International Journal of Multimedia and Ubiquitous Engineering
Vol. 9, No. 11 (2014), pp. 39-52
http://dx.doi.org/10.14257/ijmue.2014.9.11.5

# IPOMAS: An Internet Public Opinion Monitoring and Analyzing System

Guanlin Chen[1,2 *] and Panqing Huang[1]

*[1]School of Computer and Computing Science, Zhejiang University City College, Hangzhou, 310015, P.R. China*
*[2]College of Computer Science, Zhejiang University, Hangzhou, 310027, P.R. China*
*[*]chenguanlin@zucc.edu.cn*

## Abstract

*With the rapid development of Internet technology, the Internet has become an essential part of our everyday life. However, Internet can pose potential threats to government authorities, economic development or social stability in any country if false information, unsolicited or malicious public opinions are uncontrollably disseminated in large scale on the Internet, thanks to the attributes of low cost and anonymity of information exchange on the Internet. It is evident that an ineffective supervision and monitor of public opinions in the virtual cyber space can lead to serious social crisis or social instability in our real life. This paper is concluded by a summary of IPOMAS (Internet Public Opinion Monitoring and Analyzing System) implementation including the technical challenges, lessons learned and project outcomes through the lifecycle of the design, development and implementation of IPOMAS.*

***Keywords:*** *Internet public opinion, Sentiment analysis, Topic clustering, MongoDB, Node.js*

## 1. Introduction

With the rapid development of the Internet on a global scale, Internet media has become new Fourth Estate following tradition media including printing press, radio and television. Taking advantage of the convenient and superfast information highway, the netizens [1] frequently and freely exchange their personal opinions on public events. It is estimated that Chinese Wechat users will total 1,000 million by 2014 while the total number of monthly active Facebook users has amounted to 680 million by 1 January 2014. The Internet has become the provenance where people voice their opinions or disseminate information or ideas on their tablets, PCs or mobile phones. Effectively the netizens play the role of information generators in the vast Internet world where public opinions can be sent to millions of people by one simple touch or click.

Public opinion, by definition, is “an aggregate of the individual views, attitudes, and beliefs about a particular topic, expressed by a significant proportion of a community” (Britannica Encyclopedia Online). Internet public opinion has taken it one step forward in terms of its magnitude and velocity of dissemination. Personal comments or opinions about the social events or political issues in the real life can be easily read, viewed or heard on the Internet. Vice versa, strong influence, the tendency of comments or opinions can be easily broadcast from the Internet to real social life. Internet public opinion is the mirror image of

[*] Corresponding Author

ISSN: 1975-0080 IJMUE

social public opinion in cyberspace and a direct manifestation of social sentiment shared by most people.

With the unrivalled advantage over the traditional media, the Internet public opinion has increasingly significant impacts on political order and social stability in most of the countries, thanks to its expressiveness, ease of use, content diversity and extravagant interactivity. Therefore, it is useful for the healthy development of Internet society to have a mechanism for monitoring and analyzing public opinions in the cyberspace so that trends of public option can be understood and relevant strategies be developed for responding or managing the public events.

Recently there are some valuable researches on Internet public opinion. 2012, Mingjun Xin et al. proposed a quick emergency response model (QREM) for micro-blog public opinion crisis under mobile Internet environment [2]. 2013, Jianfang Wang et al. drew the community components from the replies of every post in BBS and proposed a method of extracting the opinion leader community (OLC) based on the hierarchical structure [3]. 2014, Feng Cao et al. proposed an Internet ecological monitoring and response system model based on the theory of Internet public opinion ecology system's constitutes [4].

However, these studies mostly focus on building models which respond to Internet public opinion, and there is little specific description of how to technically develop a system of monitoring and/or analyzing Internet public opinion.

The paper is intended to present a technical overview of the Internet Public Opinion Monitoring and Analyzing System (IPOMAS) to cover both of its technical and functional capabilities including system design concepts, system design framework, and primary techniques applied as well as the technical modules and their specific implementation.

Consisting of the distributed web crawlers-based on C/S model and a server bus based on B/S model, IPOMAS was designed to achieve both the collection and analysis of public opinions on the Internet using Java and Node.js languages [5], MongoDB Distributed Database and Lucene [6] keyword segmentation in Chinese language.

## 2. Overview of IPOMAS

### 2.1. Design Overview for System Requirements

The IPOMAS system design requirements can be delineated in the following aspects:

a) An integrated solution for rapidly crawling on the Internet to collect public opinions published on the Internet and achieves directional information gathering and intelligent analysis over news page, forum, blog, web comments, Weibo, Baidu Post Bar.

b) Comprehensive utilization of Internet information gathering and processing technology, intelligent information processing techniques and full-text indexing technology.

c) Provisioning multidimensional and multilayered Internet public opinion services in public opinion information indexing, the tracking and positioning of social hot spots or events and the overall statistical analysis of the relevant public opinions.

d) Informing IPOMAS users of public opinion trending in a timely manner and providing information resources for decision making.

