Computational Linguistics and Chinese Language Processing
Vol. 22, No. 1, June 2017, pp. 53-94

An Approach to Extract Product Features from Chinese Consumer Reviews and Establish Product Feature Structure Tree

Xinsheng Xu[*], Jing Lin[*], Ying Xiao[*] and Jianzhe Yu[*]

Abstract

With the progress of e-commerce and web technology, a large volume of consumer reviews for products are generated from time to time, which contain rich information regarding consumer requirements and preferences. Although China has the largest e-commerce market in the world, but few of researchers investigated how to extract product feature from Chinese consumer reviews effectively, not to analyze the relations among product features which are very significant to implement comprehensive applications. In this research, a framework is proposed to extract product features from Chinese consumer reviews and construct product feature structure tree. Through three filtering algorithms and two-stage optimizing word segmantation process, phrases are identified from consumer reviews. And the expanded rule template, which consists of elements: phrase, POS, dependency relation, governing word, and opinion, is constructed to train the model of conditional random filed (CRF). Then the product features are extracted based on CRF. Besides, two index are defined to describe product feature quantitatively such as frequency and sentiment score. Based on these, product feature structure tree is established through a potential parent node searching process. Furthermore, categories of extensive experiments are conducted based on 5,806 experimental corpuses from *taobao.com*, *suning.com*, and *zhongguancun.com*. The results from these experiments provide evidences to guide product feature extraction process. Finally, an application of analyzing the influences among product features is conducted based on product feature structure tree. It provides valuable management connotations for designer, manufacturer, or retailer.

[*] China Jiliang University
E-mail: {lionkingxxs, linjing, xiaoying, yujianzhe}@cjlu.edu.cn
The author for Correspondence is Xinsheng Xu.

Keywords：Chinese Consumer Review, Product Feature Extraction, Rule Template, Sentiment Analysis, Product Feature Structure Tree

1. Introduction

With the rapid expansion of e-commerce business, the Web has become an excellent source for gathering consumer reviews about products (Turney, 2002; Dave, Lawrence & Pennock, 2003; Dellarocas, 2003; Godes & Mayzlin, 2004; Hu & Liu, 2004a, b; Liu, Hu & Cheng, 2005; Duan, Gu & Whinston, 2008; Forman, Ghose & Wiesenfeld, 2008). Many product review websites (e.g., Amazon.com, Taobao.com) have been established to collect consumer opinions about products. Consumers also comment on products in their blogs, which are then aggregated by Blogstreet.com and AllConsuming.net etc. In addition, it has become a common practice for retailers (e.g., Amazon.com, taobao.com, jd.com) or manufacturers to provide online forums that allow consumers to express their opinions about products they have purchased or in which they are interested. Consumer reviews are essential for both retailers and product manufacturers to understand the general responses of consumers to their products. Proper analysis and summarization of consumer reviews can further enable retailers or product manufacturers to insight consumers' opinions about specific features of products (Liu *et al.*, 2005). Consumer reviews also offer retailers a better understanding of the specific preferences of individual customers. Furthermore, from a consumer perspective, consumer reviews provide valuable information for purchasing decisions.

As the number of consumer reviews expands, however, it becomes more difficult for users (e.g., product designer & manufacturers, consumers) to obtain a comprehensive view of consumer opinions pertaining to the products through a manual analysis. Consequently, an efficient and effective analysis technique that is capable of extracting the product features stated by consumers and summarizing the sentiments pertaining to specific product features automatically becomes desirable. This analysis essentially consists of two main tasks: product feature extraction from consumer reviews and opinion orientation identification for these product features (Hu & Liu, 2004a, b; Popescu & Etzioni, 2005; Jindal & Liu, 2006; Wei, Chen, Yang & Yang, 2010).

Product feature extraction is crucial to sentiment analysis, because its effectiveness significantly affects the performance of opinion orientation identification. Several product feature extraction techniques have been proposed in the literatures (Hu & Liu, 2004a, b; Kobayashi, Inui, Matsumoto, Tateishi & Fukushima, 2004; Kobayashi, Iida, Inui & Matsumotto, 2005; Popescu & Etzioni, 2005; Wong & Lam, 2005, 2008; Bahu & Das, 2015). However, product feature extraction and opinion orientation identification suffer huge challenges for Chinese consumer reviews because of the natural complexity of Chinese language (Zhang, Yu, Xu & Shi, 2011; Song, Yan & Liu, 2012; Li, 2013; Zhou, Wan & Xiao, 2013; Liu, Song, Wang, Li & Lu, 2014; Wang, Liu, Song & Lu, 2014). First, there is always no interval between the words of Chinese sentences. It leads to the difficulty of distinguishing Chinese phrases. Besides, some Chinese phrases have synonyms e.g. "电板" (electroplax), which exactly appears at Chinese consumer reviews although it is rare, is a synonym of "电池" (Battery). This kind of product features cannot be recognized and extracted based on frequency item method. Moreover, the syntactic and grammar of Chinese sentences are very complex as well as their structures, e.g. the consumer review "电池/noun 还/adverb 可以/verb" (The battery is good) always expresses the positive evaluation of consumers for the "电池"(Battery). The phrase "可以" (can)[1] is a verb but it acts as an opinion word that modifies the phrase "电池" (Battery). That means, on the context of Chinese language, verbs may also modify nouns or noun phrases and express opinion orientation. Thus, the existing methods that find product features based on adjective are also not enough for Chinese consumer reviews. In addition, there are some specific correlations among product features according to our observations. Some product features extracted from consumer reviews are the attributes of the product, components, or parts such as function, performance, quality, material, and service while some product features are product, components or parts itself. For example, "摄像头(cameral)" and "像素(pixel)" are two product features. The "摄像头(cameral)" is a component of intelligent mobile phone while the "像素(pixel)" is the attribute of "摄像头(cameral)". There is a description relation between the "像素(pixel)" and the "摄像头(cameral)". Therefore, it is always interrelated among product features. How to extract product features effectively from Chinese consumer reviews and establish the interrelations among product features are difficult tasks and huge challenges. This paper focuses on such a text mining issue of Chinese consumer reviews. More specifically, we will establish a structure tree of product features and infer the key factors of influencing the sentiment scores of product features from consumers. The goal is to provide evidences for the designer & manufacturer to improve and update their products effectively.

With these considerations, a technique framework of extracting product features from Chinese consumer reviews and its applications are proposed in which a two-stages optimizing word segmentation solution is proposed to improve the correct rate of word segmentation for supporting product feature extraction from Chinese consumer reviews, and an expanded rule template for CRF, in which two new elements namely governing word and opinion word are added, is developed to deal with complex syntaxes and grammars of Chinese language and implicit opinion words. This increases the precision of product feature extraction and is also helpful for the sentiment analysis for product features. Furthermore, product feature structure

[1] The phrase "可以" expresses a positive opinion that means "good" in Chinese language. However, its literal meaning corresponds to the word "can" in English language. The POS of it defines as verb at word segmanetation process.

tree is constructed considering the natural internal correlations among product features, and an application of inferring the key factors, that influence the preference of consumers for a product feature, is proposed based on Bayes theory whose results can be used as evidences for designers, manufacturers, or retailers to product improvement, market management, etc. Finally, 5,806 consumer reviews from *taobao.com*, *suning.com*, and *zhongguancun.com* are retrieved and used as corpus to explain the applications of these principles and methods proposed in this work. It is innovative method of implementing comprehensive applications based on Chinese consumer reviews at product feature level.

The remainder of this article is organized as follows: In Sect. 2, we review existing product feature extraction techniques and discuss their fundamental limitations to highlight our research motivation. Subsequently, a technique framework of extracting product features from Chinese consumer reviews and its applications are proposed in Sects. 3. Sects. 4 investigate the methods of extracting product features based on CRF. The quantitative characters of product feature including frequency and sentiment score are explored in Sects. 5. On the basis of these, product feature structure tree is constructed in Sects. 6. Categories of extensive experiments are conducted in Sects. 7. Sects. 8 give an example to illustrate the applications of the methods mentioned in this work. Secs.9 discuss our research works. Finally, we conclude with a summary and some future research directions in Sect. 10.

2. Literature Review

Some researchers have devoted to analyzing consumer reviews for valuable information and implementing applications based on it. These analyses and applications essentially consist of two aspects: product feature extraction and opinion orientation identification. Product feature extraction is the foundation of opinion orientation identification, and opinion orientation identification is the application based on product features.

