The Summarization of Commodity Short Comments Based on Topic Clustering

Chengcheng Yu and Ling Shu
College of Computer and Information Engineering, Shanghai Polytechnic University, No.2360, Jinhai Road, Pudong New Area, Shanghai, China.
Email: ccyu@sspu.edu.cn

Abstract. In recent years, the world economy is being transformed into a digital one. E-commerce has brought a lot of convenience people’s life, which is followed by the exponential growth of online reviews of e-commerce. These comments are very valuable for e-commerce and users. This paper proposes an effective method of summarizing short comments information based on Chinese. This method uses a series of phrases composed by noun and adjective as the final presentation. Firstly, the noun and adjective phrases in each comment were extracted, then the LDA probability model was used for content clustering, and finally the content summary of each category was extracted according the frequency of occurrence. In this paper, the real comment data of Taobao is used for experiments, and the results show that the proposed method can effectively summarize valuable information of commodities.

1. Introduction
For the past few years, the world economy is transforming towards digitization. With the vigorous development of internet technology, human society has officially entered the web2.0 era [1], and it has become a global consensus to vigorously develop the digital economy. This has been followed by a surge in online reviews of goods. For users, fragmented, incomplete, or even erroneous information makes it increasingly difficult to obtain comprehensive, objective information. This evaluation information is of high value for both e-commerce and users. Effective mining of valuable information in massive evaluation data is of great significance not only for users, but also for e-commerce and enterprises.

Based on the subject clustering method, this paper summarized the short comment information of e-commerce. The main contributions of this paper are as follows:
• The formal definition of the summary information problem of commodity comments is made, and the summary is finally displayed by a series of phrases composed by noun and adjective;
• We propose an effective summary method of commodity short comments based on Chinese;
• Experiments on real data show that the proposed method can effectively summarize the commodity short comment information.

2. Related Work
The research on the generation of critical abstracts has been abroad for decades. In 1958, Luhn H.P. proposed an automatic summarization method based on keyword frequency statistics [2]. In 1962 Doyle L B came up with the idea of looking not just at the frequency of a single keyword, but more importantly at the combination of high frequency words and phrases [3]. In 1969, H. P. Edmundson purposed a new weighting method which treat key words, pragmatic words, title and heading words and structural indicators [4].Start at the end of the 80s, some experts and scholars in Chinese colleges,
universities and research institutes began to carry out some research and practice of Chinese abstract and made some gratifying achievements [5]. Since the late 1980s, Professor Wang Yongcheng has led the research group of ‘automatic abstracts’ to conduct in depth research on it [6] [7] [8]. The automatic abstract system development by Professor Wu Lide makes statistics on the input text, which has certain text comprehension ability [9].

Recently, there has been some work on summarizing online rate contents. Yue Lu, Cheng Xiangzhai and Neel Sundareaon studied the problem of generating a rated aspect summary of short comments, which is a decomposed view of the overall ratings for the major aspects [10]. Yang Ruixin has analysed the tendency of emotion and established semantic web to summarize the comments on the air conditioning products of e-commerce [11], which requires the former manual definition of emotional weights and other work. Bao Zhiqiang and Zhou Yipeng extract fresh comment keywords through LSTM (long Short Term Memory Neural Network) [12]. Some keywords were successfully extracted in this study, but there is no attempt to cluster the aspects. None of the above can directly or well meet the objectives of this paper.

3. Problem Statement
In this section, we formally define the problem we study and briefly summarize the notations used in this paper.

For a target entity with many short comments, our goal is to generate a summary of the comment content to help users better digest the content of the comments and understand the target entity from various dimensions. The summary result can be a series of phrases composed by noun-adjective. Therefore, through Chinese word segmentation and part of speech classification, we could extract those phrases and identify head term which is a noun and the modifier which is an objective.

In this section, you can use N-A phrase to describe a phrase composed by a noun and an adjective. For later writing and reference, the important symbols used later are defined here.

In this paper, \( T = \{t_1, t_2, t_3, \ldots\} \) represents a comment, where \( t_1, t_2, \ldots\ ) is the set of all comments.

\( Ph = \langle w_h, w_m \rangle \) (nominal phrase) : \( ph \) identifies nominal phrase, which is composed of \( \langle w_h, w_m \rangle \), where \( w_h \) is a noun, \( w_m \) adjective.