The system consists of the following four main modules:

a) Distributed Web Crawler module [7]. The module manages clusters of vertical crawlers designed for both data crawling and indexing of Chinese word segmentation [8] where each crawler module targets a specific crawling domain. The crawler module follows a uniform interface conducive to the crawler data unification and the expansion of the crawler module.

b) System API (Application Programming Interface) module. The module is based on B/S model and communicates through the API with the format of JSON under Restful principles. The system API provides convenient access for the crawlers to execute data communication and configuration.

c) Distributed NoSQL (non-relational database) database module. As the result of the fact that one single relational database is insufficient to tackle big data and the high frequency reading and writing when processing massive amount of public opinion data, the system employs distributed MongoDB as a distributed NoSQL database.

d) Public Opinion Service module. The module provides users with public opinion reports and solutions by analysis and statistical data.

The basic framework of the IPOMAS is shown in Figure 1:

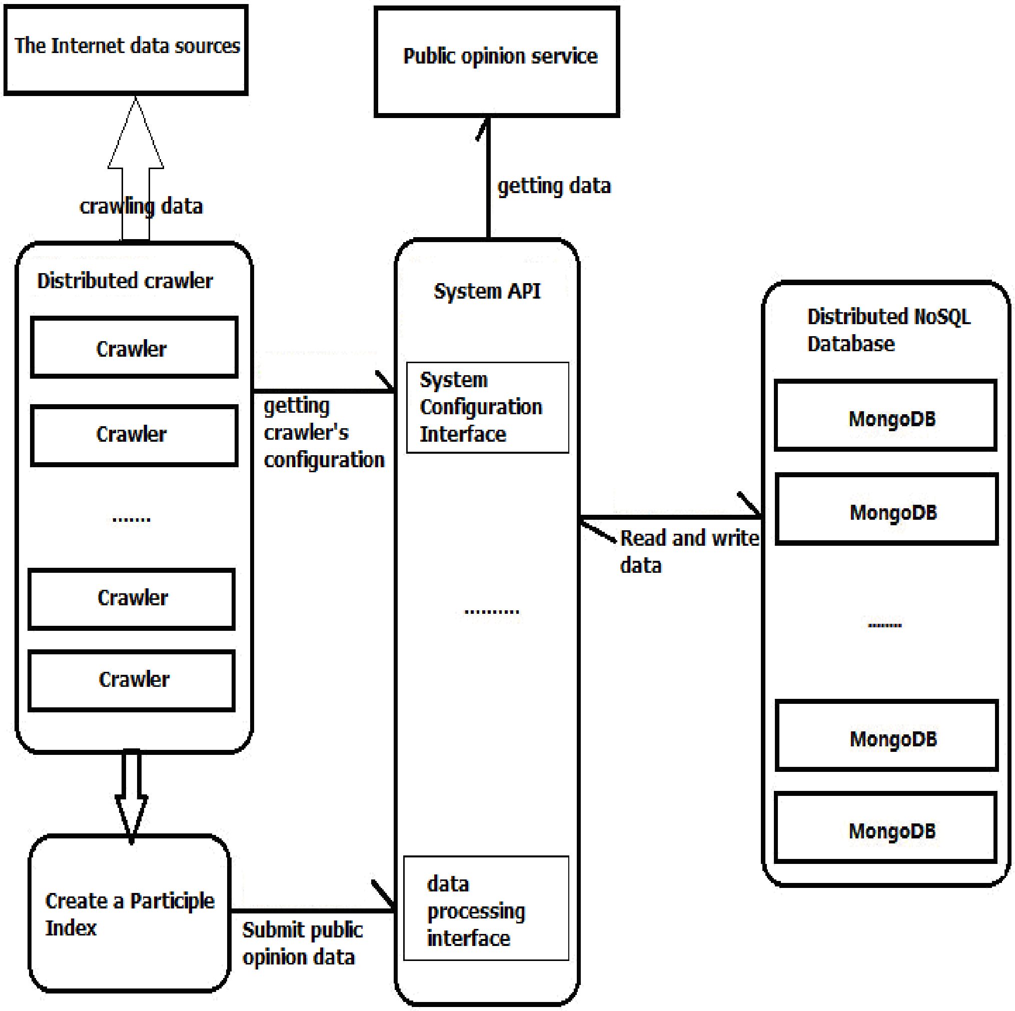

The Internet data sources
Public opinion service
crawling data
getting data
Distributed crawler
Crawler
Crawler
.......
Crawler
Crawler
System API
System Configuration Interface
getting crawler's configuration
..........
data processing interface
Distributed NoSQL Database
MongoDB
MongoDB
........
MongoDB
MongoDB
Read and write data
Create a Participle Index
Submit public opinion data

**Figure 1. The Basic Framework of the System Functions**

### 2.2. Database Design

The system uses non-relational database (NoSQL) and the collection of data is stored in JSON format. There are six main tables designed for storing monitoring library, public opinion templates, public opinion data, crawler status info, modules configuration info and system user data.