2.1 Product Feature Extraction

Hu and Liu (2004a, b) assumes that product features must be nouns or noun phrases and employs the association rule mining algorithm (Agrawal & Srikant, 1994; Srikant & Agrawal, 1995) to discover all frequent itemsets (i.e., frequently occurring nouns or noun phrases) within a target set of consumer reviews. In addition to association rule mining, other information-extraction-based product feature extraction techniques have also been proposed (Kobayashi, Inui, Matsumoto, Tateishi & Fukushima, 2004; Kobayashi, Iida, Inui & Matsumotto, 2005). Popescu and Etzioni employ KnowItAll and propose OPINE to extract product features from consumer reviews automatically (Popescu & Etzioni, 2005; Etzioni *et al.*, 2005). Using a set of domain-independent extraction patterns predefined in KnowItAll, OPINE instantiates specific extraction rules for each product class under examination and then uses these rules to extract possible product features from the input consumer reviews. Wong & Lam (2005, 2008) employ Hidden Markov Models and CRF, respectively, as the underlying learning method to extract product features from auction websites. Liu, Wu & Yao (2006) adopted supervised method to extract product features and compare variety of products for consumers based on them. Choi and Cardie (2009) presented the methods of recognizing the product feature from consumer reviews based on CRF.

2.2 Opinion Orientation Identification

Opinion orientation identification is to determine the sentiments of consumers for product features. Therefore, product feature extraction and opinion orientation identification cannot be separated in practice. Li *et al*. (2010) researched the extraction methods of opinion words for product features by integrating two CRF variables such as Skip-CRF and Tree-CRF. Htay and Lynn (2013) extracted product features and opinion words using pattern knowledge in customer reviews. Yi and Niblack (2005) worked on identifying the specific product features and opinion sentences by extracting noun phrases of specific patterns. Zhuang, Feng and Zhu (2006) proposed a supervised learning method based on dependency grammatical graph to extract product feature and opinion information. Yin and Peng (2009) studied the sentiment analysis for product features in Chinese reviews based on semantic association. Ouyang, Liu, Zhang and Yang (2015) investigated features-level sentiment analysis of movie reviews. And Chen, Qi and Wang (2012) extracted multiple types of feature-level information from consumer reviews.

In addition, topic/opinion summary is also an important aspect based on product feature extraction and opinion orientation identification. For example, Miao, Li and Zeng (2010) executed the topic extraction from movie reviews based on CRF. Turney (2002) investigated the unsupervised classification of reviews based on semantic orientation.

However, the existing product feature extraction and application techniques for English language cannot be used to deal with Chinese language directly because of the natural complexity of Chinese language mentioned above. Then some experts explore the product feature extractions and applications from Chinese consumer reviews. Li, Ye, Li and Law (2009) and Zu and Wang (2014) researched product feature extracting methods from Chinese customer online reviews. Liu and Wang (2013) proposed a keywords extraction method based on semantic dictionary and lexical chain. Ma and Yan (2014) presented the product features extraction of online reviews based on LDA model. In order to process Chinese language sentences effectively, Liu and Ma (2009) investigated the Chinese automatic syntactic parsing issues. Similarly, Li (2013) researched the Chinese Dependency Parsing for product feature extraction. Jiang *et al*. (2012) also proposed a method to enhance the feature engineering for CRF by using unlabeled data. From the perspective of applications, Chang, Chu, Chen and

Hsu (2016) investigated the linguistic template extraction for reader-emotion features based on Chinese text. Wang and Meng (2011) studied the opinion object extraction based on the syntax analysis and dependency analysis. Lv, Zhong, Cai and Wu (2014) investigated the task of aspect-level opinion mining including the extraction of product entities from Chinese consumer reviews. Besides, Hu, Zheng, Wu and Chen (2013) developed a method of extracting product characteristic from consumer reviews to provide users with accurate product recommendation. Dai, Tsai and Hsu (2014) presented a joint learning method of entity linking constraints from Chinese consumer reviews based on markov-logic network. Wang and Wang (2016) investigated comparative network for product competition in feature-levels through sentiment analysis. These literatures exactly proposed some effective methods of extracting product features from Chinese text, and used them at specific research tasks. These methods can be classified into two major approaches: supervised and unsupervised.

Supervised product feature extraction techniques require a set of preannotated review sentences as training examples while unsupervised product feature extraction approach automatically extracts product features from consumer reviewers without involving training examples. Generally, the supervised methods have better results at the precision, recall or *F*-score than those of the unsupervised methods because it can set the training samples according to specific research or application goals (Li *et al*., 2009; Zu & Wang, 2014; Ma & Yan, 2014). This work focuses on supervised product feature extraction issues and its applications.

3. Product Feature Extraction Technique for Chinese Consumer Reviews

Aiming at Chinese consumer reviews, a technique framework of product feature extraction is proposed that consists of three key phases: word segmentation and optimization, product feature extraction based on CRF, and the quantitative descriptions of product features. The proposed technique begins with the preprocessing of the inputting consumer reviews, where the preprocessing task includes word segmenting & POS tagging, reconstructing noun phrase based on N-gram, filtering and optimizing. Subsequently, product feature extraction process employs CRF to identify product features in which a train set and a rule template for constructing the *model* of CRF are developed. Based on the extracted product features and the results of word segmentation, the quantitative descriptions for product features including the frequency of product feature and the sentiment score of product feature, are constructed. On the basis of these, product feature structure tree is established based on the fact that product features are interrelated. **Figure 1** presents the framework of product feature extraction techniques for Chinese consumer reviews. In the following subsections, we will depict the detailed design and implementation of each phase.

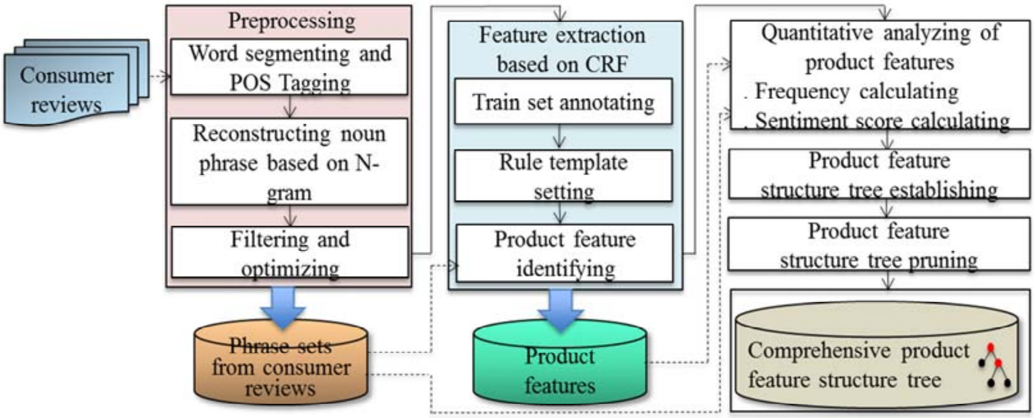

Consumer reviews
Preprocessing
Word segmenting and POS Tagging
Reconstructing noun phrase based on N-gram
Filtering and optimizing
Phrase sets from consumer reviews
Feature extraction based on CRF
Train set annotating
Rule template setting
Product feature identifying
Product features
Quantitative analyzing of product features
. Frequency calculating
. Sentiment score calculating
Product feature structure tree establishing
Product feature structure tree pruning
Comprehensive product feature structure tree

Figure 1. Overview of product feature extraction techniques for Chinese consumer reviews

3.1 Preprocessing Techniques

Preprocessing techniques consist of word segmenting and POS tagging, reconstructing noun phrase based on N-gram, filtering and optimizing.

Phase A *Word Segmenting and POS Tagging*

Word segmenting and POS tagging start with the inputting review sentence S, and end with the pairs ($wordi$, $POSi$), where $wordi$ is the ith word contained in sentence S, and $POSi$ is the POS tagging result of the $wordi$. For the convenience of presentation and measure, phrase (word), sentence, and consumer review are defined respectively as **Figure 2**. In this work, the word refers to phrase in general unless there are specific instructions. For the review sentence S "手机的屏幕很模糊(The screen of this phone is very indistinct)", the word segmenting and its POS tagging are as follows: (手机(phone), n), (的[2], ude1) , (屏幕(screen), n), (很(very), d), (模糊(indistinct), a) illustrated in **Figure 2**. At the same time, the dependency relations among these words and their governing words are also identified through syntactic parsing process based on consumer review (Liu & Ma, 2009; Wang & Meng, 2011; Li, 2013; Dai *et al.*, 2014). The objective of this phase is to divide the review sentences into discrete phrases and annotate its POS tag, and provide the data resource for the next analysis phases.

[2] It is an auxiliary word in Chinese language. There is no word corresponding to it in English language.

