A Study on the Representativeness of Sentiment Orientation of Comment Clusters in Online Public Opinion*Postprint

Authors: Yang Xiaoping, Ma Qifeng, surplus capacity, Mo Yuting, Wu Jianan, Zhang Yue

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

Abstract

【Purpose】 To verify that comment clusters possess a representative role for sentiment orientation in online public opinion. 【Method】 We propose a sentiment orientation analysis model for comment cluster objects based on social network analysis. For online news events, we take user comments on news (the complete comment set) as corpus data, perform structural processing and analysis on the corpus data, establish a knowledge graph with network nodes and topological connections by leveraging the formalized relationships among comment subjects, and identify optimal comment clusters. Focusing primarily on the comment subjects within comment clusters and their corresponding comment objects, we conduct semantic analysis on core figures and their comments within the clusters, calculate the sentiment orientation of comment clusters, and compare it with the sentiment orientation of the complete comment set for the corresponding news. 【Results】 Experimental results demonstrate that the sentiment intensity between comment clusters and the complete comment set tends to be consistent, comment clusters on news exhibit good representativeness for sentiment orientation toward the news, and can improve the performance of sentiment mining algorithms for online public opinion objects by 58%. 【Limitations】 Due to the insufficiently refined methods for sentiment feature word recognition and extraction employed by the comment cluster object sentiment orientation analysis model in this paper, it leads to minor errors in Chinese word segmentation and part-of-speech tagging, errors in grammatical dependency relationships, and degree words are not taken into consideration. 【Conclusion】 Comment clusters possess a representative role for sentiment orientation in online public opinion, can enhance the performance of sentiment computation for online public opinion objects, and can flexibly and effectively reduce the time and space complexity of public opinion analysis.

Full Text

Preamble

ChinaXiv Collaborative Journals, Issues 272/273, 2016, Issues 7/8

Comment-Clusters as Representatives of Sentiment Orientation in Online Public Opinion

Yang Xiaoping, Ma Qifeng, Yu Li, Mo Yuting, Wu Jianan, Zhang Yue
(School of Information, Renmin University of China, Beijing 100872, China)

Abstract

[**Objective**] This study verifies that comment-clusters serve as representative indicators of sentiment orientation in online public opinion. [**Methods**] We propose a sentiment orientation analysis model for comment-cluster objects based on social network analysis. For online news events, we utilize user comments (the complete comment set) as corpus data, process and analyze this data structurally, and establish a knowledge graph with network nodes and topological connections through formalized relationships among commenting entities to identify optimal comment-clusters. Focusing on the commenting subjects within clusters and their corresponding comment objects, we conduct semantic analysis of core individuals and their comments to calculate the sentiment orientation of comment-clusters, which is then compared with the sentiment orientation of the complete comment set for each news item. [**Results**] Experimental results demonstrate that the sentiment intensity between comment-clusters and the complete comment set tends to be consistent. Comment-clusters exhibit strong representativeness of sentiment orientation for news items and improve the performance of online public opinion object sentiment mining algorithms by 58%. [**Limitations**] Due to imperfect identification and extraction methods for sentiment feature words in our model, minor errors occur in Chinese word segmentation and part-of-speech tagging, syntactic dependency parsing errors exist, and degree adverbs are not considered. [**Conclusions**] Comment-clusters possess representative capacity for sentiment orientation in online public opinion, can improve the performance of online public opinion sentiment computation, and effectively reduce the time and space complexity of public opinion analysis with flexibility.

Keywords: Semantic network; Knowledge graph; Core individual; Online public opinion; Comment-cluster; Sentiment orientation computing

Classification Number: TP391

Introduction

Online public opinion represents the collective sentiment formed when the public expresses viewpoints and disseminates ideas about news events through various online channels within specific timeframes. In most cases, online public opinion

originates from diverse comment sets, where the complete set of comments on a news event reflects netizens' attitudes—also known as sentiment orientation. Meanwhile, comments that receive widespread approval or rebuttal form aggregated "comment-clusters" through likes or replies, making these clusters the core focus of public opinion sentiment research.