\( Ai = \{ph1 \text{ and } ph2, \ldots\\} \) (type): \( Ai \) represents all nominal phrases belonging to type \( i \) after clustering.

4. The Summarization of Commodity Short Comments
For the research content of this paper, the summarization of the comments is divided into the following steps.

![Figure 1. Processes of Summarization.](image)

Firstly, extracted the N-A Phrase from each comment, then the nouns in the N-A Phrase were clustered based on LDA. Finally, the results were presented by summarizing the contents.

4.1. N-A Phrase Extraction
In this paper, N-A Phrase is extracted from the short comment by Chinese word segmentation and part of speech classification.

Chinese word segmentation is to cut the Chinese characters in a sentence into individual Chinese words according to the sequence. In our research, we used the Chinese word segmentation tool of "jieba", which is an important third-party Chinese word segmentation function library. The classification of part of speech is to classify the part of speech of the words in a complete sentence, which is mainly based on the meaning of the words. We marked the part of speech in the segmentation stage.
4.2. The Topic Clustering of N-A Phrase Based on LDA

LDA (Latent Dirichlet Allocation) is a kind of document theme generation model, by Blei purposed in 2003[13]. The LDA model is a Bayesian probabilistic model which contains three layers of word, topic and document. In LDA, each document may be viewed as a mixture of various topics where each document is considered to have a set of topics that are assigned to it via LDA [14]. So, the probability of each word in the generated document is:

\[
\text{Pr}(\text{word} | \text{document}) = \sum_{\text{topic}} \text{Pr}(\text{word} | \text{topic}) \times \text{Pr}(\text{topic} | \text{document})
\]

The LDA model first selects a theme vector \(\theta\) to reorganize and determines the probability of each theme being selected. Then, in word generation, a theme \(z\) is selected from the vector, and then a word is generated according to the word probability distribution of theme \(z\).

\[\alpha\] is the parameter of the Dirichlet prior on the per-document topic distributions, \(\beta\) is the parameter of the Dirichlet prior on the per-topic word distribution. They define the document base in the model.

The LDA model adopts the Bag of Words (BOW) model. It treats each document as a word frequency vector, thus transforming text information into digital information that is easy to model. We build a thesaurus with \(V\) nouns based on the nouns of all the nominal phrases extracted from the reviews of the goods, and these \(V\) nouns are used in \((1, 2, ..., V)\) to mark. So, any given word \(v\) (1, 2, 3, ..., \(V\)) can be represented by the \(v\)-dimensional vector \(w\), where \(w_v = 1\) and \(w_u = 1, u\) does not equal \(V\).

Next, treat each comment as a document. A comment with \(N\) nouns is a document with \(N\) words. You can use the \(V\) vector \(W = (w_1, w_2, ..., w_m)\) stands for, \(w_n\) stands for the \(N\)TH word in it.

Finally, all \(M\) comments of an item constitute a set of documents containing \(M\) documents, which can be expressed by the \(m\)-dimensional vector \(D = (W_1, W_2, ..., W_m)\).

Therefore, the generation process of given document \(w\) from dataset \(D\) is as follows:

- Calculate the size of the feature words in a single document based on the Poisson distribution to generate \(N \sim \text{Poisson}(\zeta)\). \(N\) obeys the Poisson distribution, where \(N\) is the length of the document;
- Calculate the distribution probability vector of topic in a single document as being \(\theta \sim \text{Dirichlet}(\alpha)\) according to Dirichlet distribution. \(\theta\) is a \(k\) vector that follows a Dirichlet distribution;
- For any word in the document \(w_n\), randomly selected from the Multinomial distribution of vector \(\theta\) a potential theme \(z_n \sim \text{Multinomial}(\theta)\), then from \(z\) polynomial distribution in probability \(p\) (\(w_n | \text{zinc, beta}\) pick a key \(w_n\).

