# *Simple Mining Method of Rich Text Data from User Feedback and Its Application*

Cai Huali*, Wu Fang, Duan Qi, Kang Jian, Jiang Yawei
China National Institute of Standardization, Beijing, China
(chl2081@126.com)

***Abstract*—At present, many large-scale consumer goods or food suppliers would receive a lot of feedback data from consumers, which is vital to the improvement of corporate management. Nevertheless, these data fail to be fully utilized or labeled manually piece by piece due to employees' poor skills or officers' lack of awareness in most of these enterprises. This paper investigates a simple method of handling such data. First, several basic databases are built, including index system, industry topic lexicon and emotion lexicon. Then, the polarity judgment method is studied. Finally, criticism rate, appreciation rate and suggestion rate are studied and determined. All these rates have gained application in a large-scale enterprise of China and received good effect.**

**Keywords—*Rich Text, Data Mining, Complaint Rate Introduction***

## I. INTRODUCTION

Many enterprises would receive a lot of feedback comments from customers after selling their products or services. Most of the comments are offered in the format of non-structural text. Among these comments, some shows appreciation or recognition, while some indicates criticism or complaint and others express both of the above. Such comments represent the best information source for enterprises' information closed-loop and improvement of product (service) quality. However, with the continuous increase in sales, enterprises would be surrounded by massive comments data. Under the current technical background, it seems both time- and effort-consuming to identify such information piece by piece manually. Finally, such feedback data fail to be well utilized. This paper aims to solve difficulties of users' feedback data in being rapidly handled by studying the method for mining such data[1][2].

## II. RICH TEXT INFORMATION MINING METHOD

### *A. Building of a basic database*

Before data mining, a basic database needs to be designated as system input, including index system, industry topic lexicon and emotion lexicon.

- Index system

A scientific index system was designated according to industry characters. It is suggested by the author that three classes should be classified into. The Class III index is measurable. The results of both Classes I and II indexes were calculated according to the average of third-class indexes in a reversing order.

Table 1 Index System

| Class I | Class II | Class III | Keywords |
|---|---|---|---|
| $L_1$ | $L_{11}$ | $L_{111}$ | $KW_1, KW_{2...}$ |
| | | $L_{112}$ | $KW_1, KW_{2...}$ |
| | $L_{12}$ | $L_{121}$ | $KW_1, KW_{2...}$ |
| | | $L_{122}$ | $KW_1, KW_{2...}$ |
| ..... | ..... | ..... | ..... |

- Industry topic lexicon

A three-class tagged descriptive lexicon needs to be designed according to industry characters, including synonyms of individual words. Basically, one lexicon should be designed for each industry.

Table 2 Example of TV Index System Design

| Class I | Class II | Class III | Keywords |
|---|---|---|---|
| Product use | Visual appearance | Styling | Design, elegant, royal, styling, percentage, coordinated, graceful, refined, style, simple, pretty, quality, unique, thin, ultra-thin, ultra-large, large enough, large-screen, wide-screen |
| | | Color/exterior appearance | Luxurious, fashion, bright-colored, beautiful, pattern, fresh, unconventional, taste, modern, nostalgia, low profile, lavish, classical, color, matching, aesthetic, gorgeous, color, soft |

- Emotion lexicon

In addition to the traditional emotion lexicon, emotion words of special areas need to be designed, such as "murmur" and "rat-a-tat" made by the refrigerator. Also, either positive or negative polarity should be marked.

### *B. Polarity judgment method*

Statements will be divided based on users' feedback data to obtain the keywords included in such data. The three-class index coverage of industry topic words, S will be added by 1 according to the pre-built index system. Then, users' feedback will be processed according to the pre-built users' emotion lexicon to determine the emotion from such data. Finally, the

 

polarity of tags will be determined according to the emotion: for appreciation, add 1 to the positive polarity value $S_+$; for complaint, add 1 to the negative polarity value $S_-$; and for suggestions, add 1 to the neutral value (suggested value) $S_0$.

After all divided statements are subject to the above processing, the criticism rate P, appreciation rate B and suggestion rate J will be calculated.

$$P= S_+/S\times100\% \quad (1)$$

$$B= S_-/S\times100\% \quad (2)$$

$$J= S_0/S\times100\% \quad (3)$$

The specific calculation process is shown in the table below. The above results offer the most immediate reference for enterprises' improvement.

Table 3 Polarity Judgment Process

| | |
|---|---|
| 1 | Building a keyword lexicon |
| 2 | Building a topic tag tree |
| 3 | Building a user emotion lexicon |
| 4 | Obtain users' feedback data |
| 5 | Dividing users' feedback data by punctuations |
| 6 | Screening the divided statements according to the pre-built keyword lexicon to obtain the keywords included in the statement |
| 7 | Adding 1 to T, the level of designated tag with keywords according to the pre-built topic tag tree |
| 8 | Processing the words subject to screening above according to the pre-built user emotion lexicon to determine the polarity of statement |
| 9 | Determining the polarity of tags with designated classes mentioned above according to the polarity of the statement and adding 1 to the polarity value. The said polarity includes: criticism P, appreciation B and suggestion J |
| 10 | Statement output |
| 11 | When all divided statements are subject to above processing, criticism rate, appreciation rate and suggestion rate can be calculated according to the level of each designated tag, $T_{final}$, $P_{final}$, $B_{final}$ and $J_{final}$, corresponding to users' feedback data |

## III. APPLICATION

Conclusions were reached through analysis of the data for a large-scale household appliance enterprise of China, with more than 150,000 pieces of data received. As the enterprise highlights customer complaints, the processing result mainly focuses on number of complaints and complaint rate, as shown in the table 4:

Table 4 Application Example

| Class II Topic | Class III Topic | | Refrigerator | Freezer | Special refrigerator | Medical freezer |
|---|---|---|---|---|---|---|
| Sample size | | | 151005 | 27449 | 21670 | 747 |
| Visual appearance | Styling | Number of complaints | 13 | 2 | 3 | 0 |
| | | Complaint rate | 0.01% | 0.01% | 0.01% | / |
| | Color/exterior appearance | Number of complaints | 134 | 21 | 42 | 1 |
| | | Complaint rate | 0.09% | 0.08% | 0.19% | 0.13% |
| | Fine workmanship | Number of complaints | 5504 | 687 | 859 | 24 |
| | | Complaint rate | 3.64% | 2.50% | 3.96% | 3.21% |

## Acknowledgment

This work was funded by the Dean fund project of China National Institute of Standardization under grant No. 552016Y-4667, the National Key Technology R&D Program of the Ministry of Science and Technology under grant No. 2015BAK46B02 and 2015BAK46B03-3.

## References

[1] Zhang Ziqiong,etc.. Literature review on sentiment analysis of online product news[J].Journal of Management Science in China.2010(6),13:84-96

[2] Lishi etc.. Mining features of products from Chinese customer online reviews[J]. .Journal of Management Science in China 2009(2),12:144-152

 