To effectively grasp the development trends of online public opinion, comment-clusters become essential components representing sentiment orientation. The comment object refers to the subject of the comment text generated when a commenting subject expresses views on news. Key tasks in comment-cluster research include identifying opinion leaders (referred to as core individuals or core commenting subjects in this paper), extracting comment objects, and analyzing sentiment orientation. Addressing this research focus, we utilize news event comment corpora and employ social network analysis techniques, semantic network knowledge graph technology, and online public opinion sentiment mining algorithms to propose a sentiment orientation analysis model for comment-cluster objects based on social network analysis.

Research on sentiment orientation representation in online public opinion primarily includes two directions: network public opinion studies on core individual mining based on social network analysis (SNA), and corpus text mining analysis centered on machine learning that ultimately forms sentiment analysis algorithms.

First, SNA-based approaches study networks formed by information-spreading nodes to discover structural characteristics, quantitatively calculate core individuals in semantic relationship networks, and analyze public opinion propagation patterns. Shi Penghui [**?**] investigated SNA applications in online public opinion through parameter analysis. Liu Ji et al. [**?**] analyzed single-key-point, multi-key-point, and chain-type propagation modes in online public opinion, discussing the roles of strong nodes and bridge nodes in network structures. Zhao Dewei et al. [**?**] conducted holistic mining of social networks for hot topics, calculating and analyzing parameters such as degree centrality, density, diameter, and clustering coefficient, and proposed public opinion supervision recommendations.

Second, machine learning-based approaches primarily conduct semantic analysis of online disseminated information content to identify important public opinion information. Du Jiazhong et al. [**?**] proposed a network comment sentiment analysis method based on domain-specific sentiment words, constructing a feature-sentiment word ontology and comparing it with Senti-HowNet dictionary-based methods. Han Ruikai [**?**] focused on feature generation, selection, and classifier research for microblog sentiment analysis, introducing naive Bayes-based microblog sentiment analysis that treats microblogs as either single viewpoints or segmented viewpoints.

In summary, existing research in social network analysis and semantic sentiment computation demonstrates that SNA can provide rapid analysis of relevant in-

dicators for semantic sentiment computation from a non-machine-learning perspective, showing high correlation between the two approaches. Semantic sentiment analysis core indicators can be realized through social network analysis methods. This paper combines social network analysis with text mining sentiment orientation research, analyzing sentiment orientation in online public opinion comment-clusters and establishing relationships between comment-clusters and complete comment sets to verify the feasibility and effectiveness of comment-clusters as representatives of online public opinion sentiment orientation.

3. Sentiment Orientation Analysis Model for Comment-Cluster Objects Based on Social Network Analysis

During initial experiments with short text processing, we found that SVM sentiment classifiers [**?**] produced less satisfactory results for comment texts—which are short texts with sparse sentiment features and difficult-to-extract comment objects—compared to our proposed model. Therefore, in subsequent empirical studies, we developed a sentiment orientation analysis model for comment-cluster objects based on social network analysis. This model employs social network analysis knowledge inference algorithms and online public opinion sentiment mining computation through repeated experiments, verifying that comment-clusters exhibit significant sentiment orientation representativeness within complete comment sets.

Social network analysis examines network structures and attribute characteristics, including individual and holistic network properties [**?**]. This study uses complete comment sets forming online public opinion as the foundational data source, applying SNA quantitative metrics to mine core individuals (commenting subjects in our research). Based on semantic network knowledge graphs formed by information-spreading nodes [**?**] composed of commenting subjects, we derive comment-clusters corresponding to these subjects. We then apply online public opinion sentiment orientation analysis algorithms—including text preprocessing, feature extraction, structured representation of comment object sentiment, and sentiment orientation analysis—to conduct semantic sentiment orientation analysis on subjective comment texts with emotional coloring from both comment-clusters and complete comment sets, comparing the results. Through multiple experimental iterations with parameter and algorithm improvements, we achieved ideal results, accurately deriving online public opinion sentiment orientation from comment-clusters. Beyond computational performance improvements, this approach better supports public opinion heat monitoring and prediction.

The model comprises two key steps: quantitative calculation of comment-clusters and sentiment orientation calculation for comment-clusters.

