k=5 for clusters with 0-50 texts, k=10 for 51-100 texts, k=20 for 101-200 texts, k=30 for 201-300 texts, and k=40 for clusters exceeding 300 texts.

After parameter configuration, we extracted topics from each cluster, obtaining "topic-word"and "document-topic"distributions exemplified in Table 5 and Figure 4 [Figure 4: see original paper].

4.4 Topic Selection Results

Through topic selection, we extracted 5,130 topics initially. After tag word extraction, document probability mean calculation, removal of low-mean topics, and retention of maximum-mean topics with identical tags, 299 global topics remained. Example results are shown in Figure 5 [Figure 5: see original paper].

4.5 Hot Topic Identification Analysis

Following the methodology in Section 3, we consider a complaint text as supporting document for topic z if it contains at least a certain percentage of words from that topic. We set this threshold at 30%—given 10 words per topic, this means at least 3 intersecting words between the preprocessed text and topic

words. Boxplot analysis of supporting document counts (Figure 6 [Figure 6: see original paper]) yields the distribution shown in Table 6 .

Based on this analysis, we define hot topics as those with at least 3,000 supporting documents. From the 299 total topics, we selected 10 hot topics and 10 ordinary topics for comparison (Table 7). The results show mobile users are particularly concerned with "internet access,""data usage,"and "billing statements," aligning with real-world user concerns and validating our topic extraction and hot identification method.

4.6 Topic Testing Experiment Analysis

Using April 2015 data to test our algorithm' s hot topic identification effectiveness, we selected three hot topics and three ordinary topics from Table 7 based on supporting document counts (from low to high). We calculated daily supporting document counts across April' s 30 days, with trends shown in Figures 7 [Figure 7: see original paper] and 8 [Figure 8: see original paper].

Comparison reveals that hot topics generally maintain higher daily supporting document counts than ordinary topics. While ordinary topics show relatively stable trends with occasional spikes, their supporting counts remain low with minimal variation amplitude. Hot topics demonstrate more pronounced trend intensity, with variation amplitudes typically exceeding 100 and featuring prominent peaks, following a clear "emergence-climax-decline" pattern.

These phenomena reflect real-world conditions: hot topics relate closely to users' daily lives and involve frequently used services, resulting in more obvious intensity variations. Trend analysis of different topics shows patterns consistent with actual situations, demonstrating our algorithm' s effectiveness in identifying hot topics from complaint texts.

---

5. Conclusion

This study achieves promising results in hot topic identification from mobile complaint texts. During preprocessing, we constructed a domain-specific dictionary valuable for future corpus processing in this field. In the hot topic discovery phase, clustering technology strengthened intra-cluster text relationships, while LDA modeling enabled fine-grained, targeted topic expression. Topic selection considered both topic representativeness and inter-topic similarity.

As an initial exploration of topic identification and tracking in the mobile complaint domain, this work has limitations, notably the lack of semantic relationship consideration between topics, relying solely on statistical methods. Future improvements will incorporate more semantic information into topic models and investigate relationships between topics to discover their evolution dynamically.

References

[1] David M B, John D L. Dynamic Topic Model[C]//Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh. 2006: 113-120.

[2] Zhang Peijing, Song Lei. Overview on Topic Modeling of Microblogs Text Based on LDA[J]. Library and Information Service, 2012, 56(24): 120-126.

[3] Weng J, Lim E P, Jiang J, et al. TwitterRank: Finding Topic-sensitive Influential Twitterers[C]//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. ACM, 2010: 261-270.

[4] Hong L, Davison B D. Empirical Study of Topic Modeling in Twitter[C]//Proceedings of the 4th International Conference on Weblogs and Social Media. 2010.

[5] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The Author-Topic Model for Authors and Documents[C]//Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2004: 487-494.

[6] Zhao W X, Jiang J, Weng J, et al. Comparing Twitter and Traditional Media Using Topic Models[C]//Proceedings of the 33rd European Conference on Information Retrieval. Springer Berlin Heidelberg, 2011: 338-349.

[7] Ramage D, Hall D, Nallapati R, et al. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 2009: 248-256.

[8] Zhang Chenyi, Sun Jianling, Ding Yiqun. Topic Mining for Microblog Based on MB-LDA Model[J]. Journal of Computer Research and Development, 2011, 48(10): 1795-1802.

[9] Tang Xiaobo, Xiang Kun. Hotspot Mining Based on LDA Model and Microblog Heat[J]. Library and Information Service, 2014, 58(5): 58-63.

[10] Zhu Ying. Hot Topic Extraction from Microblogs[D]. Chongqing: Southwest University, 2014.

[11] Wu Wankun, Wu Qinglie, Gu Jinjiang. Hot Topic Extraction from E-commerce Microblog Based on EM-LDA Integrated Model[J]. New Technology of Library and Information, 2015(11): 33-40.

[12] Rosen-Zvi M, Chemudugunta C, Griffiths T, et al. Learning Author-topic Models from Text Corpora[J]. ACM Transactions on Information Systems, 2010, 28(1): Article No.4.

[13] Zhao W X, Jiang J, He J, et al. Topical Key Phrase Extraction from Twitter[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. 2011.

[14] Ramage D, Dumais S T, Liebling D J. Characterizing Microblogs with Topic

Models[C]//Proceedings of the 1st Workshop on Social Media Analytics. ACM, 2010: 80-88.

[15] 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.

[16] Zhu Chengwen, Li Bing, Hu Kui. Algorithm of Parameter Estimation of HMM via Gibbs Sampling[J]. Computer Engineering and Applications, 2012, 48(18): 57-60.

[17] Guan Peng, Wang Yuefen. Identifying Optimal Topic Numbers from Sci-Tech Information with LDA Model[J]. New Technology of Library and Information, 2016, 32(9): 42-50.

[18] Xu Jiajun, Yang Yang, Yao Tianfang, et al. LDA Based Hot Topic Detection and Tracking for the Forum[J]. Journal of Chinese Information Processing, 2016, 30(1): 43-50.

[19] Zhang Liangjun, Wang Lu, Tan Liyun, et al. Python Practice of Data Analysis and Mining[M]. Machinery Industry Press, 2015.

[20] jieba[CP/OL].[2016-11-23]. http://www.oschina.net/p/jieba.

[21] Stop Word Dictionary by Harbin Institute of Technology[OL].[2016-11-23]. http://more.datatang.com/data/13281.

[22] JGibbLDA: A Java Implementation of Latent Dirichlet Allocation (LDA) Using Gibbs Sampling for Parameter Estimation and Inference[CP/OL].[2016-11-23]. http://sourceforge.net/projects/jgibblda.

Author Contributions

Huang Xiaoxi, Fang Xiaofei, Chen Zhiqun: Conceived the research idea and designed the study.
Fang Xiaofei, Huang Xiaoxi, Wang Rongbo: Analyzed data, conducted experiments, and drafted the manuscript.
Wang Xiaohua, Huang Xiaoxi, Chen Zhiqun: Revised the final manuscript version.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data is self-archived by the authors. E-mail: 1484514227@qq.com.

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