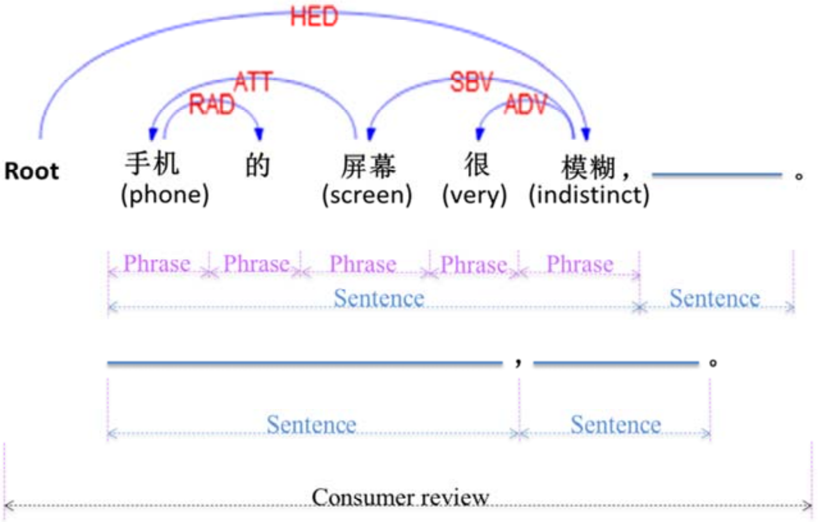


Figure 2. Word segmenting and its POS tagging for a case

Phase B *Reconstructing Noun Phrase based on N-gram*

Word segmenting process may generate some incorrect results sometimes. For example, the phrase ”分辨率(resolution)” always be divided into three kinds of independent phrases “分(divide)”, “分辨(distinguish)”, and “率(rate)”. However, the phrase “分辨率(resolution)” should be a complete phrase for digital product e.g. intelligent mobile phone. Obviously, it is an incorrect result of word segmenting. In order to deal with this problem, it is necessary to recombine these fragmental phrases into its correct form. A reconstruct method based on *n*-gram is introduced which consists of two steps: (a) Identifying the number n of the *n*-gram method reasonably; (b) Constructing new phrases according to giving number n. Using $wordi$ as an example and assuming *n*=3, then new phrases can be generated by recombining it with adjacent words from left and right directions, respectively. For example, the reconstructing new phrases based on *3*-gram method are as follows ($word_{(i-1)}word_i$, $word_{(i)}word_{(i+1)}$, $word_{(i-2)}word_{(i-1)}word_i$,$word_{(i)}word_{(i+1)}word_{(i+2)}$, $word_{(i-3)}word_{(i-2)}word_{(i-1)}word_i$, $word_{(i)}word_{(i+1)}word_{(i+2)}word_{(i+3)}$. After this reconstructing process, the phrase “分辨率(resolution)” that was incorrect segmented will be restored to its correct from. Likeness, all other incorrect segmented phrases can also be restored to their correct forms through this kind of reconstructing process.

Unfortunately, this phase may also lead to other error phrases due to over-combination. Thus we also need to optimize the results generated from reconstructing phase.

Phase C *Filtering Algorithms*

In order to remove the over-combination phrases from ***Phase*** B, a series of filtering algorithms are employed.

(Ⅰ) Frequency filtering. In general, some new combination phrases which are incorrect such as "屏幕很(screen very)" or "的屏幕('s screen)" seldom occurrence at consumer reviews. Therefore, we can remove them through frequency filtering process by setting a reasonable threshold. An expression for frequency filtering is generalized as follows:

$$\text{If } Number(word_i') \le Q_1 \text{ then remove it from } \Omega \tag{1}$$

where $word_i'$ is a phrase generated from *Phase B*. And Ω is the phrase group of $word_i'$'s. $Number(word_i')$ is the function that calculates the number of the $word_i'$ appearing at consumer reviews. Q_1 is the threshold of frequency filtering process.

This filtering rule means that the $word_i'$ whose frequency appearing at consumer reviews less than Q_1 will be removed from Ω.

(Ⅱ) Cohesive filtering. However, there are another kind of phrases such as "就这样(That's it)" which consist of two frequency words "就[3]" and "这样(this/it)", and is also a frequency phrase because of the expression habit of Chinese. But it is not a valid phrase. This kind of phrases still cannot be removed through frequency filtering process only.

According to our observation, the constitute elements of a phrase, for example "分辨(distinguish)" and "率(rate)" are two constitute elements of the phrase "分辨率(resolution)", are always strongly coupled among them. That means the cohesive among them is very strong. However, the cohesive among the constitute elements of the over-combination phrases generated from *Phase* B is weak because the combination form of these elements is seldom or may not exist at consumer reviews at all. Therefore, we can use cohesive to remove these phrases from the results of *Phase B*. The cohesive among the constitute elements of a phrase is generalized as follows (Li *et al*., 2009):

$$Coh(word_i') = {Fre(word_j'')_{(word_i')}}\Big/{(Fre(word_i') + 1)} \text{ and } word_j'' \subset word_i' \tag{2}$$

where $Fre(word_j'')$ is the frequency of phrase $word_i'$ occurring at the results of original word segmentation. $word_j''$ is one of the constitute elements of phrase $word_i'$. $Fre(word_j'')_{(word_i')}$ is the frequency of the constitute elements $word_j''$ of phrase $word_i'$ occurring at the results of original word segmentation.

Then, the expression of cohesive filtering is generalized as follows:

[3] It is an auxiliary word in Chinese language. There is no word corresponding to it in English language.

$$\text{If } Coh(word_i') \geq Q_2 \quad \text{then } word_i' \text{ is not a correct phrase} \tag{3}$$

Through cohesive filtering process, the over-combined frequency phrases that consist of two frequency words can be removed from phrase set $word_i'$.

(Ⅲ) Left entropy and right entropy filtering. In addition, a complete phrase always has various neighbors including left neighbors and right neighbors. If a phrase has a fixed neighbor either left neighbor or right neighbor, it is always not a complete phrase. For example, phrase "诺基亚" (Nokia: a band of mobile phone) should be a complete phrase. But it is always divided into two separated words "诺基[4]" and "亚[5]". Although the process of reconstructing phrase can generate its complete form "诺基亚(Nokia)", but some incorrect word segmentation results such as "诺基" and "亚" still exist at the original word segmentation results. Therefore, it is necessary to remove these phrases from the original word segmentation results to keep the accuracy of word segmentation results.

The calculation models of the left entropy and the right entropy are defined as follows, respectively (Li *et al*., 2009):

$$\textbf{Left entropy: } H_L(U) = \sum_i \frac{C_{Li}}{n} \times log\frac{C_{Li}}{n} \tag{4}$$

where C_{Li} is the number of the ith left neighbor appearing at the results of the original word segmentation. n is the number of the current phrase appearing at the results of the original word segmentation. $H_L(U)$ is the left entropy of the current phrase.

$$\textbf{Right entropy: } H_R(U) = \sum_i \frac{C_{Ri}}{n} \times log\frac{C_{Ri}}{n} \tag{5}$$

where C_{Ri} is the number of the ith right neighbor appearing at the results of the original word segmentation. $H_R(U)$ is the right entropy of the current phrase.

On the basis of these, an expression of the left entropy and right entropy filtering is generalized as follows:

$$\text{If } H_L(U) \geq Q_3 \text{ or } H_R(U) \geq Q_3 \text{ then } word_i' \text{ is not a complete phrase} \tag{6}$$

where Q_3 is the threshold of the left entropy and right entropy filtering process.

3.2 Optimizing Word Segmentation Process

Through reconstructing phrase and three filtering processes, some incorrect word segmentations are removed from the results of original word segmentation and some fragmented phrases are restored also. Besides, some valuable new phrases corresponding to specific research object can also be found during these processes. By adding these new

[4] 诺基亚 is a transliteration word of brand name of mobile phone in Chinese language. There is no word corresponding to "诺基" in English language.

[5] Likeness, there is no word corresponding to "亚" in English language.

phrases into the user dictionary which is the important evidences of word segmentation process, then the word segmentation process will restart again based on this extended user dictionary. Thus, the process of word segmentation in this work contains two stages which is presented in **Figure 3**. These two stages can optimize the results of word segmentation to provide valid data resources for the next product feature extraction process.