3.1 Quantitative Calculation of Comment-Clusters

Quantitative calculation of comment-clusters is the critical step for identifying all core commenting subjects (core individuals) within the complete comment set and deriving corresponding comment-clusters C1 from them. Based on large volumes of unstructured comment text sets C from news items, this process employs social network analysis and knowledge inference theories to construct knowledge graphs according to interconnections among commenting subjects. Using node prestige, centrality, and comment quality metrics from the knowledge graph, we identify individuals at the network core with active participation. From these core individuals, we derive associated network relationship graphs containing each node' s corresponding commenting subject and comments, thereby identifying optimal comment-clusters C1. Since "like" and "dislike" counts significantly influence comment viewpoint orientation, we extract comments with like/dislike counts $\geq$ N as comment-clusters C2, with threshold N set accordingly. The metrics for identifying core commenting subjects in knowledge graphs are prestige, centrality, and comment quality.

(1) Prestige
Based on magnitude metadata extracted from original comment web pages, we apply SNA methods to calculate and rank the ratio of a node' s indegree to all network nodes' indegree, deriving core individuals' prestige [**?**]. Indegree represents the sum of arcs pointing to the node—in this study, the number of replies a comment receives. Using ratio methods, prestige serves as a core individual metric, with node indegree directly proportional to prestige. Higher indegree indicates higher prestige, meaning the user' s comment content receives more replies, placing the user in a more important forum position. The prestige calculation formula [**?**] is:

P(v)

where xi- represents node vi' s indegree.

(2) Centrality
Higher node centrality indicates more interaction between the commenting subject and other subjects, demonstrating greater forum activity. Active commenting subjects can drive overall network comment engagement. Following this approach and employing SNA techniques, we use the Pajek analysis platform with manual analysis and calculation to obtain standardized network centrality data, identifying highly active commenting subjects.

(3) Comment Quality
Comment quality, denoted as M, requires normalization due to significant differences in reply counts among commenting subjects, as shown in Formula (2). Defining comment quality enables more precise core individual identification, distinguishing it from prestige by focusing on standardized comparison of reply counts within local social networks. Higher weight values indicate higher comment quality.

where m represents a commenting subject' s reply count, mmax represents the maximum reply count among all commenting subjects, and mmin represents the minimum reply count.

3.2 Sentiment Orientation Calculation for Comment-Clusters

The sentiment orientation calculation phase transforms unstructured comment texts obtained from the quantitative calculation stage into structured formats. It extracts comment objects and sentiment feature words from comment texts for structured representation, then performs sentiment orientation analysis through semantic analysis algorithms. This analysis covers both complete comment sets and comment-clusters. The main steps include text preprocessing, feature extraction, structured representation of comment object sentiment, and sentiment orientation analysis.

(1) Text Preprocessing
Text preprocessing converts non-structured comment texts from complete comment sets and comment-clusters into structured data, facilitating subsequent binary group analysis. Core steps include dividing comment paragraphs into sentences, splitting sentences into phrase groups using delimiters, performing word segmentation and part-of-speech tagging using the ICTCLAS program [**?**], removing stop words, conducting word frequency statistics, and performing co-occurrence word merging analysis on both complete comment sets and quantitatively derived comment-clusters.

(2) Feature Extraction
Feature extraction involves extracting comment objects from news comment texts and their corresponding sentiment feature words [**?**]. Based on word frequency statistics, co-occurrence word merging, and grammatical part-of-speech analysis from preprocessing, we identify news comment objects. Syntactic and dependency parsing then derive reasonable phrase matching patterns for each object' s viewpoint feature words. Each short text is marked by comment object to extract semantic words representing sentiment, using , , and as sentiment orientation feature words [**?**] to obtain binary group structures:

Lexicon = (Object, Feature)

where Object represents the extracted comment object and Feature represents sentiment orientation feature words for that object.

(3) Structured Representation of Comment Object Sentiment
For the obtained binary group structure table, we use the sentiment lexicon ontology from Dalian University of Technology's Information Retrieval Laboratory [**?**] for polarity and degree matching, producing a quadruple structure:

Lexicon = (Object, Feature, Polarity, Degree)

where Polarity refers to the polarity of Object' s corresponding Feature in the sentiment lexicon ontology, and Degree refers to the polarity degree.

