A Public Opinion Analysis System for Urban Management Information

Guanlin Chen
School of Computer and Computing Science, Zhejiang University City College, Hangzhou, 310015, China
College of Computer Science, Zhejiang University, Hangzhou, 310027, China
Email: chenguanlin@zucc.edu.cn

Shengquan Li*
Digital Urban Management Information Center, City Management Committee, Hangzhou, 310003, China
*Corresponding author. Email: GPLEE1975@163.com

Xiaoyang Shen
School of Computer and Computing Science, Zhejiang University City College, Hangzhou, 310015, China
Email: kg20042008@126.com

Yujia Zhang
Citigroup Software Technology and Services (China) Limited, Shanghai, 201203, China
Email: yjzhang.academic@gmail.com

Gang Chen
School of Computer and Computing Science, Zhejiang University City College, Hangzhou, 310015, China
Email: hz_chengang@163.com

***Abstract*—Digital urban management has been a fast developing trend in modern cities. And as the development of the urban management, there are an increasing number of people finding that public opinion information is quite valuable and important for this process. So in this paper, we design a public opinion analysis system for urban management information (POASUMI), in which we implement certain functionalities like public opinion information acquisition, Chinese word segmentation, analysis and statistic of sentiment and hotspot of public opinion information. The experimental results indicate that the results generated by POASUMI can match actual condition very well, definitely providing quite valuable information for urban management.**

***Index Terms*—public opinion, sentiment analysis, hotspot analysis, digital urban management.**

I. Introduction

In China, the construction of public opinion thought and policy has quite a long history, however, little research endeavor was put on this field before 2003, and the research on Internet public opinion was begun on 2005.

Recent years, there are some valuable achievements emerging, which are definitely very meaningful for understanding and research on Internet public opinion. 2011, Feng Zhao et al. utilized vector space model to present the text orientation of web information and offer data-mining approaches to analyze public opinion's orientation [1]. In the same year, Xiaoling Liu et al. proposed a novel model to quantify topic experts and a new approach to identify the related blog communities on that topic [2]. 2012, Yongping Du and Changqing Yao adopted the vector center model to represent the text document for detecting and tracking the cyberspace public opinion [3]. Meanwhile, Mingjun Xin et al. proposed a quick emergency response model for micro-blog public opinion crisis oriented to mobile Internet services [4]. 2013, Jianfang Wang et al. proposed a method of extracting the opinion leader community based on the hierarchical structure [5]. In the same year, Amandeep Kaur and Vishal Gupta described the survey on main approaches for the existing opinion mining techniques [6].

However, after analyzing these methods, we found that little research on public opinion analysis has been done for urban management in Chinese languages.

In this paper, we introduce a method of the implementation of public opinion analysis system for urban management information.

As the statistics indicate, in China, there have been hundreds of cities that are carrying out digital urban management. Take Hangzhou, Zhejiang for a typical example. On Aug. 2006, Hangzhou, as one of the first batch of 'digital urban management' experimental cities, successfully passed the inspection of the Ministry of Construction of China, so that its creative model was named 'Hangzhou Model' [7]. By the end of 2010, Hangzhou had successfully built up a unified digital urban management platform that covers the whole city.

doi:10.4304/jcp.9.9.2022-2028

As a new research field, public opinion analysis for digital urban management has drawn more and more attention. By analyzing the citizens' opinion and suggestion on urban management, Public Opinion Analysis System for Urban Management Information (POASUMI) builds a mechanism to share knowledge with relevant mediums, and gathers and analyzes the public opinion of urban management to discover possible problems of urban management and law executing, which definitely provides a good reference for the promotion of urban management.

In order to provide overall, intuitive statistics and analysis for urban management, we designed and implemented a public opinion analysis system for urban management information, which integrates certain crawler and Chinese word segmentation techniques.

II. Overview of POASUMI

A. System Design Overview

All data of POASUMI are collected from information posted on various popular portals and forum websites, like the Baidu forum, Sina microblog etc. The data are obtained with the crawler of Nutch which is a well known open-source web-search software project, being preprocessed with Paoding Analyzer which is a powerful Chinese word segmentation programming library. Combining with current common sentiment and hotspot analysis methods, we finally figured out an algorithm to extract valuable information from the data, and generated more figures to implement the analysis functionality quantitatively to make the result more intuitive.