In this paper, LDA generally uses Gibbs sampling, a special case of Markov chain monte Carlo algorithm, to approximate the main parameters. Gibbs sampling was used to estimate the parameters of the LDA model, according to the following:

\[
P(z_i = s | Z_{-i}, W) \propto \frac{n_{s,i} + \beta_i}{\sum_{l=1} n_{s,l} + \beta_i} \times (n_{s,i} + \alpha_s)
\]

Where, \(Z_i = s\) represents the probability that word \(w_i\) belongs to the topic of \(s\), \(z_i\) represents the probability of all other words, \(n_{s,i}\) represents the number of words \(w_i\) without the current word.
assigned to the current topic \( z_s \), and \( n_{(s,j)} \) represents the number of words \( d_j \) without the current document assigned to the current topic \( z_s \).

Empirical values are pre-set for prior parameters \( \alpha \) and \( \beta \). \( \alpha = \frac{50}{K} \) and \( \beta = 0.1 \), and the parameter \( K \) is selected with the degree of confusion, which is a common evaluation standard in the statistical language model, let \( K = 50 \) [15].

4.3. Content Summarization

Finally, the results should be presented as N-A Phrases. Each of it contains a noun and an adjective. This section describes how we choose the noun and select the adjectives.

- **Choose the noun:** the reason we choose a noun first is that we summarize the content by nouns in N-A Phrases. First to select a few contents for each type of \( A_i \). The noun selection is based on the probability that each noun belongs to the type which is summarized by LDA. Therefore, in each \( A_i \), nouns are selected according to their probability to obtain the first few contents with the highest probability, so as to determine the set of nouns to be displayed for each type.

- **Select adjectives:** for each noun, there may be multiple adjectives to describe it. Here we simply choose adjectives based on how often they appear in selected N-A Phrase. For all adjectives corresponding to the selected noun, calculate its occurrence frequency \( f \),

\[
F(A_i, w_h \,'
\}) = \{ph/A(\{ph\}) = i, ph.w_h = w_h \}
\]

\[
f(w_m, w_h) = c(w_m, F(A_i, w_h))
\]

Where, \( F(A_i, w_h \) \) is the list of all N-A Phrase \( ph \) of type \( A_i \), \( w_h \) is the noun, and \( c(w_m, F(A_i, w_h)) \) is the occurrence number of adjective \( w_m \) in all N-A Phrase \( ph \) of \( F(A_i, w_h) \).

According to this formula, the frequency of each adjective corresponding to each noun in each type is calculated, and the selected adjective with the highest frequency is selected as the selected noun in each selected type. The final noun - adjective pairing is successful to all classifications shown in the summary results.

5. Optimization

5.1. Data Pre-processing

There are abnormal contents, repeated comments and automatic evaluation in the original comment data. The value of information in this part is low and its data structure is chaotic. Before building a model, data cleansing is essential. Secondly, Chinese word segmentation and part of speech classification are optimized.

![Figure 3. Processes of Data Pre-processing Optimization](image)

- **Text de-duplication:** delete text over repeated comment data so that it remains unique. When using pandas, use the unique method to quickly remove duplicate data.
• **Clean up trash comments:** trash comments are content that has no real meaning in subsequent processing and can impact model results, including comments automatically generated by the system and the content contained in the hypertext markup language.

• **Mechanical compression of words:** in the original data, some text comments have the continuous repetition of words, which is lack of practical significance and will affect the results summarized by the model. This situation needs to be corrected, for example, a sentence like ‘very satisfied, very satisfied, very satisfied’, needs to be corrected to be ‘very satisfied’.

• **Short sentence deletion:** The fewer words a comment contains, the less meaning it has and the less information it can mine. We decided that those sentences which contains less than three characters will be ignored.

5.2. *Optimization of Chinese Word Segmentation and Part of Speech Classification*

• **Set stop word list:** stop word is the word that is considered meaningless in the subsequent processing, and it is the word that needs to be eliminated in the word segmentation stage. In this paper, we set a stop word list containing 1419 stop words.

• **Dictionary customization and word frequency adjustment:** the Chinese word segmentation tool "jieba" supports specifying a custom dictionary to contain words that are not in the "jieba" vocabulary. The custom dictionary used in this paper comes from the dictionary provided by the sogou input method. In the end, we built a new dictionary with 50,313 new words. At the same time, the word frequency of 16 words is improved through the analysis of the comment content.

