Research and Application Development for SMS Extraction and Analysis System

To cite this article: Jiawei Mao et al 2018 J. Phys.: Conf. Ser. 1069 012068

View the article online for updates and enhancements.

Related content
- Outlier Detection on Hotspots Data in Riau Province using DBSCAN Algorithm
  Pristi Sukmasetya and Imas S. Sitanggang
- Hotspots for carbon offsets in agricultural landscapes: Identifying threats, opportunities and policy responses to widespread reforestation
  Neville Crossman, P Butler, D Summers et al.
- Pressure and magnetic field effect on transport gap of black SmS
  K Imura, K Matsubayashi, H S Suzuki et al.
Research and Application Development for SMS Extraction and Analysis System

Jiawei Mao, Hui Xu¹ and Yawei Hu
School of Computer Sci. & Tech, Nantong University, Nantong, 226019, China
Email: xu.h@ntu.edu.cn
¹Corresponding author

Abstract. Mobile phone SMS play an indispensable role in people's work, life and friendship, and accumulated text messages contain rich information. In this paper, we study the method of visually identifying the amount of texts in different places on the map and the method of visualizing the amount of text messages of different individuals in a certain period of time using a bar chart. The system research and development is based on the technologies of SMS extraction, information filtering and storage. On the one hand, the system can enable users to query short messages according to their territories and time periods. On the other hand, the visualization of this information intuitively reflects the contact heat of different regions, and the heat of contact of different people in a mobile phone contact with the outside world, and so indirectly can obtain far-reaching, useful information such as business hotspots, hotspots for dating, change of contact's ownership, etc.

1. Introduction
The intelligence of smart phones has greatly extended their functionality [1]. Mobile phones are not merely electronic devices to contact with others; they have become people’s indispensable assistants in work, lives, and interpersonal relationships. Whoever has the capacity of using a cell phone will deem it a necessity in daily life. The “intelligence” of mobile phones applied in many software with the same real-time calling functions as mobile phones, such as IP phones, WeChat voice, QQ voice, covering the original functions of mobile phones; What’s more, the real-time text transmission of WeChat or QQ, plays the same role as SMS, and it will cost less than SMS. However, WeChat, QQ, etc. all rely on special application software. When these softwares are not started, the information channel cannot be established. On the other hand, the message will be unblocked as long as the phone is on. Therefore, SMS plays an irreplaceable role due to its timely, convenient and real-name mobile phone system.

Message socialization is an efficient social method that transmits information accurately, receives timely, and responds quickly. It has been an important daily routine for many organizations to send notifications via SMS [2]. Today, the SMS function of mobile phones is not just a text transmission tool in a general sense, but an irreplaceable tool for authentication, notification of important issues, timely notification, etc. The extensive use of text messages have made it possible for text messages to contain more and more information, and the precipitation of complicated text messages is a treasure house. This paper investigates (1) Extraction of SMS and the methods to visualize results of the analysis of SMS possession. (2) Classic statistics and visualization methods sorted by individuals at specific time periods. The features of this software are:

- Data visualization, presenting analysis results to users through charts;
- Automatically filter invalid information to ensure the accuracy of the analysis results;
The content of the analysis is more comprehensive by combining geographical information with time information;
· Higher security, critical data does not flow into the Internet;
· Easier to operate, less user restrictions.

2. System Analysis and Design

2.1. Data Analysis
System processing object are: contacts and messages. The properties of contacts are: phone number, name, attribution place, account etc. The properties of messages are date, content etc. The relationship between the two is one to many, which means, a contact related text message is multi-conditional. The E-Rdiagram is shown in Figure 1.

![E-R diagram](image)

**Figure 1.** E-R diagram.

The contact list table and the short message table are established according to the above analysis. The table definition is shown in table 1 and table 2.

| Field   | Type  | Key |
|---------|-------|-----|
| Phone   | String| Yes |
| Name    | String| No  |
| Place   | String| No  |
| Count   | Integer| No |

**Table 1.** Contacts.

| Field   | Type  | Key |
|---------|-------|-----|
| Phone   | String| Yes |
| Detail  | String| No  |

2.2. Analysis and Visualization of SMS Possession
Definition: Possession, in this paper, means to the area where the phone number sending the message belongs to.