As shown in Figure 1, the main functionalities of the system involve the crawler, preprocessor, analyzer for sentiment and management system etc.

- Crawler for public opinion takes charge of collecting data from certain websites like the Baidu forum etc.
- Preprocessor implements word segmentation for public opinion information. Besides, it processes computing of public opinion meta-information and term frequency count, based on the source of the public opinion.
- Analyzer takes charge of sentiment analyzing for the information collected and the hotspots, as well as for the comparison among the public opinion.
- Public opinion report generator outputs the analysis results in form of chart or Excel report.
- Management system, customized specially for administrators, is able to monitor the analysis results from normal users, and has rollback functionality for public opinion [8].

B. Key Techniques

Following chief techniques are involved in POASUMI.

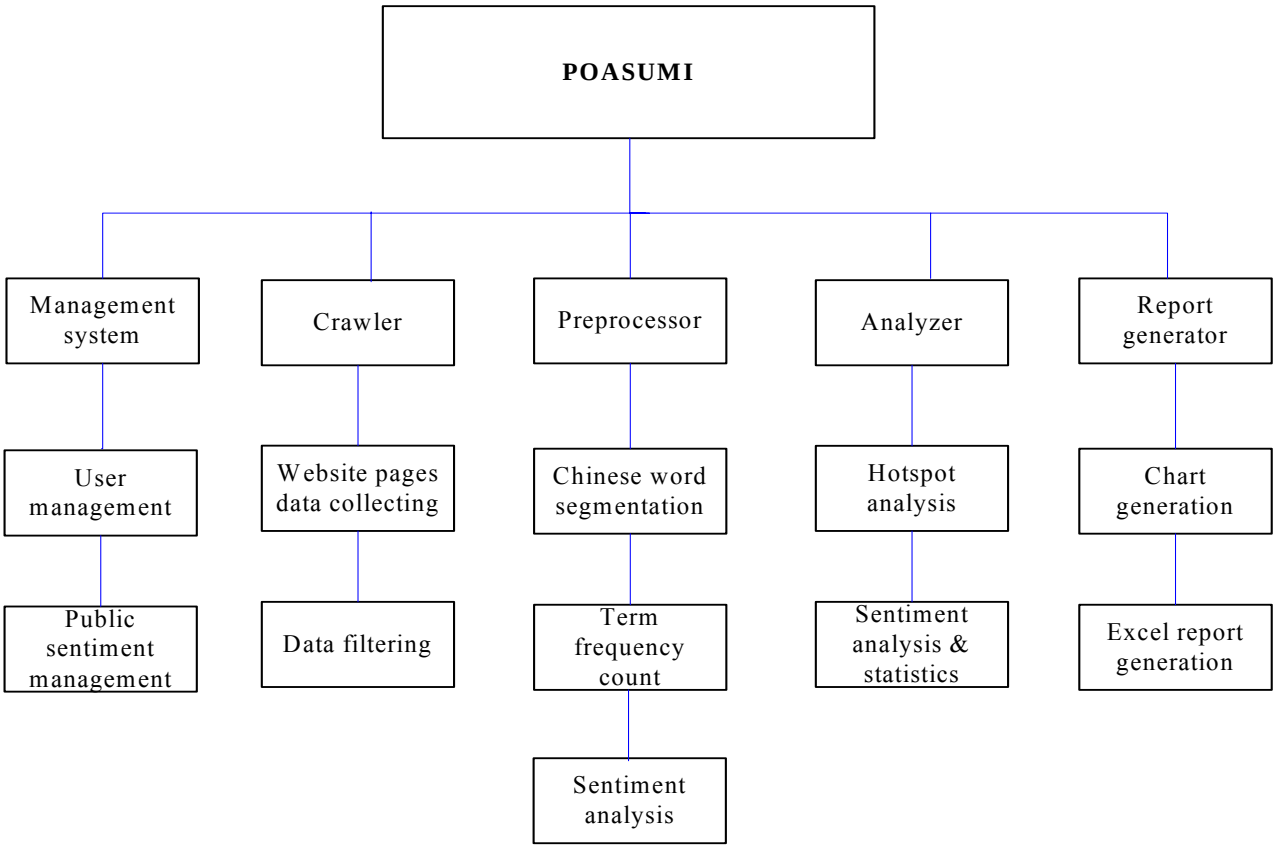

Figure 1. The basic framework of POASUMI.