(4) Sentiment Orientation Analysis of Comment Objects
Separating positive and negative sentiment intensity values rather than combining them through difference operations better reflects the two-dimensional characteristics of sentiment objects, providing more specific experimental reflection and objectivity. Since the importance of comment-clusters C1 and C2 affects experimental accuracy, we adjust weight values for C1 and C2 during experiments to obtain reasonable values. Based on the quadruple structure and sentiment calculation formula, we derive sentiment orientation intensity values for comment objects in comment-clusters.

$\mathrm{SO(Object)} = \alpha \times \Sigma(\mathrm{polarity(i)} \times \mathrm{degree(i)})$ for positive words
$\mathrm{SO(Object)} = \beta \times \Sigma(\mathrm{polarity(j)} \times \mathrm{degree(j)})$ for negative words

where SO (Sentiment Orientation) represents the sentiment intensity value of sentiment feature words corresponding to comment objects; pword and nword represent positive and negative sentiment orientation words respectively; polarity(i) represents sentiment orientation word polarity; degree(i) represents sentiment orientation word degree; represents comment-cluster weight factors taking values α or β, with α for sentiment orientation words in comment-cluster C2 and β for those in comment-cluster C1.

4. Empirical Research and Results Analysis

We selected representative news comment corpora from three different domains: "Chengdu Female Driver Beaten" (2,802 comments), "Fudan Poisoning Case" (12,717 comments), and "Faye Wong and Nicholas Tse to Marry in Dali Next Month" (30,959 comments).

4.1 Comment-Cluster Extraction

For each news page' s complete comment data set, we extracted characteristic items for all commenting subjects: subject identifier, IP address, comment content, and like/dislike counts. Each news page' s complete comment set C served as the experimental comparison source data. Each commenting subject corresponded to one node, establishing a knowledge graph with network nodes and topological connections. Using three parameters—prestige, centrality, and comment quality—for each node, we identified core individual nodes. [Figure 1: see original paper] shows parameter value analysis for the three data sets' nodes, with only partial listings due to the large number of commenting subjects.

Based on these three parameter analyses, we identified core individual nodes and located all related network relationships from the knowledge graph [**?**]. [Figure 2: see original paper] presents representative network relationship graphs. Combining parameter analysis and network relationship graphs, we identified commenting subjects and comments corresponding to core individual nodes, thereby locating optimal comment-clusters C1.

For comment-clusters C2, we extracted comments with like/dislike counts $\geq$

N. Since hot news comment threads rank top comments by descending like counts, this experiment used like count as the single threshold indicator. Analysis of extensive comment data determined threshold N = 200, yielding optimal comment-clusters C2.

Since comment-clusters C1 and C2 lack objective fixed experimental weight values, we conducted repeated experiments to determine these parameters. Based on theoretical considerations, we established eight experimental parameter groups for α, β optimization. Using these eight parameter groups, we calculated one-dimensional object sentiment orientation values for the three comment-cluster sets and compared them with corresponding complete comment set C values, using Euclidean distance as the evaluation function d (see Formula (6)). The comparison sought values where comment-cluster sentiment orientation most closely matched complete comment set sentiment orientation, ultimately determining $\alpha = 0.8$ and $\beta = 0.2$.

4.2 Sentiment Orientation Calculation for Comment-Clusters

We performed word segmentation, part-of-speech tagging, stop word removal, word frequency statistics, co-occurrence word merging analysis, and grammatical part-of-speech analysis on obtained comment-clusters Ci to identify main comment objects in comment texts. Comment texts may contain pronouns like "he" or "she," which we resolved using a coreference resolution tool developed by Renmin University Information School' s Text Mining and Data Analysis Research Group before conducting syntactic and dependency parsing. This process yielded binary group relationship tables with , , and as primary sentiment orientation words, which were then matched using Dalian University of Technology' s sentiment lexicon ontology to obtain quadruple relationship tables. shows partial representative feature attribute relationships. We performed identical analysis steps on the three news items' complete comment set source data C. Error cases in identifying sentiment feature words for evaluation objects primarily resulted from insufficient program optimization causing incorrect syntactic dependency parsing for short texts, particularly in complex sentences with multiple evaluations where distances between evaluation objects and sentiment feature words were too large or feature word pairing errors occurred among different evaluation objects. These errors were corrected using the SentiRuc sentiment dictionary from our research group' s existing 成果, with synonym replacement via HowNet for minority cases where ontology matching failed, minimizing error values.