- Public Opinion Templates table stores public opinion templates data, including public opinion categories and public opinion template array etc.
- Monitoring Library table stores self-defined monitoring keywords that serve as system triggers for collecting and analyzing data.
- Public Opinion Data table stores data about the collected public opinion data including the data source, the keyword triggers, public opinion template arrays, crawling timestamp and the content of public opinion, etc.
- Crawler Status Info table stores crawler status logs, including the crawler’s name, IP address, status information, and communication timestamp.
- Modules Configuration Info table stores the configuration information for each module, including the name of configuration data.
- System User Data table stores the user data in the system, including user types, user permissions, account, password, last login timestamp, and user authorization token.

Among the six database tables, the Public Opinion data table is the most important. Table 1 indicates the detail of the public opinion data table.

**Table 1. The Table of Public Opinion Data**

| Field | Type | Descriptio**ns** |
|---|---|---|
| _id | ObjectID | Public opinion information ID |
| MD5 | String | Digital fingerprint |
| documentType | String | Data sources |
| triggerKeyWord | Array | The trigger of monitoring words |
| Category | Array | The categories of public opinion |
| createDate | Double | The creation timestamp of data |
| analyzerText | String | The content of public opinion |
| relationInfo | Document | The associated data of public opinion |

To prevent data repeating, the MD5 algorithm is applied to determine whether the data is duplicated. The algorithm is described as MD5 (data source + publisher’s name + the content of public opinion) [10]. The “relationInfo” field is used to store the associated data of public opinion, such as publisher’s name, web page address of data source, timestamp, *etc*. In order to achieve the data format generality, different types of public opinion data use different ‘relationInfo’ table structure.

## 3. System Design and Implemention

The system is based on distributed design complimented by the Java script-based crawlers and HTTPClient for Internet public opinion collection as well as IK Analyzer for analysis and Lucene for Chinese word segmentation and indexing. It passes the indexed outcome to the service bus developed by using Node.js which distinguishes data classification and processes data de-duplication. The final report and public opinion in the form of raw data are fed back to the IPOMAS users, allowing the users to understand the development trend of public opinion.

A detailed description about such key modules as the integrated management system, public opinion information pretreatment and information analysis is presented as follows.

### 3.1. Design and Implementation of Distributed Data Crawlers

The system is developed by using Java language for crawlers that operate in modules. The system also integrates Chinese word segmentation and Lucene full-text indexing and searches.

To make the crawler scalable and easily maintainable, the IPOMAS is designed to be web-crawler modular operated with some core tasks (such as Chinese word segmentation, indexing, service bus access) packaged. Developers can easily add new crawler modules in this system. IPOMAS requires a developer to develop or implement class extension StandedSpider ISpider interface and rewrite the run method if the developer wants crawlers to execute the developed modules. The approach is the main access to the modular functions.

If ISpider is used as the interface for developing modules, the approach of rewriting startAnalyzer for submitting PO data should be taken into consideration. The approach can apply similar techniques as PredatorHttpClient (packaging HTTP GET Request and POST requests) for submitting PO data to the service bus.

Crawler modular codes are shown below.

```
  public class BaiduWeiboSpider extends StandedSpider {
      private BaiduWeibo spider;// Data crawler （The object needs to be implemented by
the module developer to address specific needs.）

      public BaiduWeiboSpider() {
              this("./SpiderIndexCache/baiduWeibo");// Index cache storage location
      }

      public BaiduWeiboSpider(String indexFile) {
              super(indexFile);
              spider = new BaiduWeibo();
      }

      /***
       * Modular Main Access
       */
      @Override
      public void run() throws LuceneCoreException, DataAnalyzerException,
                      SpiderException {
              /***
               * TODO Call Crawlers Crawling Data  （Crawling process is to be
developed by the module developer.）
```

```
                 */
                // start public opinion analysis and submit data
                this.startAnalyzer();
        }
    }
```

The above example shows that a standard crawler consists of two components, namely run and startAnalyzer. Since StandedSpider has implemented StarAnalyzer, there is no need for the system users to rewrite it. However, at the end of the run approach, startAnalyzer must be called again, or the system would fail to execute PO analysis. What's more, the modular structure should include the passing of the parameter that serves as the buffering path of the indexing directory. If non reference structure is required to be implemented, the above code can be referenced as the designated default parameter.

Before the crawlers are uploaded to the module, putSpiderModule should be called in the module manager. The first parameter of the above method is module objects that must be the sub category of ISpider interface. The second parameter is the name of the module that corresponds with the module name in the configuration file.