(1)Referencing existing popular sentiment analysis methods, we designed an algorithm for sentiment analysis and statistics by calling the existing sentiment vector library. Besides, we combined with existing URL neighbor algorithm to implement the analysis and statistic for the hotspot of web pages.

(2)Paoding Analyzer framework is adopted to implement Chinese word segmentation functionalities, based on specific dictionary [9]. With interfaces provided by a class named Knives, word segmentation points can be located in the dictionary by the way of bidirectional matching. Besides, Knives and relevant dictionary can be turned off with the configuration file, so that they can be integrated closely together with the whole system.

(3)The system is built mainly based on the Java EE platform with the MVC design pattern. Moreover, some well-known open source development frameworks, like Struts, Hibernate, Spring, ExtJS etc, are applied to the system, which definitely improves the cohesion degree of the whole system, reducing the degree of the coupling and redundancy.

C. Database Design

Database design is quite important for the system's performance. There are four chief tables designed for storing of public opinion information, vocabulary, basic filtering words and term frequency count.

TABLE I.
THE TABLE OF PUBLIC OPINION INFORMATION

Field	Type	Meaning
TIPID	Varchar(100)	Public opinion information ID
TIPCONTENT	Varchar(4000)	Public opinion information content
TIPDATE	Varchar(50)	Public opinion published date
FROMWEB	Varchar(300)	Public opinion source website information
TIPEMOTION	Varchar(100)	Sentiment value
FROMWEB_URL	Varchar(100)	Public opinion source
HAS_WORDDIV	Varchar(2)	Word segmentation status

• Public opinion information table is mainly used for storing data collected by crawler from the Internet and some other information inputted manually.

• Vocabulary table is used to store information of words from word segmentation of public opinion information, like source information of words. Besides, after sentiment analysis, the sentiment value for each word will also be stored in this table.

• Basic filtering words table is a vocabulary table that needs to be updated constantly. Insignificant public opinion words are stored in this table, so that they can be used to filter those noisy words.

• Term frequency count table is used to store hotspot meta-information after term frequency counting. Combined with the vocabulary and corresponding source, term frequency will be counted and classified [10].

Among these four tables, the public opinion information table and term frequency count table are most important, which are related to the next processes of sentiment and hotspot analysis. Table 1 indicates the detail of the public opinion information table.

Column TIPEMOTION and column HAS_WORDDIV are the two most important fields for this table.

• The default value of the column TIPEMOTION is NULL. The process of sentiment analysis involves word segmentation of public opinion information, double times of filtering and counting after sentiment value computing. And after the administrator processes the transaction on the front end page of the system, the result of word segmentation will be cleaned up.

• HAS_WORDDIV column is a status field, that is, to mark if word segmentation has been processed. When the transaction is a rollback, this field will be set as NULL, and all the relevant transaction will be removed.

III. SYSTEM DESIGN AND IMPLEMENTATION

POASUMI is developed based on the Java EE platform with MVC design pattern [11]. Following is the detail of the system design and implementation [12].

A. Management System

This module is specially designed for administrators, mainly providing the functionalities like data adding, deleting and updating. And it also allows administrators to maintain information of users and their authorizations. Besides, administrators are able to audit the analysis result from normal users and process rollback for some specific transactions.

System handles the request from the front end mainly through the transaction tag of Service tier, which centralized the functionalities, reducing the degree of coupling. Moreover, in Service tier, the design for database transaction also involves logics of rollback for multi-table and joined deletion among several tables.

B. Public Opinion Preproces

Public opinion preprocess is the first, and also the most important step of public opinion analysis. With the dictionary files of the Paoding Analyzer, noisy words can be filtered out. At the same time, using the existed sentiment vectors to analyze positive and negative words, analysis for a single word will be processed. Then the second step of filtering is conducted, filtering sentiment words with basic words dictionary. Finally, term frequency is counted within rest words.

(1) Word segmentation for public opinion information

The system mainly applies the dictionary and word segmentation class named Knives provided by Paoding

Analyzer to implement the analysis of public opinion information. The dictionary is loaded through the configuration of the mapping file, which is shown as Table 2.