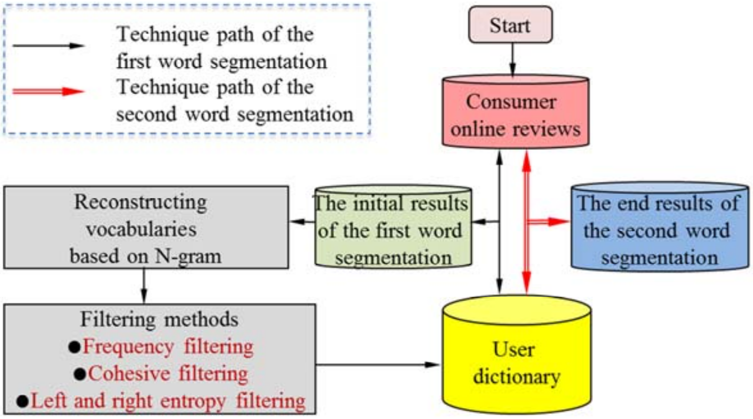

Technique path of the first word segmentation
Technique path of the second word segmentation
Start
Consumer online reviews
Reconstructing vocabularies based on N-gram
The initial results of the first word segmentation
The end results of the second word segmentation
Filtering methods
●Frequency filtering
●Cohesive filtering
●Left and right entropy filtering
User dictionary

Figure 3. Two-stages word segmentation process

4. Product Feature Extraction based on CRF

The CRF (Lafferty, McCallum & Pereira, 2001; Jakob & Gurevych, 2010) is a sequence modeling framework that can solve the label bias problem in a principled way. CRF has a single exponential model for the joint probability of the entire label sequence given the observation sequence which assign a well-defined probability distribution over possible labeling, trained by maximum likelihood or MAP estimation. Therefore, the weights of features at different states can be traded off against each other. CRF perform better than HMMs and MEMMs when the true data distribution has higher-order dependencies than the model, as is often the case in practice (Zheng, Lei, Liao & Chen, 2013; Zhang & Li, 2015). With these considerations, CRF is employed to extract product features from Chinese consumer reviews in this work. The principles of CRF can be described as follows:

Let X is a random variable over data sequences to be labeled. Y is a random variable over corresponding label sequences. And $X=(x_1,x_2,\cdots,x_t)$ might range over natural language sentences, and x_i denotes the ith phrase in X. $Y=(y_1,y_2,\cdots,y_t)$ range over POS taggings of those sentence Xs, and y_i is the POS tag of the phrase x_i. It is illustrated in **Figure 4**. The random variables X and Y are jointly distributed. CRF, with the known observation data sequence X, calculate the conditional probability $p(Y|X)$. As a result, the POS tag sequence Y that corresponds to the maximum value of the conditional probability $p(Y|X)$ will be label sequence of the X.

The conditional probability $p(Y|X)$ can be calculated as follows:

$$p(Y|X) = \frac{1}{Z(x)} \exp(\sum_i \sum_k \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_i \sum_k u_k s_k(y_i, x, i)) \tag{7}$$

where $t_k(y_{i-1}, y_i, x, i)$ is the transfer character function. It denotes that the label corresponding to the $(i-1)$th element in the observation sequence X is y_{i-1}, and the label corresponding to the i th element in the observation sequence X is y_i. $s_k(y_i, x, i)$ is the status character function. It denotes that the label corresponding to the ith element in the observation sequence X is y_i. λ_k and u_k are the weights for the transfer character function and the status character function, respectively.

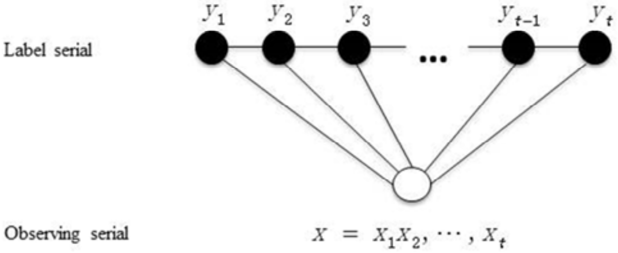


Figure 4. Undirected graph of conditional random fields

According to the principle of CRF, the process of extracting product feature from the results of word segmentation mainly contains two tasks: annotating train set and designing rule template.

4.1 Annotating Train Set

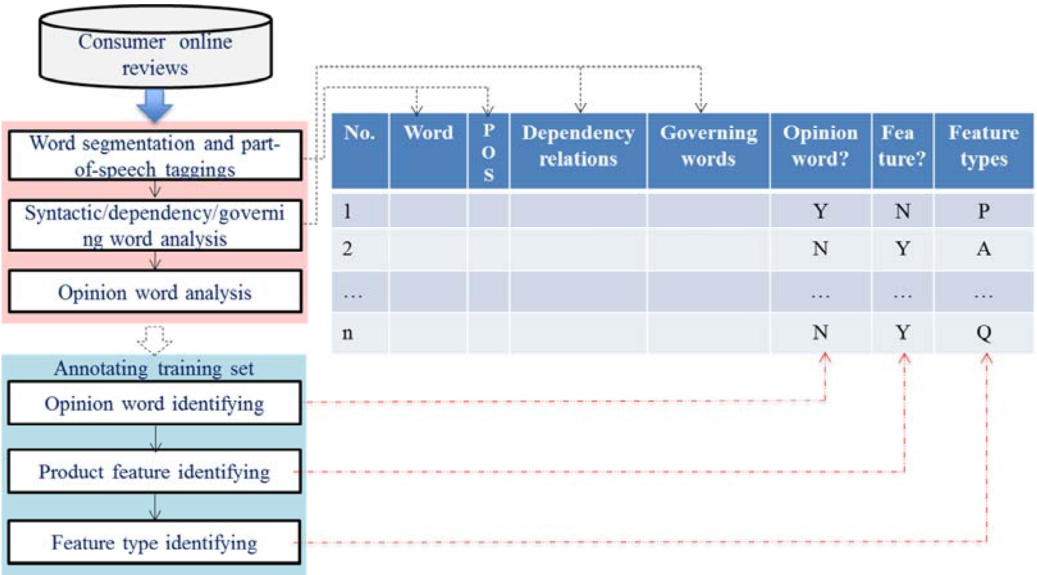
Annotating train set, based on the results of the preprocessing phase including POS tag, dependency relations, and governing words, is to identify the opinion words, product features and their types that is presented in **Figure 5**.

4.1.1 Opinion Word Identifying

For Chinese language, opinion words may also be other kinds of POSs, not just adjective. For example, Chinese phrase “可以(can)[6]” is a verb but it may express a positive opinion of consumer sometimes. This is one of the notable differences between Chinese language and English language. However, these phrases are usually not included in traditional opinion word set. This leads to the inaccuracy of the sentiment analysis for product features inevitably,

[6] The phrase “可以” expresses a positive opinion that means “good” in Chinese language. However, its literal meaning corresponds to the word “can” in English language. Therefore, the POS of it defines as verb at word segmanetation process.

especially for Chinese product features. In order to analyze Chinese product features effectively, it is necessary to identify this kind of opinion words. Table 1 presents these unusual opinion words (partial) based on the analysis for Chinese language at preprocessing phase. Using them, many nouns or noun phrases can be identified and evaluated. This is high significant for product feature extraction from Chinese consumer reviews and its sentiment analysis.


Consumer online reviews
Word segmentation and part-of-speech taggings
Syntactic/dependency/governing word analysis
Opinion word analysis
Annotating training set
Opinion word identifying
Product feature identifying
Feature type identifying

No.	Word	POS	Dependency relations	Governing words	Opinion word?	Feature?	Feature types
1					Y	N	P
2					N	Y	A
…					…	…	…
n					N	Y	Q

Figure 5. Annotating train set process

Table 1. Unusual opinion words at Chinese consumer reviews

No	POS	Phrases	Sentences
1	v	可以(can), *et al*.	“手机的分辨率还可以” (The resolution of this phone is good)
2	n	战斗机(fighter), *et al*.	“手机中的战斗机” (It is a **fighter** among phones)
…	…	…	…

4.1.2 Product Feature Identifying

Product feature identifying is a crucial step for supervised feature extraction method. It will affect the validity of product feature extraction directly. A reasonable size of train set is necessary to keep the accuracy of product feature extraction. Therefore, it is a time-consuming manual annotating process.

4.1.3 Feature Type Identifying

In general, the product features extracted from consumer reviews include contents and types. For example, some product features refer to the product, and some product features refer to the components/parts constituting this product while some product features refer to the attributes of the product or the components/parts. Furthermore, these attributes can be grouped into the function, performance, quality, and service and so on. Distinguishing these product features carefully can help designer, manufacturer, or retailer to insight into the correlation and influence characters among them. It provides evidences for deep comprehensive applications based on product features. Therefore, identifying feature type is very necessary.

Considering the types of product features and their classifications as well as the interrelations among them, a hierarchical structure for product features can be constructed which is presented in **Figure 6**. This hierarchical structure consists of two parts: basic product structure and the product features describing the attributes of the nodes in basic product structure such as function, quality, and (or) service. Basic product structure consists of root node (product), components, and parts which may also be extracted from consumer reviews and are product features. And the attributed product features including function, quality, and service are the expanded descriptions to corresponding product, component, or part. This product feature structure tree connects the attributed product features with corresponding product, component or parts. It is the foundation for implementing deep comprehensive applications based on product features.