For the crawling module to run smoothly, the configuration file (SpiderSwitch.properties) as well as the crawler configuration file (sys.properties) for the module loader should be modified.

SpiderSwith.properties configuration file has two options, namely Key and Value. Key is the name of the module and Value is subdivided into two types: Close Value and Open Value. If the module is Open, the loader will execute the module whereas if the module is closed, the loader will terminate the module.

There are three configuration files in the SpiderSwitch.properties:

a) apiURL is the API access address for the server bus.
b) spiderName is the name of crawler.
c) spiderIP is the IP address for the crawler.

An example of the configuration file is given as follows:

```
apiURL=http://10.66.3.234:3000/api/
spiderName=DeveloperSpider
spiderIp=10.66.3.234
```

Once the development and configuration of the crawler modules are completed, the crawlers can run as normal.

### 3.2. Design and Implementation of System Service Bus

In B/S structure, the Public Opinion Server Bus uses MVC [11] architecture and Express Framework and Node.js language to enable crawler and server interaction and data processing.

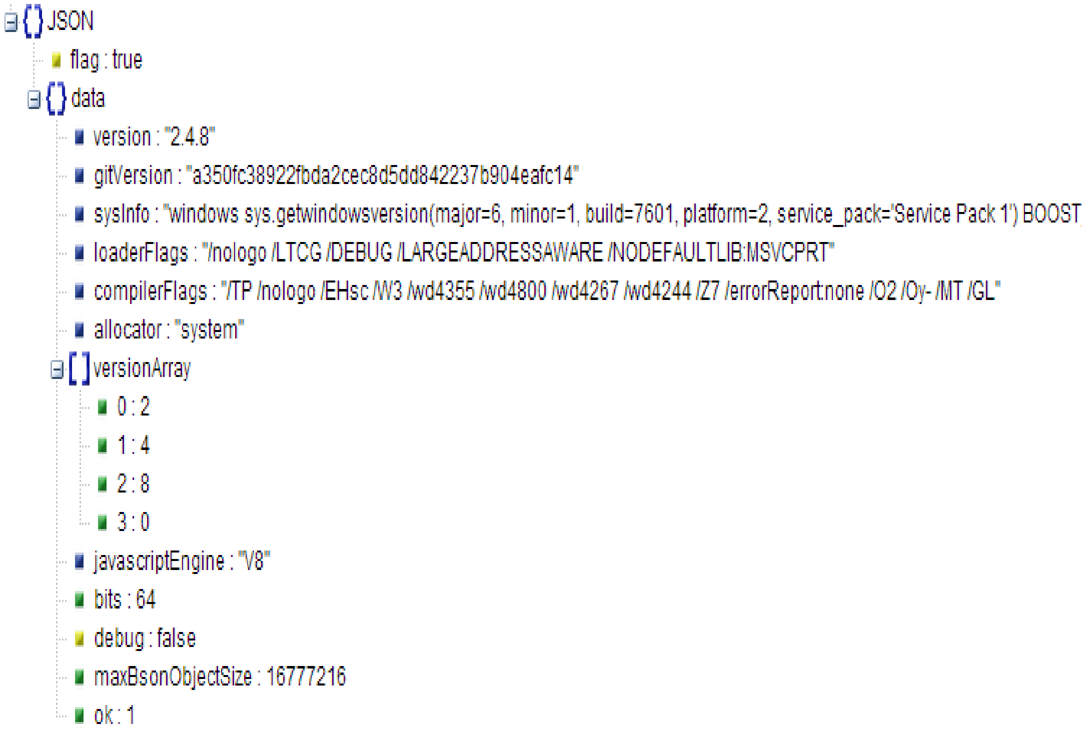
In order to facilitate future extensions of the system and subsystem access, IPOMAS uses the popular JSON format API for interaction. Data security is achieved through API certificate referencing OAuth certificate. The system users must be authenticated before using API. After certification by the system a Token key is assigned for the user, which is the permission for using all APIs in IPOMAS.

Example for system user security authentication is given below:

```
# Use Get to Access
/api/user/login?username=User Name &password=Password MD5

# If incorrect password,  the system returns JSON data：
{
        "flag": false
}
# if correct password,  the system will return Token key,  shown as follows：
{
        "flag": true,
"token": "e7011a15297bd4c8ff02ad4f88181d8a"
}
```