TABLE II.
THE DICTIONARY FILES OF PAODING ANALYZER

File name	Description
x-noise-charactor.dic	To filter the insignificant part of a word
x-noise-word.dic	To filter a whole word with insignificant part
t-base.dic	To filter all the base words in it
festival.dic	To filter the word about festivals
paoding-dic-names.properties	To choose any files end up with '.dic' to be dispatched or not

(2) Sentiment calculating for public opinion

We designed an algorism to calculate the sentiment value for public opinion as following. Sentiment analysis is carried out through analyzing words in a triple and finally returning a sum of the analyzing results. From the common practice on the Internet, we counted that there are few people using the inversion of word order when they are posting, therefore we did not take this case into consideration in our experiment. Normally, a sentence with sentiment mainly involves three parts, that is, adjunct word for sentiment (e.g. "HEN" meaning "quite", "JIQI" meaning "extremely", "FEICHANG" meaning "very" in Chinese etc.), negative word and sentiment direction word (e.g. "GEILI", "BU GEILI" etc.). Nevertheless, the different orders of these three types of words and the default language habits of people may bring about various analyzing results, like "HEN BU HAO" (meaning "very bad" in Chinese)" and "BU SHI HEN HAO" (meaning "not very good" in Chinese). In this case, the same word meat-information is used in these two phrases, but due to the different orders of words, they express a totally different meaning, at least on the degree of sentiment [13].

Here, that the central word of the triple is counted as the negative word is taken as an example. And following is the implementation of the sentiment calculating for the process.

```
public int gettriplevolget(String a, String b, String c) {
int res1 = 0, res2 = 0, res3 = 0, res = 0;
if(tipsneganal(b)){
                res2 = -1;
if(isdeny(a)){
                        res1 = -1;
                        res = res2 * res1;
                        return res;
                }
                else if(tipsemotionadv(a) > 0){
                        res1 = tipsemotionadv(a);
                        res = res1 * res2;
                        return res;
                }
                else{
                        res = res2;
                        return res;
                }
        }
…
}
```

In this piece of codes, the triple is applied to calculate sentiment. Firstly, each unit of the triple is set with 0, and then if the central word is negative, the second unit will be set to -1. For the first unit, if it is negative, it will also be set to -1. Thus, if both of the values are -1, they are multiplied with each other to be a positive value. Besides, if there is adjunct word before the negative words, they will be multiplied by corresponding weights to indicate a negative value. If the central word is positive, the calculating method is nearly the same as above.

(3) Term frequency counting for public opinion information

Term frequency counting is mainly based on hotspot meta-information. Hotspot meta-information is not identified with a single primary key, but represented in form of two-dimension with the words after being filtered and the corresponding sources of them.

The basis of the process is just the information of each word in the table of basic filtering words mentioned above. This table is specially used for comparison. If any same cases exist, they will be filtered out, and not be involved in the analyzing and counting processes in the next steps.

Figure 2 shows the result of the term frequency counting.

C. Analysis for Public Opinion Information

	Hotspot Meta ID	Hotspot Meta Content	Hotspot Source ID	Hotspot Frequency
1	2c9bb825-7ca4-4543-...	黑海	358	1
2	0370bec2-c558-4190-...	赏雪	358	1
3	ee1ac1ef-e1c2-4dfc-a...	贝加尔湖	358	1
4	2f68f677-35d8-4bba-...	湖面	358	1
5	18de3973-7602-4d5e...	海北	358	1
6	730e609b-1676-4670...	指天	358	1
7	a174238b-93f9-42a6-...	张弓	358	1
8	da979588-8550-43b9...	库页岛	358	1
9	a24e39d3-78bc-4693...	岛上	358	1
10	166db3ba-7d73-4855...	山西	358	1
11	79102a23-3276-4418...	天山	358	1

Figure 2. The result of the term frequency counting.

The analysis for public opinion information is mainly carried out for the sentiment and the hotspots, which is the core part of the whole system. The degree of accuracy depends on the preprocessing for the public opinion, whether the result of the analysis for public opinion unit and hotspot is accurate or inaccurate.

(1) Sentiment analysis for public opinion information

The data source of the sentiment is coming from the analytical data of the public opinion unit. Here, we apply methods of two dimensional analyzing and single dimensional result indicating, which definitely makes the analysis algorithm more accurate and the representation of the result more understandable.

Moreover, the source of the public opinion and the values of the sentiment direction are indicated with the bar charts. And we apply the well-known framework ExtJS to implement those charts, encapsulating the data with JSON, and the data will be sent to the client end, finally being shown in form of bar charts.