Getting possession of SMS, statistics the account of messages classified by possession, and marked by different colors in the map. Users can get detailed information of messages by clicking certain areas. The processing flow is shown in figure 2.
Figure 2. Processing flow of SMS possession.

**Step 1: Traverse the database of inbox.** Access the phone’s internal text message via URI "content://sms/inbox" and address book via URI "Contacts Contract. Common Data Kinds. Phone. CONTENT_URI[3]".

**Step 2: Filter and save phone numbers.** Filter out invalid mobile number, save new mobile number.

1) Use regular expression [4] to extract key information and determine whether the SMS number exists in the address book. If it exists, then filter and goes to 4); else, to 2).

2) Determine if the length of the number is 14 digits (including the international number "+86") or 11 digits (does not contain the international number). If not, then filter it and goes to 4); else to 3).

3) The mobile phone number is sent to the server through the API for judgment. If the returned reason is 0, it is not filtered; otherwise, it is filtered.

4) The filtered short message is passed through the network API [5] to obtain the Json data related to the cell phone number attribution and resolve the attribution.

5) Save the processed data to the database through object-oriented database operation framework Litepal.

**Step 3: Data statistics.** The system provides the number of SMS statisticied according to different time periods and different territories. The system selects all the short messages from the selected time to the system time through the time selected by the user. At this time, the user can click to query the number of short messages of a certain place in the time period.

**Step 4: Data visualization display.** The data is visualized and displayed on the map in different colours based on the number of different types of SMS. Areas with a lot of text messages are filled with darker colours. In this way, you can see at a glance which regions have more text messages.

1) **Draw map.** Use custom maps to show analysis results. Custom maps are mainly obtained by parsing map point set information in SVG (Scalable Vector Graphics) format [6]. This system mainly uses the d attribute of the most complex <path> element in the basic graphics of SVG to complete the drawing work. There are two types of command: straight line command and curve command. Line commands such as "M" indicate moving brush; "L" connects two points to draw a straight line, "H" draws parallel lines, "V" draws vertical lines.; "Z" closes the path; curve commands such as "C" indicate three times The Searle curve, "Q" indicates a quadratic Bezier curve, and "S" and "T" indicate abbreviated Bezier curves. For example, "Beijing */" M421.139, 189.75L420.782, 186.894L419.95, 184.989L425.045, 182.863" stores information on some of the shapes of Beijing in the map.

2) **Filling areas with colours.** Predefined a set of colours, ordered by the descending order of the number of text messages.

3) **Frame.** We need to draw a border after completing the vector content. Observing the Middle Ages map, Heilongjiang is the easternmost and northernmost province in China, Xinjiang is the westernmost provinces in China, and Hainan is the southernmost province in China. Therefore, the map boundary is (0, 0)->(Heilongjiang.right,0),(0,Hainan.bottom)->(Heilongjiang.right.bottom), the map width is “Heilongjiang. Right”, and the map height is “Hainan. Bottom”. Besides (0, 0) represents the position of the brush in the vector diagram, located in the upper left corner of the map.

**Step 5: Query the specific SMS.** Users click on the domain to display the detailed text message number and text message list in the chart.

1) **Set the event as the reaction to the click.** By clicking, on the one hand, the area is selected in the background; the database is searched by the area information, and the analysis result of the number of short messages in the corresponding area is obtained; on the other hand, the result display form is popped up.

2) **Set map click events.** The click reacts to zoom maps, adjust the size and position of screen maps and tables.

3) **Set the click event (click on tab named datail).** A list of text message is got by clicking.
2.3. Classification Statistics and Visualization based on Time and Individual

The user sets the time query conditions, the system counts the number of messages according to the sender, and the result is displayed as a bar graph. Click on a bar to view detailed text message records. The processing flow is shown in figure 3.