In example of Database Information Acquisition API with the assumption that the system user has got the Token value, the system will return JSON data as shown in Figure 2, the acquired Token value must be passed as a parameter when calling the API.



```
⊟ { } JSON
   flag : true
  ⊟ { } data
     version : "2.4.8"
     gitVersion : "a350fc38922fbda2cec8d5dd842237b904eafc14"
     sysInfo : "windows sys.getwindowsversion(major=6, minor=1, build=7601, platform=2, service_pack='Service Pack 1') BOOST
     loaderFlags : "/nologo /LTCG /DEBUG /LARGEADDRESSAWARE /NODEFAULTLIB:MSVCPRT"
     compilerFlags : "/TP /nologo /EHsc /W3 /wd4355 /wd4800 /wd4267 /wd4244 /Z7 /errorReport:none /O2 /Oy- /MT /GL"
     allocator : "system"
    ⊟ [ ] versionArray
       0 : 2
       1 : 4
       2 : 8
       3 : 0
     javascriptEngine : "V8"
     bits : 64
     debug : false
     maxBsonObjectSize : 16777216
     ok : 1
```

**Figure 2. The Returned Data from the Database Information Acquisition API**

### 3.3. System Backend Design and Implementation of Public Opinion Service

The system backend of Public Opinion Service is one of the most important components of the system. It is the back end service with UI for all system users. The backend service system is designed for provisioning reporting services for the system users to understand the current trends of the public opinion monitored and predict their future development and track the information sources of the public opinions. The backend system also provides flexibility for the IPOMAS users to configure crawler information, modify or update morning thesaurus and public opinion templates and export public opinion reports and data.

Using B/S architecture and based on API development for public opinion service bus, the system backend adapts Ajax and interfaces with server bus API. The front end uses Bootstrap UI framework and HTML5, providing dynamic dashboards by using chart.js.

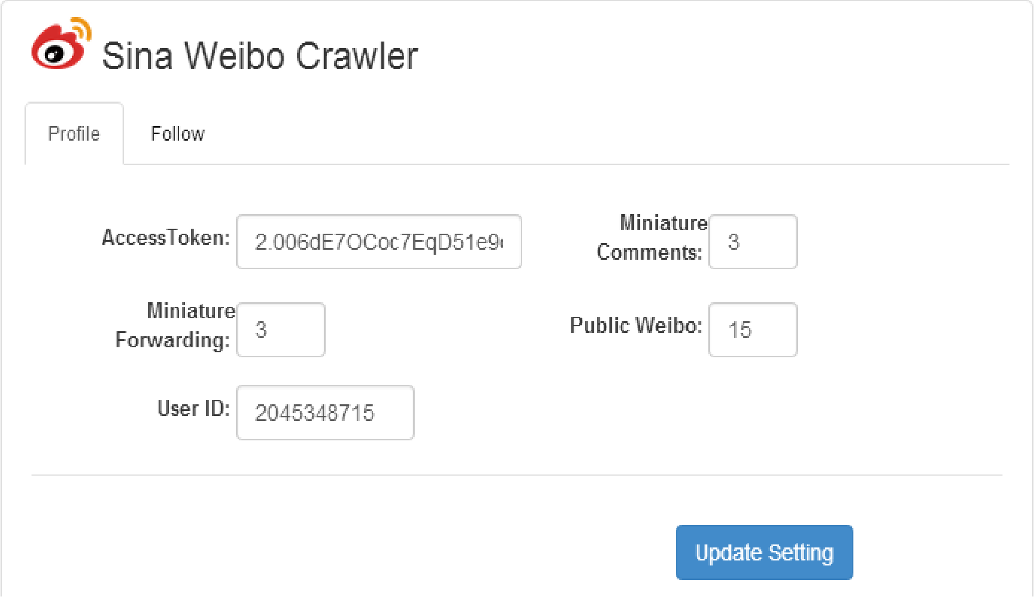
(1) Crawler Management

Armed with modular-based distributed crawling techniques, IPOMAS uses a Crawler Management Module (CMM) for efficient crawler management and configuration. CMM consists of the following four main functions: data maintenance, crawler setting, Sina Weibo Configuration and Comment crawlers.

The Crawler Status Info is captured by using Ajax. Once the crawler's status data is obtained, the system writes the status info to the current web page in HTML format. Therefore, a real time status of the crawlers is displayed on the system user's screen without requiring refreshing to facilitate easy management of distributed crawlers.

Constrained by the Sina Weibo interface, the crawlers can only use the API provided by Sina Weibo to access Sina Weibo. IPOMAS can configure the Sina Weibo crawlers and set monitoring focus on a microblogging object in Weibo. Once the monitoring object is defined, the crawlers will check the objects every so often on any new updates and collect the new updates to the system if updates occur.

The configuration UI of Sina Weibo Crawlers is shown in Figure 3.



AccessToken: 2.006dE7OCoc7EqD51e9

Miniature Comments: 3

Miniature Forwarding: 3

Public Weibo: 15

User ID: 2045348715

**Figure 3. Sina Weibo Crawler Configuration UI**

(2) Monitoring Setting

There are four main functional modules in the monitoring settings, namely: public opinion templates, monitoring thesaurus, sentimental tendencies, and Internet hot-word detection.

IPOMAS uses template based public opinion identification algorithm which not only quickly determines specific information on the subject of public opinion via pre-entry templates but also collects, processes, and classifies public opinions to allow user-friendly data analysis and monitoring.