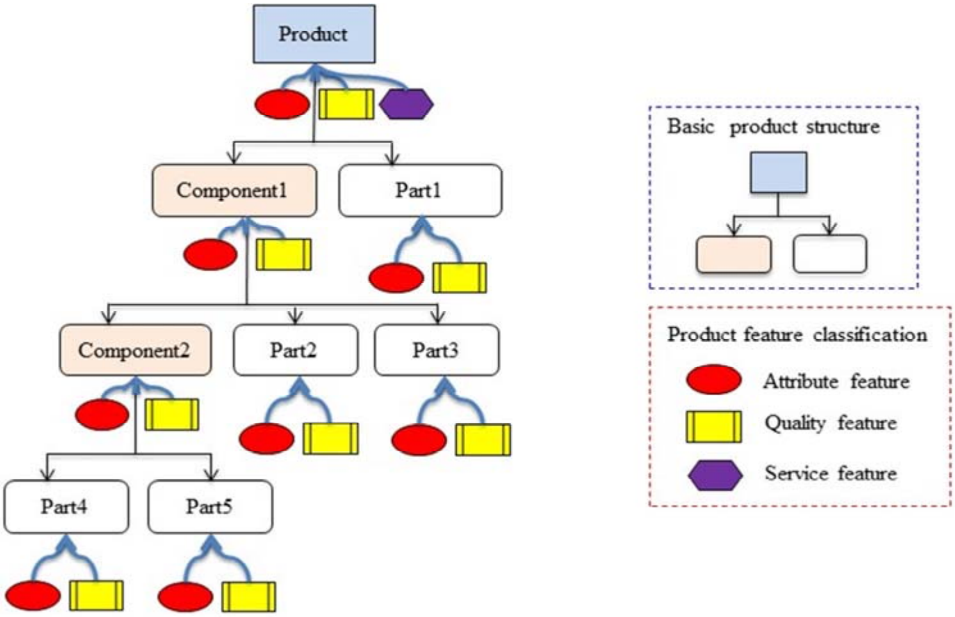


Figure 6. Product feature classifications and its hierarchical structure

4.2 Rule Template Designing

Product feature extraction based on CRF need a rule template to train its *model* which is the core module of CRF to guide product feature extraction process. According to the requirements of our research works, an approach of designing the rule template for Chinese product feature extraction is proposed. It mainly includes three aspects of works such as the core elements of rule template, the unit structure of rule template, and the organization form of rule template. Considering the characters of Chinese language, the core elements that consist of rule template are presented in **Table 2** which contains word elements (including phrase, POS, and context), syntactic elements (including dependency relations and governing words), and sentiment element (opinion words). Each element is also explained in detail in **Table 2**.

Table 2. Core elements of rule template and their explains

Elements	Contents	Explains
Word form elements	Phrase	Element denotes a phrase
	POS	Element denotes the POS of the current phrase
	Context (front or back)	Element denotes the phrases that locate at the front of the current phrase or at the back of the current phrase
Syntax elements	Dependency relation	Element denotes the dependency relation between the current phrase and its governing word
	Governing word	Element denotes the governing word that belong to the dependency relation between them
Opinion elements	Opinion words	Governing word is an opinion word or not

These elements describe the current phrase and the concerned information around it that are very useful to identify product feature. The utilization unit of these elements can be described as a three tuple $< p, \Omega, "T" >$ which is explained in **Figure 7**.

Where p denotes the position information of the elements. Ω denotes the content information of the element such as phrase (0), POS (1), dependency relation (2), governing word (3), and opinion word (4). And T denotes the value corresponding to the element that is determined based on p and Ω. For example, the unit [1, 1,”n”] means that the POS of the phrase that is next to the current phrase is a noun. Using this mode, we can design the contents at a given position to deal with the various expression forms of Chinese language. In practice, the elements in **Table 2** are always combined when establishing the rule template to increase the accuracy and efficiency of extracting product features. The combination forms of elements and its implications are presented in **Figure 8**.

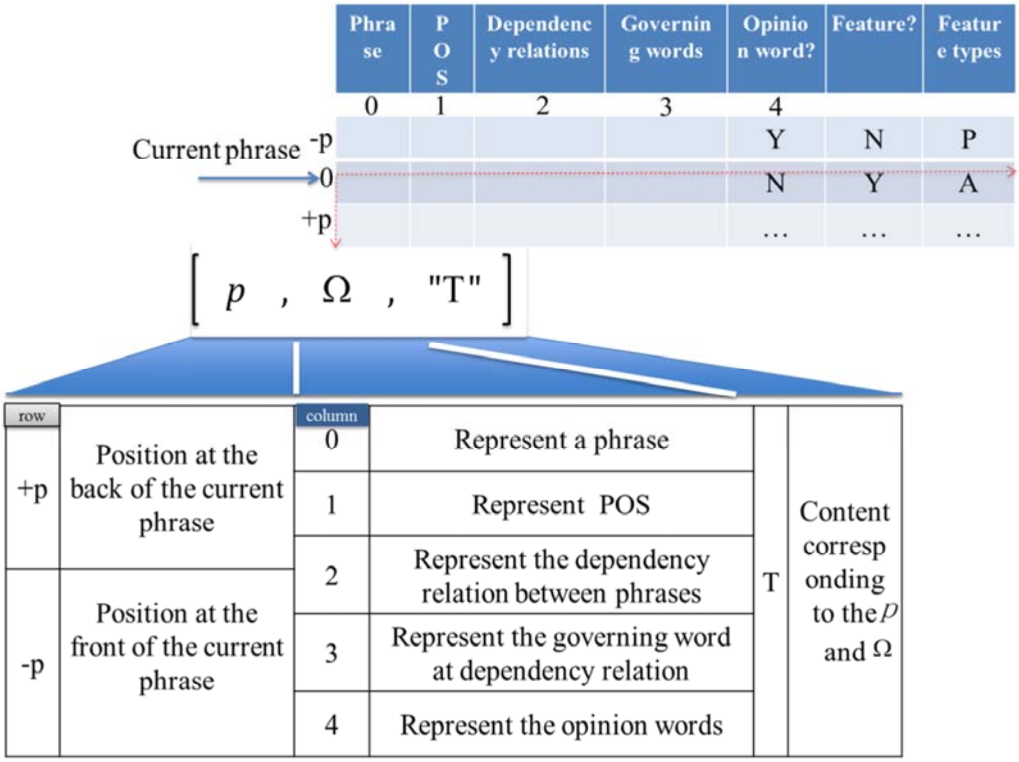

Phrase
POS
Dependency relations
Governing words
Opinion word?
Feature?
Feature types
0
1
2
3
4
Current phrase
-p
0
+p
Y
N
P
N
Y
A
…
…
…
[p , Ω , "T"]
row
+p
Position at the back of the current phrase
-p
Position at the front of the current phrase
column
0
Represent a phrase
1
Represent POS
2
Represent the dependency relation between phrases
3
Represent the governing word at dependency relation
4
Represent the opinion words
T
Content corresponding to the p and Ω

Figure 7. Unit structure and its explains

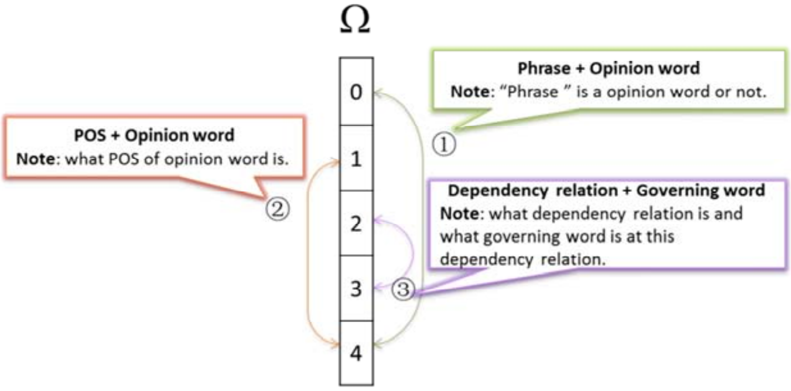

Ω
0
1
2
3
4
Phrase + Opinion word
Note: "Phrase " is a opinion word or not.
①
POS + Opinion word
Note: what POS of opinion word is.
②
Dependency relation + Governing word
Note: what dependency relation is and what governing word is at this dependency relation.
③

Figure 8. Combination forms of the elements and its implication

There are mainly three kinds of combination forms in this work namely ***phrase*** + ***opinion word***, ***POS*** + ***opinion word***, and ***dependency relation*** + ***governing word***. The combination "***phrase*** + ***opinion word***" describes whether the current phrase is an opinion word or not. The combination "***POS*** +***opinion word***" describes what is the POS of the opinion word. And the combination "***dependency relation*** + ***governing word***" describes the dependency relation between two phrases and what is the governing word of this dependency relation. These combination utilizations of the elements, together with their sole utilizations, form a complex architecture of rule template which is illustrated in **Figure 9**. Based on it, product feature extraction for specific task can be achieved well.

$$\begin{array}{l}
\cdots \\
[\; p_k \;,\; \Omega_i \;,\; \text{"T"} \;] \\
\cdots \\
[\; p_k \;,\; \Omega_i \;,\; \text{"T"} \;]\cdots[\; p_{k+h} \;,\; \Omega_i \;,\; \text{"T"} \;]\cdots \\
\cdots \\
[\; p_k \;,\; \Omega_i \;,\; \text{"T"} \;]\cdots[\; p_{k+h'} \;,\; \Omega_{i'} \;,\; \text{"T"} \;]\cdots \\
\cdots
\end{array}$$

$$\begin{array}{l}
s.t. \\
i,i' \in \{0,1,2,3,4\} \quad and \quad i \neq i' \\
k,h,h' = const.
\end{array}$$

Figure 9. General organization form of rule template

Based on train set and rule template, the models of CRF can be established through in-depth learning process which is presented in **Figure 10**. This learning process constructs a large amount of function sets which will be used in models to calculate the conditional probability of elements co-occurring with the form of rule unit description at consumer reviews. Then these results are used to calculate the probabilities at the transfer character and those of the state character in **Equation (7)**, respectively.