It is worth mentioning that we compress the value extension of the axis of the charts and apply desired value method to solve the problem of chart overflow. And the unit of the analysis is based on the detail information of each public opinion, which will be indicated with the X axis of the final chart, and the corresponding value will be presented with the Y axis. Figure 3 shows the analysis result of one case of our experiment.

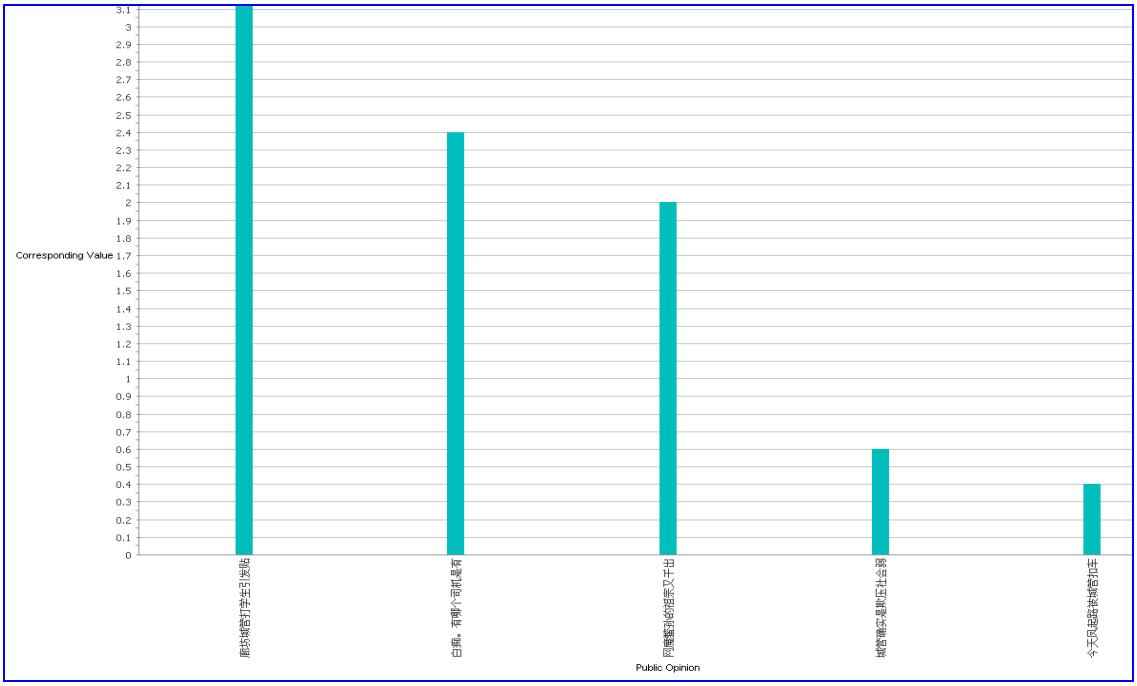

Figure 3. Sentiment analysis result of public opinion information.

(2) Hotspot analysis for public opinion information

Hotspot analysis is another analysis process based on filtering after term frequency counting. Because going through the filtering of the Paoding Analyzer and the comparing filtering from term frequency counting with basic filtering words table, the vocabulary information has been more practical and to-the-point. However, some noisy words still exist, so we should restrain the length of the hotspot words, limiting the length of the words between 2 and 6, which can filter out many invalid words.

Hotspot analysis is based on two dimensions, time and source. The algorithm that is involved in the analysis is mainly based on the simplified URL neighbor algorithm. Because current common URL neighbor algorithms are relatively rough, which are based on the comparison between the characters of the URL. Therefore, we do some modification for the algorithm, cutting the protocol head of the URL, and do the comparison with the real content of the URL, calculating the amount of the separating characters and the difference of characters, and finally we can get a specific factor value to calculate the hotspots' neighbor degree. And then hot degree for each hotspot is calculated by multiplying their hot degree by the factor value, so the result of the hotspot analysis is figured out, which can be seen in Figure 4.

(3) The statistic of sentiment for public opinion information

The statistic module is built based on the analysis module. Because the sentiment values are quite different from each other, we also transform the positives value to desired values to qualify the data and finally present them. The result is indicated in Figure 5.