The feature of Public Opinion Template is one of the most important algorithms of the system implementation. The algorithm is based on Lucene index statement in support of conditional statements and determines the public opinion data classification in accordance with the preset templates (conditional statement). More efficiency and accuracy for determining the public opinion classification can be achieved by using preset templates than sentimental tendency library. The algorithm code is as follows.

```
    /***
        * Matching character conversion
        *
        * Keyword matching operator allows AND OR NOT conditional query,  Improve
the accuracy of public opinion analysis
        *
        * @param keywords
        *          Keywords
        * @return String Keyword matching character conversion
        */
       public static String matchingChar(String keywords) {
              String matchedStr = keywords.trim();
              String andRegex = "\\b AND \\b";
              String orRegex = "\\b OR \\b";
              String notRegex = "\\b NOT \\b";
              matchedStr = matchedStr.replaceAll(andRegex, "* AND ")
                            .replaceAll(orRegex, "* OR ").replaceAll(notRegex, "* NOT
");
              return matchedStr;
       }
```

The algorithm, using the asterisk symbol for template matching, not only improves the efficiency of Lucene indexing but also rapidly increases public opinion data classification.

However, due to the fact that it is insufficient to ascertain the precision of public opinion data classification by single use of asterisk character, IPOMAS uses the secondary indexing algorithm to classify collected public opinion data. In other words, the public opinion classification indexing precision is ensured via matching indexing twice. The algorithm process flowchart is shown in Figure 4.

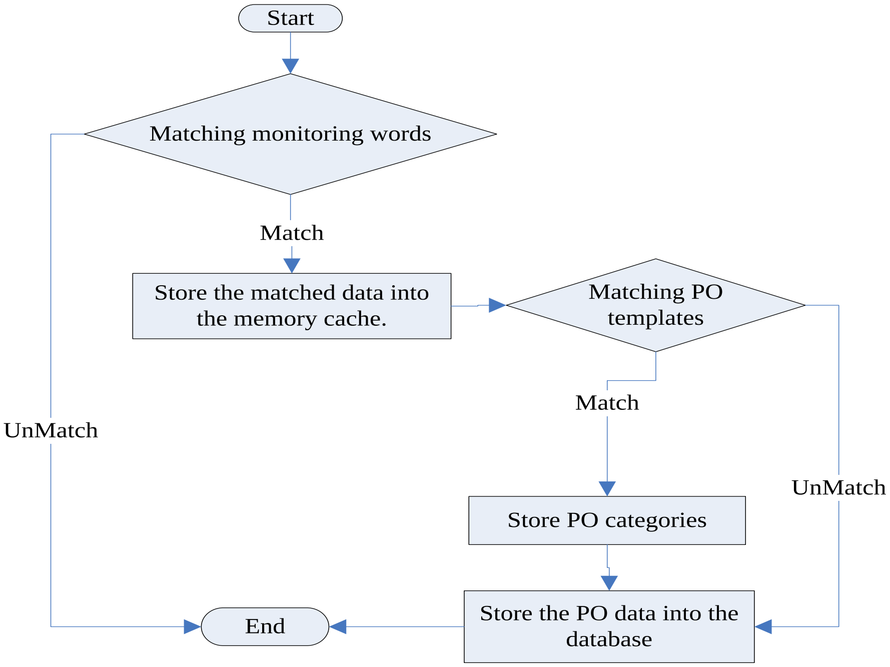

Start
Matching monitoring words
Match
Store the matched data into the memory cache.
Matching PO templates
Match
UnMatch
Store PO categories
UnMatch
Store the PO data into the database
End

**Figure 4. The Secondary Matching Algorithm Flowchart**

The recognition accuracy depends on the level of details of the public opinion template. The better the template is defined, the more detailed and accurate the classification on the public opinions will be.