6. Experiments

6.1. *Data Set Preparation*

We create a data set by collecting feedback comments from Taobao. All the data used in this paper is Taobao real comment data, and some statistics of the data is shown in Table1.

| Product Name                       | Affiliated Business           | Number of Comments |
|------------------------------------|-------------------------------|--------------------|
| Nintendo Switch Consoles           | Tmall Global Official Store   | 1980               |
| Nintendo Switch Consoles           | Dasbabylandde Overseas        | 1443               |
| Midea F60-15WB5(Y) Water Heater    | Midea Meihao Specialty       | 1980               |
| Midea F60-15WB5(Y) Water Heater    | Midea Official Store          | 576                |

Data is the real data obtained by crawling through the web crawler we designed and implemented. In the Taobao website, through itemId, spuId and sellerId, we specifically locate a specific commodity in the Taobao mall and access its corresponding comment information. In total, more than 5,979 comments were obtained.

6.2. *Sample Result of Short Comments Summarization*

A sample short comments summarization result of the commodities are shown in the following tables. For each commodity’s comparison, two tables will be shown. The first table of each commodity contains the difference with and without the optimization. The second table of each commodity contains the summarization result from different merchants.

The difference between the results before and after the optimization process mentioned above for the ‘Nintendo Switch Consoles’, which is owned by ‘Tmall Global Official Store’, is shown in the Table2.
Table 2. Sample Comparison of Optimization for ‘Nintendo Switch’

| Topic | Before Optimization                                                                 | After Optimization                                                                 |
|-------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| topic0: | 税/快,机器/好,美版/不错,游戏/贵,有点,开机/红 | 税/快,物流/慢,税/好,客服/好,海关/久,价格/不错 |
| topic1: | 评论/好,客服/好,用户,卖家/旺旺,老公/惊喜,宝贝 | 机器/不错,游戏/贵,物流/慢,税/好,美版不错,开机/红 |
| topic2: | 物流/慢,东西,海关/好,天猫/好,速度/快,棒棒/厉害 | 卖家/好,速度/快,购物/愉快,态度/好,质量/好,店家/好 |

The difference between the ‘Nintendo Switch Consoles’ from ‘Dasbabylandde Overseas’ and ‘Tmall Global Official Store’ is shown in the Table3.

Table 3. Sample Comparison of ‘Nintendo Switch’ from Different Merchants

| Topic | Dasbabylandde Overseas                                                                 | Tmall Global Official Store                                                                 |
|-------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|
| topic0: | 物流/慢,客服/耐心,发货/慢,购物/好,体验/不错 | 退税/快,物流/慢,税/好,客服/好,海关/久,价格/不错 |
| topic1: | 游戏/不错,评价/不错,机器/好,感觉/不错,质量/好,服务态度/好 | 机器/不错,游戏/贵,物流/慢,税/好,美版不错,开机/红 |
| topic2: | 正品/好,机子/不错,开机/红,游戏机/好,手柄/松动 | 卖家/好,速度/快,购物/愉快,快,态度/好,质量/好,店家/好 |

Through a simple analysis of their comments, it can be concluded that for the goods from ‘Dasbabylandde Overseas’, buyers have more complaints about delivery and logistics, and they have a positive attitude towards the quality of the goods and the service attitude of the merchants. For commodities from ‘Tmall Global Official Store’, buyers have more diverse opinions on logistics, but they also mention matters related to customs and tax refund.

The difference between the results before and after the optimization process mentioned above for the ‘Midea F60-15WB5(Y) Water Heater’, which is owned by ‘Midea Official Store’, is shown in the Table4.

Table 4. Sample Comparison of Optimization for ‘Midea Water Heater’

| Topic | Before Optimization                                                                 | After Optimization                                                                 |
|-------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| topic0: | 热水器/好,速度/快,购物/舒服,宝贝,评价/好,手机 | 美的/好,热水器/好,客服/好,人/好,热水/快,发货/快 |
| topic1: | 东西,售后/不错,质量/好,爸爸,空间/凉快,顾客/急急 | 安装/好,物流/快,服务态度/好,师傅/好,收费高,安装费/贵 |
| topic2: | 师傅/好,物流/快,评价/好,用户,客服/好,美的/好 | 购物/舒服,售后/不错,速度/快,质量/不错,掌柜/好 |