![Figure 3. Processing flow of SMS analyzed by time period and individual.](image)

2.3.1 Time Selector. The time selector is a dialog based custom control. According to system requirements that time selector need to be accurate to the day. So the design needs to be able to allow the user to freely choose the year, month and date. The custom method mainly includes the title content, time format, interface style, animation effect, and monitoring content. The tool class WheelTime, which is used to display the scroll time, contains three private variables "WV_YEAR", "WV_MONTH", and "WV_DAY", which correspond to the year, month, and day items that can be scrolled. For the reason that the number of days in different months is different, in order to achieve the linkage effect, dynamic update data needs to be implemented for whether it is a leap year and a specific number of days of each month. So each time the scroll completed, data updated. For example, the "year" changes can affect the "month" and "day", changes to the "month" can affect the "day". The time selector achieve the most accurate display by dynamic refreshing when clicked. In addition to this, it is necessary to set the "Year", "Month", and "Day" prompt words after each scroll entry, and achieve the most appropriate visual effect by setting the horizontal distance between the prompt words and the scroll controls.

2.3.2 SMS Statistics. The statistics of the individual text messages are collected from the local database according to the time period set by the time selector.

2.3.3 Histogram Plotting and Data Filling. 1) Determine the height and width of the graphics area. In order to have a more comfortable visual experience, the x-axis width accounts for 70%-85% of the container width, and the y-axis accounts for 80%-90% of the container height.

2) Determine the coordinate scale. The minimum scale unit of the y-axis is 2, and the scale of the x-axis is described dynamically using the abscissa of the input data.

3) Plot the bar. Assuming that the incoming data value is value, the origin coordinate is yPoint, and the y-axis unit length is y-Scale, the height of the cylindrical height corresponding to the y-axis scale is obtained through height = yPoint-value/2*y.

4) Data retrieval. Retrieving the number of different individual messages according to the time period statistics from the database. Converting the quantity into an array of string types to be stored as the values, converting the contact names into arrays of string type as x-Label, and passing the values and x-Label through the initialization method. Passed as a parameter to achieve histogram drawing.

3. Function Display

3.1. Possession Visual Display of SMS Possession

Draw a map at system initialization. The first displayed content is the statistical result of historical data. Users can also use the time selector to display statistical results for a certain period of time. Click on an area and the number of specific text messages in the area can be seen in the table, as shown in figure 4. Click on the details, users can see a list of text messages from this area, including the phone number and text message summary; click on one of the items to display detailed text message content. As shown in figure 5.
3.2. SMS Histogram and SMS Inquiry
The initial session is shown in the figure 6. Set the time period through the time selector. The system presses the sender to count the number of messages, and the result is displayed as shown in the figure.
7. Users can click on its column to check a person's text message and it’s similar to the display interface.

![Time Selector](image)

**Figure 6.** Time Selector.

![Histogram](image)

**Figure 7.** Histogram.

4. **Conclusion**

This paper presents a visual representation of the classified statistics of SMS by region and time. Accumulated text messages contain a lot of useful information. For example, the SMS analysis technology given in this paper can tell us: 1) Hotspot business. As a salesperson, the business spread allover the country, it means more business contacts with the region in which there are more text messages. Indicating more business and longer cooperation with a person who sends more text messages and has more time. 2) Change of place of residence or change of place of work. People who use the mobile phone to talk more often to choose to use the package. The package has distinctive geographical attributes. If a contact shows statistical results for a considerable period of time before a certain point in time or after, the SMS message is clearly concentrated in two different areas, indicating that the person's place of residence or workplace has changed. There is no much work is done on extracting text messages. This paper is just served as a trigger.

5. **Reference**

[1] Baidu Encyclopedia. https://baike.baidu.com/item/%E6%99%BA%E8%83%BD%E6%89%8B%E6%9C%BA/94396?fr=Aladdin.2018.03.05
[2] Han, Yumin, Yang Tengfen, and Liu Anzhan. Design and implementation of schedule management software based on parsing messages content [J]. *Journal of Microcomputer Application*, 34 (5), 2015, pp.92-94. (in Chinese)

[3] Fra~~kaka's Blog. https://www.cnblogs.com/kakafra/archive/2012/10/06/2713327.html

[4] Baidu Encyclopedia. https://baike.baidu.com/item/%E6%AD%A3%E5%88%99%E8%A1%A8%E8%BE%BE%E5%BC%8F/1700215?fr=Aladdin

[5] http://apis.juhe.cn/mobile/get?phone=17805058405&key=22a6ba14995ce26dd0002216be51dabb.

[6] Zhang Xing-xu. Deeply grasp the Bezier curve instruction of path of SVG path. http://www.zhangxinixu.com/wordpress/2014/06/deep-understand-svg-path-bezier-curves-command/