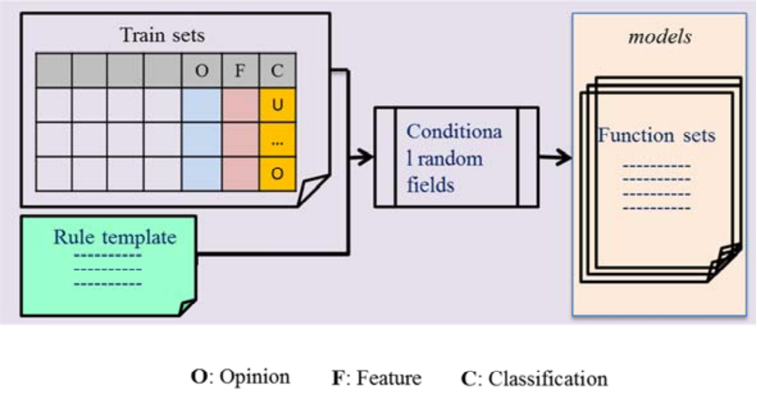


Figure 10. Training the models of CRF

5. Quantitative Characters of Product Features

Quantitative description is the foundation of analyzing product features precisely. In this work, the quantitative characters of product features are investigated from two aspects: the frequency of product feature and the sentiment score of product feature which reflect the extent of consumer paying attention to them, and the positive or negative feeling of consumer for them, respectively.

5.1 Frequency of Product Feature Occurring at Consumer Reviews

The frequency of product feature occurring at consumer reviews reflects the extent of customer paying attention to it. For example, consumer maybe like a product feature very much or disappoint very much when the frequency of it is very high. The frequency of a product feature occurring at consumer reviews is generalized as follows:

$$Num\_F_i = \sum_s^{n_s} k_{is} \tag{8}$$

where n_s denotes the number of all consumer reviews. k_{is} denotes the number of the ith product feature appearing at the sth consumer review. Num_F_i denotes the frequency of the ith product feature occurring at consumer reviews.

5.2 Sentiment Score of Product Feature

Generally, the evaluation of consumers to a product feature is either positive or negative, and its strength is different as well. How to describe this kind of distinguishes and how to measure its strength are very important to insight into the preference of consumers precisely.

After analyzing 3,000 consumer reviews, we find that the language pattern of consumer evaluating a product feature is mainly manifested as follows:

$$(adv_., adj_., pf) \tag{9}$$

where pf denotes a product feature. $adj.$ denotes the adjective that modifies product feature pf. And $adv_.$ denotes the adverb that modifies the adjective $adj.$. The adverb $adv_.$ and the adjective $adj.$ modify the product feature pf together.

The adverb $adv_.$ and the adjective $adj.$ that modify the product features are always qualitative descriptions at consumer reviews. In order to describe the strengths of these adjectives $adj.$ and their polarity as well as those of adverb $adv_.$ for the goal of calculation and comparison, the adjective $adj.$ and the adverb $adv_.$ should be transformed to numerical value according to their strength and polarity. In this work, the adjective $adj.$ is defined as the range [-9, +9], and the adverb $adv_.$ is also defined as the range [-9, +9]. From 1 to 9, strength is increasing gradually. And the minus sign denotes opposite polarity (namely negative). Then, the sentiment score of the ith product feature pf is generalized as follows:

$$\begin{aligned} Sco_{F_i} &= \frac{1}{T}\left\{Sco_{F_{iP}} + Sco_{F_{iN}} + Sco_{F_{iM}}\right\} \\ &= \frac{1}{T}\{\textstyle\sum_{x=1}^{a}(Strong_{Px} + Strong_{PxA}) - \sum_{y=1}^{b}(Strong_{Ny} + Strong_{NyA}) + \\ &\textstyle\sum_{z=1}^{c}[\sum_{z1=1}^{pz}(Strong_{Mz_{Pz1}} + Strong_{Mz_{Pz1A}}) - \\ &\textstyle\sum_{z2=1}^{nz}(Strong_{Mz_{Nz2}} + Strong\_Mz\_Nz2A)]\} \end{aligned} \tag{10}$$

$$T = a + b + \textstyle\sum_{z=1}^{c}(pz + nz) \tag{11}$$

where Sco_F_i denotes the sentiment score of the ith product feature F_i. a, b, and c denote the number of the positive consumer reviews concerned with the ith product feature F_i, the number of the negative consumer reviews concerned with the ith product feature F_i, and the number of the neutral consumer reviews (the consumer review that has multiple different polarity opinion words is defined as neutral consumer review in this work because it is difficult to identify its exact polarity) concerned with the ith product feature F_i, respectively. $Strong\_Px$ denotes the score of the adjective nearby the ith product feature F_i at the xth positive consumer review. And $Strong\_PxA$ denotes the strength of the adverb that modifies the nearest adjective at the xth positive consumer review. $Strong\_Ny$ denotes the sentiment score of the adjective nearby the ith product feature F_i at the yth negative consumer review. $Strong\_NyA$ denotes the strength of the adverb that modifies the nearest adjective at the yth negative consumer review. pz is the number of the positive adjective that correspond to product feature F_i at the zth neutral consumer review, and nz is the number of the negative adjective that correspond to product feature F_i at the zth neutral consumer review. $Strong\_Mz\_Pz1$ denotes the sentiment score of the $z1$th positive adjective of the zth neutral consumer review, and $Strong\_Mz\_Pz1A$ denotes the strength of the adverb that modifies the $z1$th positive adjective at the zth neutral consumer review. Likeness, $Strong\_Mz\_Nz2$ denotes the sentiment score of the $z2$th negative adjective of the zth neutral consumer reviews, and $Strong\_Mz\_Nz2A$ denotes the strength of the adverb that modifies the $z2$th negative adjective at the zth neutral consumer review.

The sentiment score reflects the preference of consumers to a product feature and its extent comprehensively. It can provide the evidences for retailer, designer, or manufacturer to precisely implement product improvement, and market strategy *et al*.

6. Product Feature Structure Tree Constructing

Product features that correspond to the attributes of the product, components, or parts should be connected with relevant objects (namely product, components, or parts) in order to implement in-depth analysis and comprehensive applications at product features level. According to the classifications of product features, product features form a tree structure in general which is presented in **Figure 6**, namely product feature structure tree. In order to construct this product feature structure tree, a basic product structure is employed which is an existing product structure and used as frame, and the nodes of it are also the potential parent nodes for attributed product features. It needs to be noted that the nodes of the basic product structure should also the product features extracted from consumer reviews. Therefore, the key effort of constructing product feature structure tree is to find corresponding parent nodes for each attributed product feature from the results of word segmentation, and compare the corresponding parent nodes with the nodes of basic product structure.

6.1 Finding Potential Parent Node for Current Product Feature

The potential parent nodes of current product features are always the parts, components, or even product. They are noun phrases. And they always co-exist with these attributed product features. Besides, considering the expression habits of Chinese consumer reviews e.g. some consumers may mention the parts or product first when they comment an object, and then evaluate its attributes for example "照相机的像素太低(The pixels of the camera is too poor)" while some other consumers may evaluate the attributes of the parts or product first, and then mention the parts or product for example "续航时间长(long battery life)，电池杠杠的(the battery very good)". Thus, keeping the current product feature pf_i as a central point, the process of finding potential parent node for current product feature pf_i is to search the phrase that satisfies with specific POS and type requirements, or the dependency relation with current product feature pf_i based on given step-length from left and right direction illustrated as **Figure 11**. The pseudo-code description for the algorithm of finding potential parent node for current product feature is presented in **Figure 12**.