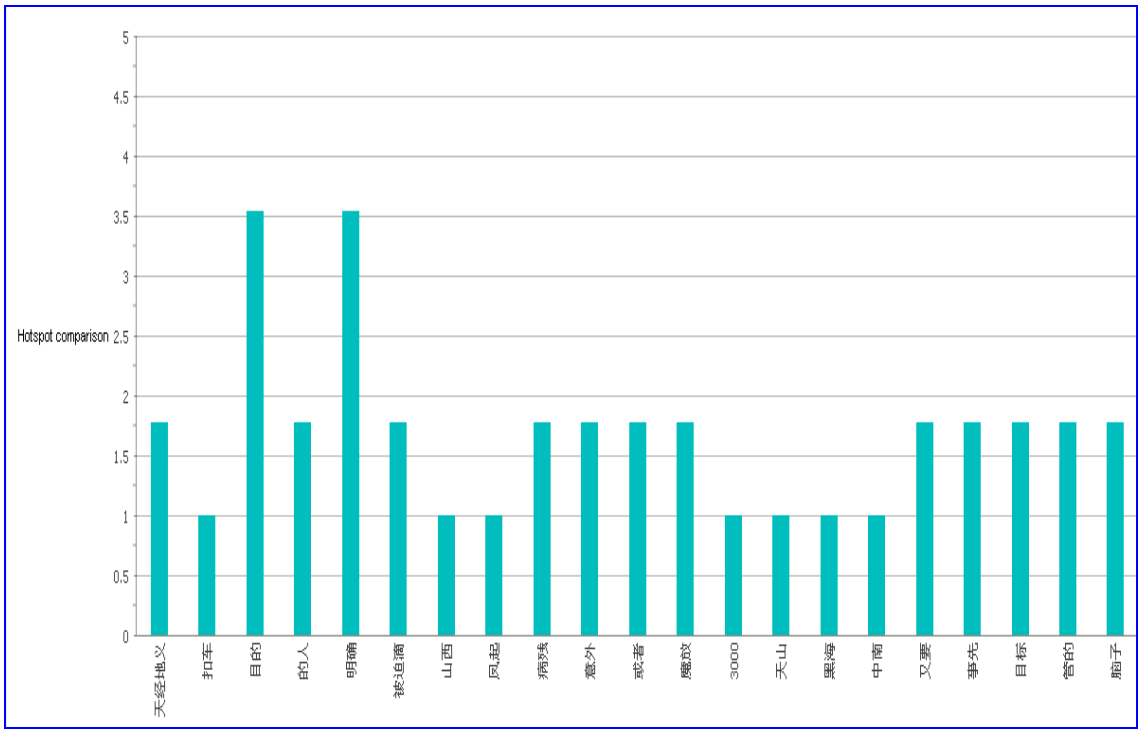

Figure 4. Hotspot analysis result of public opinion information

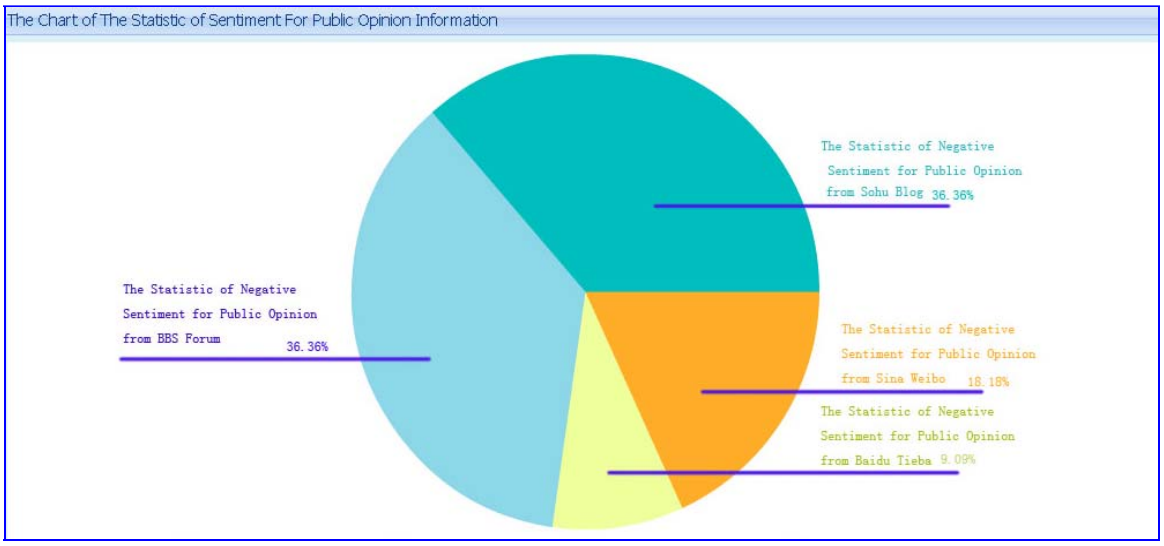

Figure 5. Sentiment analysis result of public opinion information.