(3) Public Opinion Reporting

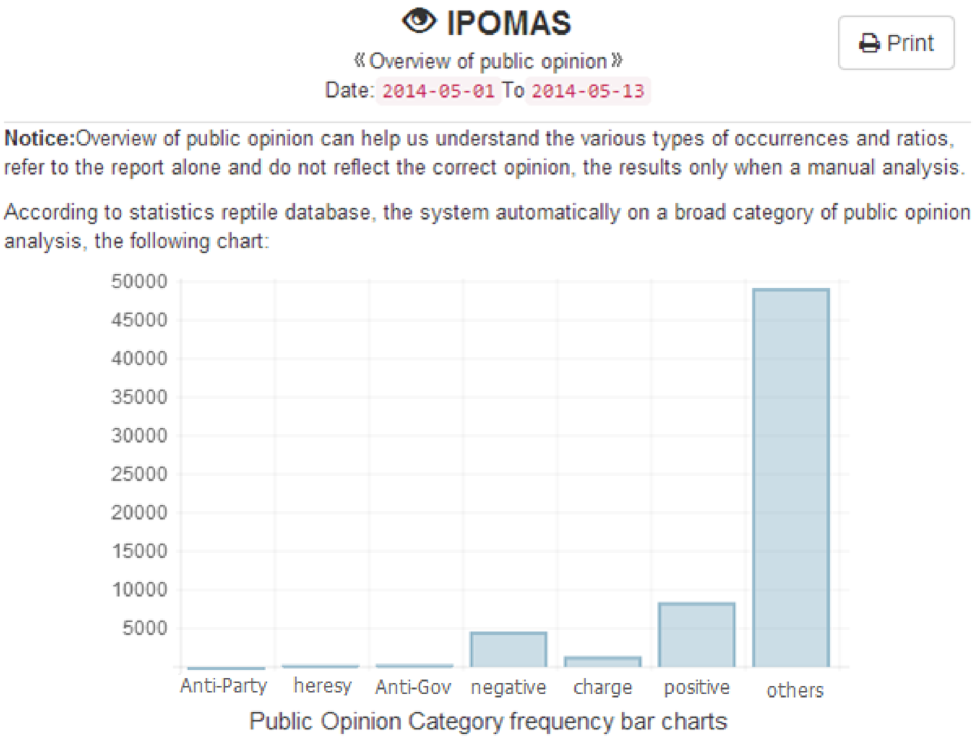
The Public Opinion Reporting component comprises of public opinion dashboard and public opinion tracking reporting. The feature is one of the most important functions of the system. It is through these two functions that the IPOMAS users acquire the understandings of public opinion trends as well as undertaking monitoring and analysis of public opinions on the Internet.

The public opinion diagram and chart on the dashboard is mainly based on Chart.js development techniques. The component uses HTML5 Canvas [12] rendering with animated charting and diagramming capabilities. IPOMAS can put the chart JSON data to Chart.js without worrying about how the chart and diagram dashboards are generated.

The public opinion reporting capability is one of the core services of the system which enables IPOMAS users with the flexibility of setting monitoring keywords and the public opinion template information and obtaining a quick understanding of the trends of public opinion and its development on the Internet through chart and text dash-boarding.

Public opinion is one of the core services reporting capabilities of the system, monitoring the words and opinion by setting the template information. Users can quickly understand the network developments and trends of public opinion through a chart and text.

Figure 5 is an example of the system-generated Public Opinion Category and Frequency Bar Chart.



IPOMAS

《Overview of public opinion》

Date: 2014-05-01 To 2014-05-13

**Notice:**Overview of public opinion can help us understand the various types of occurrences and ratios, refer to the report alone and do not reflect the correct opinion, the results only when a manual analysis.

According to statistics reptile database, the system automatically on a broad category of public opinion analysis, the following chart:

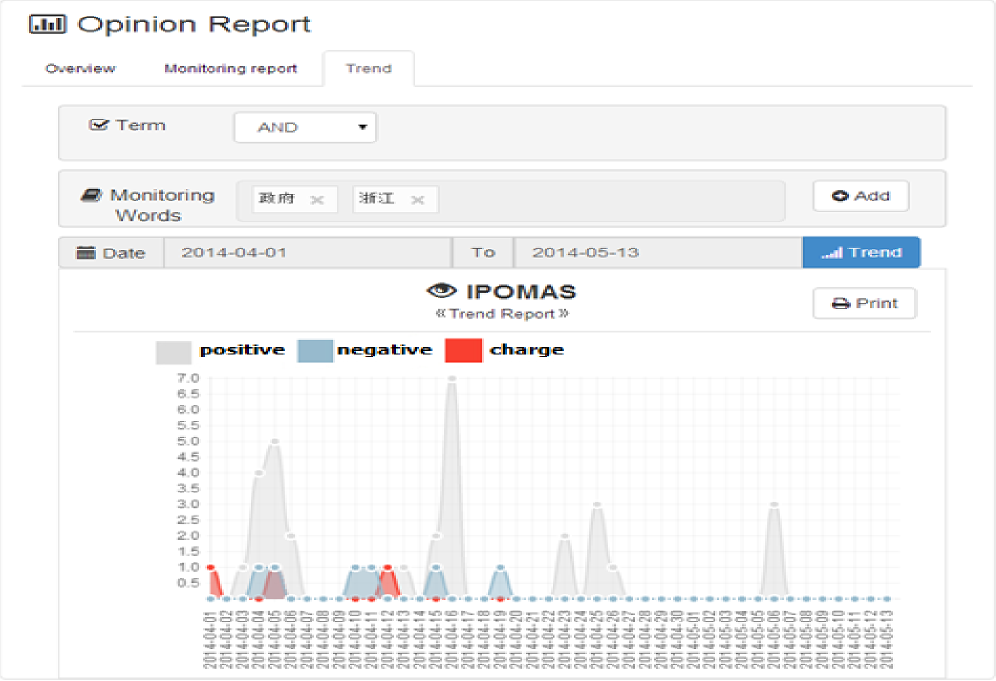
**Figure 5. Public Opinion Category and Frequency Bar Chart**

Public opinion information packet using the Map-Reduce MongoDB grouping function, which is able to instantly process and group tens of thousands of public opinion data entries. Currently there are tens of thousands of entries of public opinion data in the database. Using the Map-Reduce techniques can sort the data in a flash, greatly speeding up the data processing capability of the system.