4.3 Comparative Analysis of Experimental Results

Based on the quadruple structure table and sentiment calculation formula, we calculated sentiment intensity values for comment objects in both complete comment sets C and comment-clusters Ci within their respective news network public opinions. [Figure 3: see original paper] compares these calculated sentiment intensity values.

[Figure 3: see original paper] demonstrates that after analysis and processing of the three news comment text source data sets using our model, the sentiment orientation intensity values for comment objects show consistent positive and negative sentiment intensity between comment-clusters Ci and complete comment sets C. In News 1, main comment objects were "female driver" and "male driver" ; in News 2, "Lin Senhao," "Huang Yang," and "law" ; in News 3, "Faye Wong," "Nicholas Tse," and "Cecilia Cheung." The calculation results using our social network analysis-based comment-cluster sentiment orientation analysis model yield three conclusions:

(1) Through repeated improvements to unreasonable and omitted issues in our model experiments, sentiment intensity values from comment-clusters and complete text sets for the three representative hot topics across different domains maintain proximity within ideal error ranges. Under identical source data sets, SVM sentiment classifiers produce less satisfactory results than our model. From an empirical perspective, our research arguments and model are feasible and effective.

(2) Current general calculations for comment object sentiment orientation in online public opinion derive from complete comment set source data. [Figure 3: see original paper] shows that positive and negative sentiment intensity for respective comment objects tends to be consistent between comment-clusters and complete comment sets across the three news items. Directly using comment-clusters for online public opinion sentiment orientation analysis improves computational performance by 58% and effectively reduces time and space complexity.

(3) Theoretically, highly representative public opinion reflects the common orientation of overall commentary [**?**]. This experiment verifies that comment-clusters possess ideal sentiment orientation representativeness in online public opinion, providing theoretical and practical significance for 深入研究 common orientation analysis in online public opinion.

Conclusion

Based on extensive comment corpora, this paper proposes a sentiment orientation analysis model for comment-cluster objects using social network analysis. Combining SNA and online public opinion sentiment mining algorithms, we established extraction rules for comment objects and corresponding comment features, effectively calculating comment object sentiment orientation intensity and verifying comment-clusters' representativeness theory in online public opinion. Using authentic news comment text sources with correct theories and methods ensures research authenticity and reliability. Therefore, directly employing comment-clusters to represent complete comment sets improves computational performance by 58% and effectively enhances online public opinion analysis efficiency, providing effective theoretical guidance and practical significance for 深入研究 online public opinion. However, due to imperfect sentiment feature word

identification and extraction methods, minor errors occur in Chinese word segmentation and part-of-speech tagging, syntactic dependency parsing errors exist, and degree adverbs are not considered. Future research will optimize this algorithm, incorporate degree adverbs modifying comment objects into sentiment intensity calculations, further optimize Chinese word segmentation and part-of-speech tagging, enrich ontology lexicon vocabulary, and create a more complete research system to improve sentiment information extraction completeness and sentiment intensity value accuracy.

References

[1] Shi Penghui. Empirical Studies of Network Public Opinion Based on Social Network Analysis[J]. Journal of Modern Information, 2013, 33(2): 27-31.

[2] Liu Ji, Li Lei. Analysis of Public Opinion Propagation Mode Based on Repost Behavior of Microblog Users[J]. Journal of Intelligence, 2013, 32(7): 74-77.

[3] Lin Jianghao. Research on Key Techniques of Chinese Micro-blog Sentiment Analysis[D]. Guangzhou: Guangdong University of Foreign Studies, 2013.

[4] Zhao Dewei, Xu Zhengqiao. Network Public Opinion Data Mining Based on Social Network Analysis[J]. Fujian Computer, 2014, 15(8): 15-16, 50.