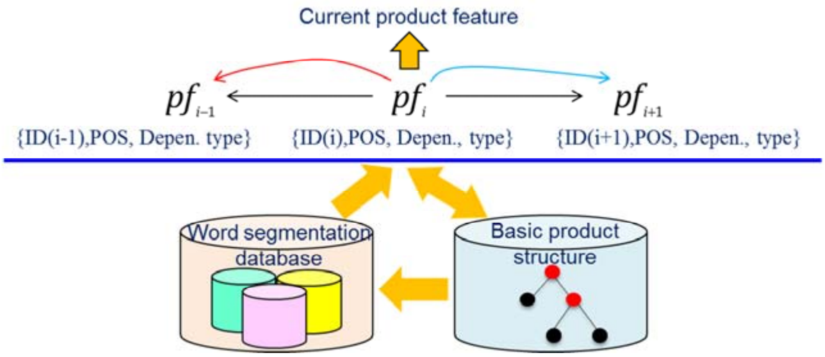
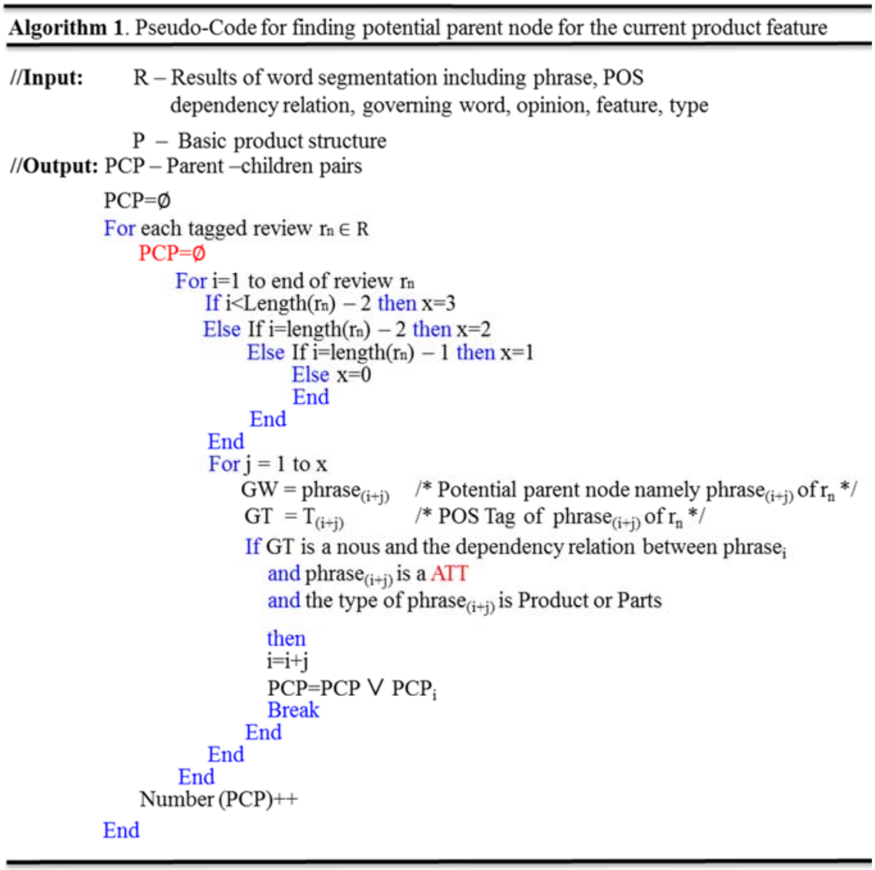

Current product feature
pf_{i-1}
pf_i
pf_{i+1}
{ID(i-1),POS, Depen. type}
{ID(i),POS, Depen., type}
{ID(i+1),POS, Depen., type}
Word segmentation database
Basic product structure

Figure 11. Principle of finding the parent nodes for current product features



Algorithm 1. Pseudo-Code for finding potential parent node for the current product feature

```
//Input:    R – Results of word segmentation including phrase, POS
                dependency relation, governing word, opinion, feature, type
            P  –  Basic product structure
//Output: PCP – Parent –children pairs
            PCP=Ø
            For each tagged review $r_n \in R$
                PCP=Ø
                    For i=1 to end of review $r_n$
                        If i<Length($r_n$) – 2 then x=3
                        Else If i=length($r_n$) – 2 then x=2
                            Else If i=length($r_n$) – 1 then x=1
                                Else x=0
                                End
                            End
                        End
                        For j = 1 to x
                            GW = $phrase_{(i+j)}$    /* Potential parent node namely $phrase_{(i+j)}$ of $r_n$ */
                            GT  = $T_{(i+j)}$          /* POS Tag of $phrase_{(i+j)}$ of $r_n$ */
                            If GT is a nous and the dependency relation between $phrase_i$
                              and $phrase_{(i+j)}$ is a ATT
                              and the type of $phrase_{(i+j)}$ is Product or Parts

                              then
                              i=i+j
                              PCP=PCP ∨ $PCP_i$
                              Break
                            End
                        End
                    End
                Number (PCP)++
            End
```

Figure 12. Pseudo-code of finding potential parent node

6.2 Similarity between Potential Parent Nodes and the Nodes of Basic Product Structure

It is necessary to confirm whether a potential parent node of the current attributed product feature exists at basic product structure or not before adding the attributed product features into basic product structure. Comparing the similarity between the potential parent nodes and the nodes of basic product structure is a valid measure. Considering the characters of Chinese language, the similarity between the potential parent nodes and the nodes of basic product structure is calculated from two aspects: literal similarity and context similarity.

6.2.1 Literal Similarity

Word is the basic unit of constructing a phrase. For Chinese language, many phrases whose meanings are similar always contain the same words (Xia, 2007). Based on these facts, the similarity between potential parent nodes and the nodes of basic product structure can be calculated through the status of words appearing at these nodes (product features) namely literal similarity which is influenced by two factors: quantitative and position (Wang, Zhou & Sun, 2012).

Let pf_A and pf_B are two product features that the similarity between them need to be calculated. The literal similarity $LitSim(pf_A, pf_B)$ between pf_A and pf_B is generalized as follows (Xia, 2007; Wang *et al.*, 2012):

$$LitSim(pf_A, pf_B) = \alpha \times (\frac{|SameHZ(pf_A,pf_B)|}{|pf_A|} + \frac{|SameHZ(pf_A,pf_B)|}{|pf_B|})/2$$

$$+\beta \times d_p \times (\frac{\sum_{i=1}^{|pf_A|} Weight(pf_A,i)}{\sum_{i=1}^{|pf_A|} i} + \frac{\sum_{j=1}^{|pf_B|} Weight(pf_B,j)}{\sum_{j=1}^{|pf_B|} j})/2 \quad (12)$$

and $0 \leq LitSim(pf_A, pf_B) \leq 1$.

where α and β are the weights that describe the importance of quantitative factor at the literal similarity calculation and the importance of position factor at the literal similarity calculation respectively, and $\alpha + \beta = 1$. In addition, d_p defines the ratio of the number of words at these two product features.

$$d_p = \min\{\frac{|pf_A|}{|pf_B|}, \frac{|pf_B|}{|pf_A|}\}$$

$Weight(pf_A, i)$ denotes the weight of the ith word of the product feature pf_A.

$$Weight(pf_A, i) = \begin{cases} i, \text{ if } pf_A(i) \text{ at } SameHZ(pf_A, pf_B) \\ 0, \ others \end{cases}$$

where $|pf_A|$ and $|pf_B|$ denote the number of words at product feature pf_A and product feature pf_B , respectively. $pf_A(i)$ denotes the i th word of product feature pf_A . $SameHZ(pf_A, pf_B)$ denotes the set of the words that are contained in both product feature pf_A and product feature pf_B at the same time. $|SameHZ(pf_A, pf_B)|$ is the number of the set $SameHZ(pf_A, pf_B)$.

6.2.2 Context Similarity

In addition, some Chinese phrases are similar at sematic but they don’t contain any the same words such as “外观(appearance)” and “样子(shape)”. In order to calculate the similarity of these kinds of product features, it should make full use of the context information around these product features because the phrases which modify the same sematic phrases are always similar (Tu, Zhang, Zhou & He, 2012). Thus, the similarity calculation between product features based on context can be generalized as follows:

Each product feature pf_i is described as a n dimensional vector.

$$pf_i = (S_{i1}, S_{i2}, \cdots, S_{ij}, \cdots, S_{in}) \tag{13}$$

where S_{ij} is the co-occurrence frequency between product feature pf_i and the jth modified phrase.

Thereupon, the similarity calculation among product features is transformed into the similarity between two vectors. It is generalized as follows (Tu *et al.*, 2012):

$$Sim(pf_a, pf_b) = \frac{\sum_{k=1}^{n} S_{ak} \times S_{bk}}{\sqrt{\sum_{k=1}^{n} S_{ak}^2 \sum_{k=1}^{2} S_{bk}^2}} \tag{14}$$

where S_{ak} denotes the co-occurrence frequency between product feature pf_a and the kth modification phrase. S_{bk} denotes the occurrence frequency of product feature pf_b and the kth modification phrase. n is the total number of the modification phrases in an existing group. And $Sim(pf_a, pf_b)$ is the similarity between product feature pf_a and product feature pf_b.