Map-Reduce implementation code is show as below:

```
function Map() {
  emit(
          this.city,                                    // how to group
          {count: 1, age: this.age} // associated data point (document)
  );
}
```

Based on monitoring the condition, the IPOMAS users can also query the system regarding the public opinion trending a period of time in either the single and mixed query mode. An example of public opinion trending analysis is shown in Figure 6.



Opinion Report

Term: AND

Monitoring Words: 政府 浙江

Date: 2014-04-01 To 2014-05-13

IPOMAS
«Trend Report»

**Figure 6. Public Opinion Trending Analysis**

In order to facilitate forensic evidence, every single bit of the information associated with public opinion crawled by IPOMAS will be stored in the database, including the information about the publisher account, the original web address, publishing time, IP address, *etc*. The system will classify all data sources in accordance with where data comes from. The associated information presented to the system users varies from one data source to another.

## 4. Conclusion

This paper presents the concept design and the implementation of a scalable and low-maintenance Internet public opinion monitoring and analysis system for monitoring and analyzing public information in cyberspace with the functionality of providing users with public opinion monitoring reports for decision-making on various public events.

However, the public opinion data has been exponentially increasing on daily basis. Statistically people and brands on Twitter send more than 340 million tweets a day at the moment. IPOMAS is currently focusing on the analysis on a portion of data associated with the keyword triggers. In future research work, the further development of IPOMAS will exceed this limitation and expand the scope of monitoring and analysis so that the system can cope with the ever-changing public opinion information and further enhance the value of the system.

## Acknowledgement

This work is supported by the 2013 Construction Research Project of Zhejiang Provence, China (No. 18).

## References

[1] H. Gao, M. Zhou and Y. Fu, “Analysis of Netizen's Affective Tendency on Public Opinion”, Journal of Convergence Information Technology, vol. 5, no. 6, **(2010)**.

[2] M. J. Xin, H. X. Wu and Z. H. Niu, “A Quick Emergency Response Model for Micro-blog Public Opinion Crisis Based on Text Sentiment Intensity”, Journal of Software, vol. 6, no. 7, **(2012)**.

[3] J. F. Wang, X. Jia and L. B. Zhang, “Identifying and Evaluating the Internet Opinion Leader Community Through k-clique Clustering”, Journal of Computers, vol. 9, no. 8, **(2013)**.

[4] F. Cao, Z. J. Zhang, Y. C. Jing and X. L. Guan, “A model of ecological monitoring and response system for Internet public opinion”, International Journal of Multimedia and Ubiquitous Engineering, vol. 5, no. 9, **(2014)**.

[5] S. Tilkov and S. Vinoski. “Node, is: Using JavaScript to Build High-Performance Network Programs”, IEEE Internet Computing, vol. 14, no. 6, **(2010)**.

[6] Q. Wu, H. X. Xia, G. H. Zhao, and C. Y. Liu, “Application and Improvement of Lucene-based Search Engine”, Journal of Wuhan University of Technology, vol. 7, **(2008)**.

[7] M. Thelwall, “A web crawler design for data mining”, Journal of Information Science, vol. 27, no. 5, **(2001)**.

[8] Z. Y. Luo and R. Song, “Disambiguation in a modern Chinese general-purpose word segmentation system”, Computer Research and Development, vol. 43, no. 6, **(2006)**.

[9] H. C. Chang and H. S. U. C. Chieh, “Using topic keyword clusters for automatic document clustering”, IEICE Transactions on Information and Systems, vol. 88, no. 8, **(2005)**.

[10] S. W. Chen and J. C. Hui, “Research on the multi-message modification techniques on MD5”, Journal on Communications, vol. 8, **(2009)**.

[11] S. Z. Fang, “Research on Framework Developing Technology Based on MVC”, Journal of Wuhan Institute of Shipbuilding Technology, vol. 1, **(2009)**.

[12] D. X. Yang, Y. M. Yun and S. H. Cha, “Virtual reality contents based on X3D and HTML5 Canvas”, International Journal of Advanced Media and Communication, vol. 5, no. 2, **(2014)**.

## Authors

### Guanlin Chen

Born in 1978, Ph.D., associate professor, chenguanlin@zucc.edu.cn. His main research interests include computer networks, E-government and information security.