The difference between the ‘Midea F60-15WB5(Y) Water Heater’ from ‘Midea Meihao Specialty’ and ‘Midea Official Store’ is shown in the Table5.
### Table 5. Sample Comparison of ‘Midea Water Heater’ from Different Merchants

| topic0          | Midea Meihao Specialty                                                                 | Midea Official Store             |
|-----------------|----------------------------------------------------------------------------------------|---------------------------------|
| topic1          | 安装/好, 师傅/好, 物流/快, 服务态度/好, 价格/正好                                      | 安装/好, 物流/快, 服务态度/好, 收费高, 安装费/贵 |
| topic2          | 速度快/好, 货/好, 购物/好, 热水/快, 效果/好, 保温/好             | 购物/舒服, 售后/不错, 速度/快, 质量/不错, 掌柜/好 |
| topic0          | 美的/好, 热水器/好, 客服/好, 热水器/快, 客服/耐心, 卖家/好, 质量/好            | 热水器/好, 客服/好, 热水/快, 发货/快 |

Through a simple analysis of their comments, it can be concluded that buyers have a positive attitude towards the product quality and installation services of the two merchants. However, the evaluation of ‘Midea Official Store’ includes complaints on its price, while the evaluation of ‘Midea Meihao Specialty’ is mostly positive.

### 7. Conclusions and Future Work

In this paper, we formally defined the problem of summarization of commodity short comments, in which the summary was described in the form of noun and adjective phrases. Then, we propose an effective summary method of commodity short review based on Chinese. Furthermore, we designed and implemented a web crawler for obtaining Taobao data. Finally, the experiment on real data shows that the proposed method can effectively summarize the commodity short comments. In our research process, we also found that the particle results should be re-correlated here, and the accuracy of phrase recognition still needs to be improved. These will be interesting directions for future work.

### 8. Acknowledgement

This work is supported by of the Key Disciplines of Computer Science and Technology of Shanghai Polytechnic University under Grant No.XXKZD1604.

### 9. References

[1] LU Kun. Research on Externalization of Tacit Knowledge Based on Web 2.0[J]. *Information Science*, 2008(02):247-251.
[2] Luhn H P. The automatic creation of literature abstracts[J]. *IBM Journal of research and development*, 1958, 2(2): 159-165.
[3] Doyle L B. Indexing and abstracting by association[J]. *American documentation*, 1962, 13(4): 378-390.
[4] Edmundson H P. New methods in automatic extracting[J]. *Journal of the ACM (JACM)*, 1969, 16(2): 264-285.
[5] Lu Xiaoping. An Overview of Researching on the Techniques of the Automatic Creation of Literature Abstracts[J]. *Journal of Modern Information*, 2009, 29(10):221-225.
[6] Wang Yongcheng, Wang Jicai. Automatic Compilation of Chinese Text Summaries [J]. *Microcomputer Applications*, 1998, (2):13 -16.
[7] Wang Yongcheng, Xu huimin. Oa-1.4 Version of Chinese Automatic Summary System [J]. *High Technology Letters*, 1998, (1):19 -23.
[8] Mo Yan, Wang Yongcheng. Automatic Compilation of Abstracts of Chinese Literature [J]. *New Technology of Library and Information Service*, 1993, (3):10 -12.
[9] Wang Wenxin.Huang Xuanjing,Wu Lide. A Chinese Automatic Abstracting System Based on Statistical Method[J]. *Computer Applications and Software*, 2000(09):28-33.
[10] Lu Y, Zhai C X, Sundaresan N. Rated aspect summarization of short comments[C]//Proceedings of the 18th international conference on World wide web. ACM, 2009: 131-140.

[11] Yang Ruixin. Sentiment Analysis of Electricity Supplier Air Conditioning Products[D]. Shanxi University, 2017.

[12] Bao Zhiqiang, Zhou Yipeng. Keywords Extraction of Fresh Product Reviews Based on LSTM[J]. Fujian Computer, 2018, 34(10): 90-92.

[13] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.

[14] Zhong Rongfei. Research on Network Public Opinion Monitoring Based on Topic Model[J]. Computer Knowledge and Technology, 2015, 11(7): 90-93.

[15] Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, Sheng Tang. A density-based method for adaptive LDA model selection [J]. Neurocomputing, 2008, 72 (7).