[5] Du Jiazhong, Xu Jian, Liu Ying. Research on Construction of Feature-Sentiment Ontology and Sentiment Analysis[J]. New Technology of Library and Information Service, 2014(5): 74-82.

[6] Han Ruikai. Research and Application of Network Consensus Guidance System Using Community Detection[D]. Beijing: Beijing Jiaotong University, 2010.

[7] Xiao Zheng, Liu Hui, Li Bing. SVM Sentiment Classifier Based on Semantic Distance for Web Comments[J]. Computer Science, 2014, 41(9): 248-252, 284.

[8] Huang Xiaobin, Zhao Chao. Application of Text Mining Technology in Analysis of Net-Mediated Public Sentiment[J]. Information Science, 2009, 27(1): 94-99.

[9] Li Zhuozhuo, Ding Zihan. Exploring Online Opinion Leadership Based on Social Network Analysis-Public Opinion of College Student Employment Taken for Example[J]. Journal of Intelligence, 2011, 30(11): 67-70.

[10] Nie Hui. Content-oriented Evaluation and Detection for Product Reviews[J]. Library and Information Service, 2014, 58(13): 83-89.

[11] ICTCLAS Chinese Segmentation System[DB/OL]. [2013-07-02]. http://ictclas.nlpir.org.

[12] Liu Jianyi, Wang Jinghua, Wang Cong. Keyword Extraction Using Language Network[C]. In: Proceedings of the 3rd National Conference on Information Retrieval and Content Security. 2008.

[13] Yang Jing, Lin Shiping. Emotion Analysis on Text Words and Sentences Based on SVM[J]. Computer Applications and Software, 2011, 28(9): 225-228.

[14] Zhang Shengsheng, Yang Aimin, Zhou Yongmei, et al. Method of Sentiment Orientation Analysis for Micro-blogging Product Reviews[J]. Journal of Shanxi University: Natural Science Edition, 2015, 38(2): 215-222.

[15] Du Weifu. Research on Sentimental Lexicon Construction for Text Sentiment Analysis[D]. Harbin: Harbin Institute of Technology, 2010.

[16] Xu Linhong, Lin Hongfei, Pan Yu, et al. Constructing the Affective Lexicon Ontology[J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(2): 180-185.

[17] Dang Lei, Zhang Lei. Method of Discriminant for Chinese Sentence Sentiment Orientation Based on HowNet[J]. Application Research of Computers, 2010, 27(4): 1370-1372.

[18] Yang Jing, Xin Yu, Xie Zhiqiang. Semantics Social Network Community Detection Algorithm Based on Topic Comprehensive Factor Analysis[J]. Journal of Computer Research and Development, 2014, 51(3): 559-569.

[19] Ren Ren, Ji Xiaowei, Yang Bin. The Characteristics, Function and Control Method of Network Public Opinion[J]. Legal System and Society, 2013(11): 173-175.

Author Contributions

Yang Xiaoping: Proposed the research proposition
Yu Li: Designed the research methodology
Mo Yuting, Wu Jianan, Zhang Yue: Collected, cleaned, and analyzed data
Ma Qifeng: Implemented the research methodology, conducted experiments, drafted and revised the final manuscript

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data [1-3] are available in the journal' s online version at http://www.infotech.ac.cn. Supporting data [4] can be obtained via email from Dalian University of Technology Information Retrieval Laboratory at irlab@dlut.edu.cn.

[1] Yang Xiaoping, Ma Qifeng, Yu Li, Mo Yuting, Wu Jianan, Zhang Yue. news_{url}.rar. Web links for three news items from different domains.
[2] Yang Xiaoping, Ma Qifeng, Yu Li, Mo Yuting, Wu Jianan, Zhang Yue. newscluster.rar. News comment test datasets and processing.

[3] Yang Xiaoping, Ma Qifeng, Yu Li, Mo Yuting, Wu Jianan, Zhang Yue. paper.rar. Chart data from the article.
[4] Dalian University of Technology Sentiment Lexicon Ontology.

Received: January 25, 2016
Revised: April 1, 2016

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