VI. CONCLUSIONS

The analysis and statistic for public opinion of urban management information is a new research field, which definitely has quite a large development space. In this paper, we introduce a public opinion analysis method for urban management, and at the same time, we finally implement a public opinion analysis system for urban management information with functionalities of public opinion crawler, preprocessor and analyzer. In future work, the endeavor should be put on how to extend the analysis dimension, optimize the algorithm for each dimension and improve the reliability of the analysis of the system.

ACKNOWLEDGMENT

This work is partially supported by the 2013 Construction Research Project of Zhejiang Provence, China (No. 18) and the Key Research Project of Zhejiang Federation of Humanities and Social Sciences Circles, China (No. 2012Z53).

REFERENCES

[1] Feng Zhao, Qianqiao Hu, Xiaolin Xu, "Orientation Mining-Driven Approach to Analyze Web Public Sentiment", Journal of Software, vol. 6, no. 8, pp.1417-1428, 2011.

[2] Xiaoling Liu, Yitong Wang, Yujia Li, Baile Shi, "Identifying Topic Experts and Topic Communities in the Blogspace", Lecture Notes in Computer Science, vol. 6587, pp.68-77, 2011.

[3] Yongping Du, Changqing Yao, "Performance Evaluation of the Cyberspace Public Opinion Detection and Tracking", Journal of Computers, vol. 7, no. 5, pp.1284-1288, 2012.

[4] Mingjun Xin, Hanxiang Wu, Zhihua Niu, "A Quick Emergency Response Model for Micro-blog Public Opinion Crisis Based on Text Sentiment Intensity", Journal of Software, vol. 7, no. 6, pp.1413-1420, 2012.

[5] Jianfang Wang, Xiao Jia, Longbo Zhang, "Identifying and Evaluating the Internet Opinion Leader Community Through k-clique Clustering", Journal of Computers, vol. 8, no. 9, pp. 2284-2289, 2013.

[6] Amandeep Kaur, Vishal Gupta, "A Survey on Sentiment Analysis and Opinion Mining Techniques", Journal of Emerging Technologies in Web Intelligence, vol. 5, no. 4, pp.367-371, 2013.

[7] Guanlin Chen, Meiying Zhao, Wenyong Weng, "Analysis on the Development Stages and Trend of Digital Urban Management in Zhejiang Province", in Proceedings of the International Conference on E-Business and E-Government(ICEE 2010), pp.3734-3737, 2010.

[8] Quanlong Guan, Saizhi Ye, Guoxiang Yao, Huanming Zhang, Linfeng Wei, Gazi Song, Kejing He, "Research and Design of Internet Public Opinion Analysis System", in Proceedings of the 2009 IITA International Conference on Services Science, Management and Engineering(SSME 2009), pp.173-177, 2009.

[9] Chengliang Wang, Juanjuan Chen, Xichuan Wu, "Dictionary Chinese Word Segmentation Research a Method Combined with CRFs", in Proceedings of the 5th International Conference on Computer Sciences and Convergence Information Technology(ICCIT 2010), pp.962-965, 2010.

[10] Yanjun Li, Soon M. Chung, John D. Holt, "Text Document Clustering Based on Frequent Word meaning sequences", Data & Knowledge Engineering, vol. 64, no. 1, pp.381-404, 2008.

[11] Askar S. Boranbayev, "Defining Methodologies for Developing J2EE Web Based Information Systems", Nonlinear Analysis: Theory, Methods & Applications, vol. 71, no. 12, pp.1633-1637, 2009.

[12] N. Hritonenko, Yu.Yatsenko, "Creative Destruction of Computing Systems: Analysis and Modeling", Journal of Supercomputing, vol. 38, no. 2, pp.143-154, 2006.

[13] Jianping Zeng, Shiyong Zhang, Chengrong Wu, Jianfeng Xie, "Predictive Model for Internet Public Opinion", in Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery(FSKD 2007), pp.7-11, 2007.