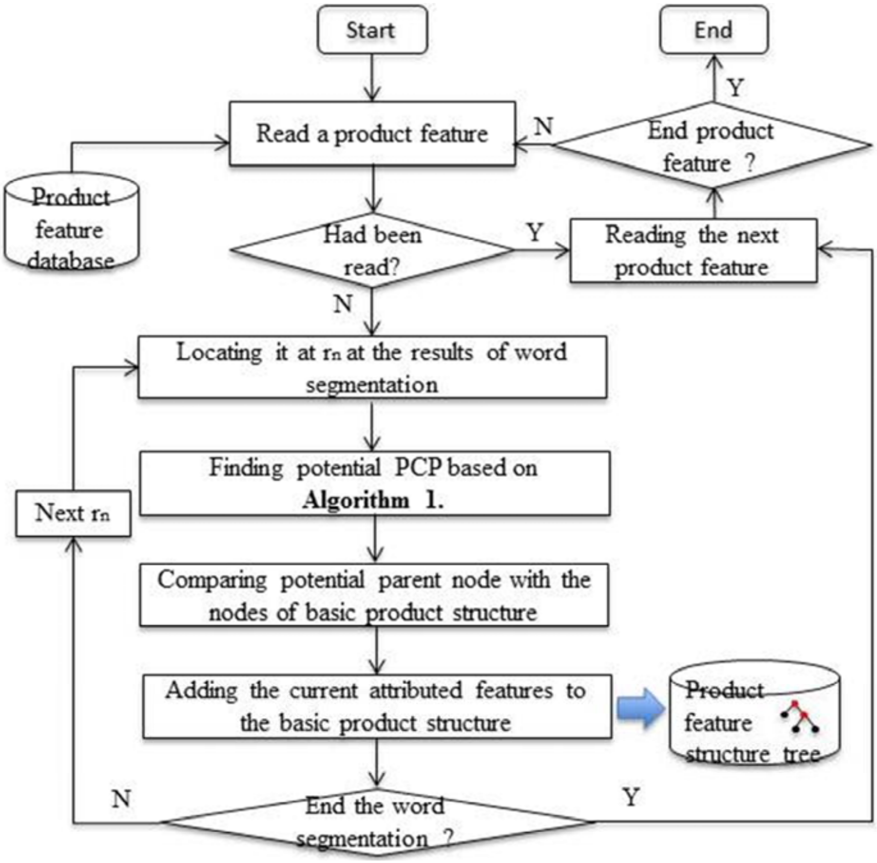

Start
End
Y
Read a product feature
N
End product feature ?
Product feature database
Had been read?
Y
Reading the next product feature
N
Locating it at rn at the results of word segmentation
Finding potential PCP based on Algorithm 1.
Next rn
Comparing potential parent node with the nodes of basic product structure
Adding the current attributed features to the basic product structure
Product feature structure tree
N
Y
End the word segmentation ?

Figure 13. Process of constructing product feature structure tree

Based on potential parent node searching and similarity calculating, the process of constructing product feature structure tree is presented in **Figure 13**. First, picking out a product feature from product feature database, and locating it at the results of word segmentation e.g. the r_nth consumer reviews. Second, searching the potential parent-child pairs (PCP) by calling **Algorithm 1**, and then comparing the parent nodes of potential PCP with the nodes of basic product structure based on similarity analysis. If exists, adding the product features (namely attributed product feature) into the corresponding nodes of basic product structure as its children. Repeating this process, until all the attributed product features are added into basic product structure. This process connects not only the attributed product features but also their quantitative descriptions such as frequency and sentiment score with their parent nodes.

7. Experimental Analysis

Product feature extraction from Chinese consumer reviews is a complicated task and is also a crucial task because its results influence the efficiency of similarity analysis and comprehensive applications directly.

Many factors influence the results of product feature extraction. In order to insight into these factors and provide evidences to control the process of product feature extraction effectively, we design extended experiments from different perspectives based on 5,806 Chinese consumer reviews retrieved from e-commerce platforms ***Taobao.com***, ***Suning.com***, and ***Zhongguancun.com***.

7.1 Results of Word Segmentation

The results of word segmentation provide the data resources for product feature extraction and product feature structure tree constructing. Therefore, a valid word segmentation should keep enough correctness. In this work, a two-stage optimizing word segmentation process is proposed which is presented in **Figure 3**. In order to show the effectiveness and necessity of two-stage optimizing word segmentation process, we designed two experiments: the word segmentation based on tool *ictcals* only and the word segmentation based on our proposed two-stage optimizing word segmentation method. And then the correct rate, which is defined as the ratio between the number of correct word segmentation and the number of total word segmentation result, is used as index to evaluate the effectiveness of different word segmentation methods and different data sources such as ***taobao.com***, ***suning.com***, and ***zhongguancun.com***, respectively. These results are presented in **Figure 14**. Black rectangles describe the correct rates of product features that are extracted based on *ictcals* system only from ***taobao.com***, ***suning.com***, and ***zhongguancun.com*** respectively (*taobao*:90.16%, *suning*:90.5%, and *zhongguancun*:95.29%.). Red rectangles describe that correct rates of product features that are extracted based on our two-stage optimizing word segmentation from ***taobao.com***, ***suning.com***, and ***zhongguancun.com*** respectively (*taobao*:98.39%, *suning*:95.97%, and *zhongguancun*:97.65%.). Obviously, the correct ratios of red rectangle are all higher than those of black rectangle.

Furthermore, we also calculate the average correct rate of word segmentation based on the total data from ***taobao.com***, ***suning.com***, and ***zhongguancun.com*** which is illustrated in **Figure 15**. The correct rate is also increased by 6.16%. Therefore, it is very necessary to implement two-stages optimizing word segmentation in order to increase the correctness of Chinese consumer reviews and provide valid data sources for product feature extraction.

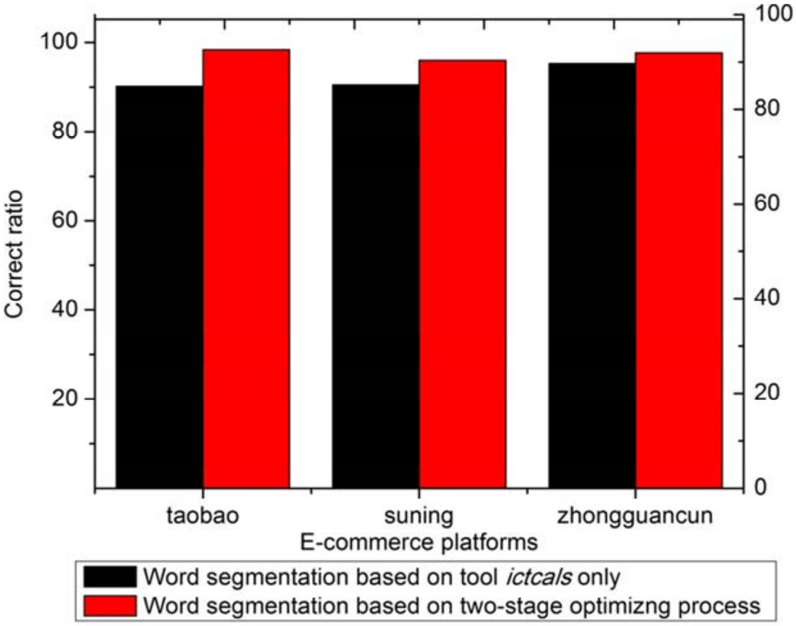


Figure 14. Correct rates of two word segmentation methods for three data sources

7.2 Contents of Rule Template

The elements of rule template and its organization form determine the solution of extracting product features. Different rule templates will lead to different effectiveness of product feature extraction. In order to explore a valid rule template including elements and its organization form for our work, 10 rule templates that are developed based on different elements which are presented at **Table 2** and **Figure 7** and organization forms are designed. The efficiency of product feature extraction based on these 10 rule templates are evaluated respectively based on existing popular index such as precision, recall, and *F*-score which is illustrated in **Figure 16.**

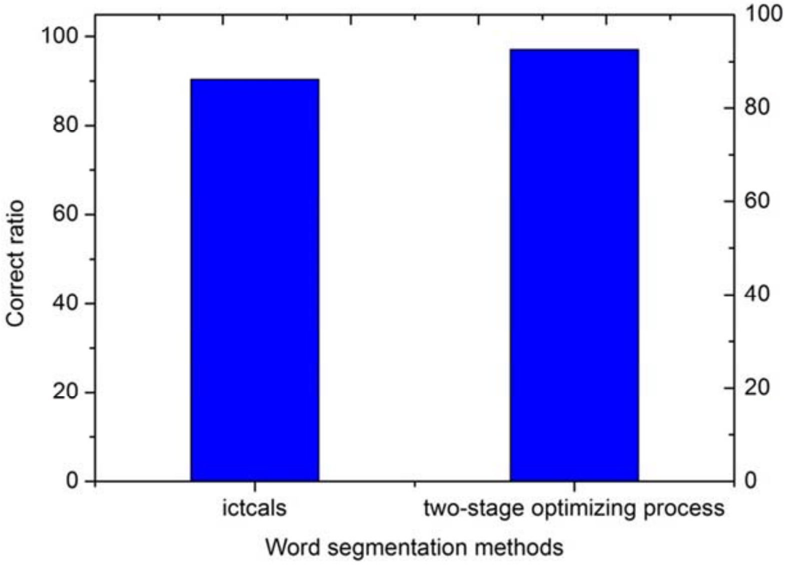


Figure 15. Correct rates of two word segmentation methods for total data

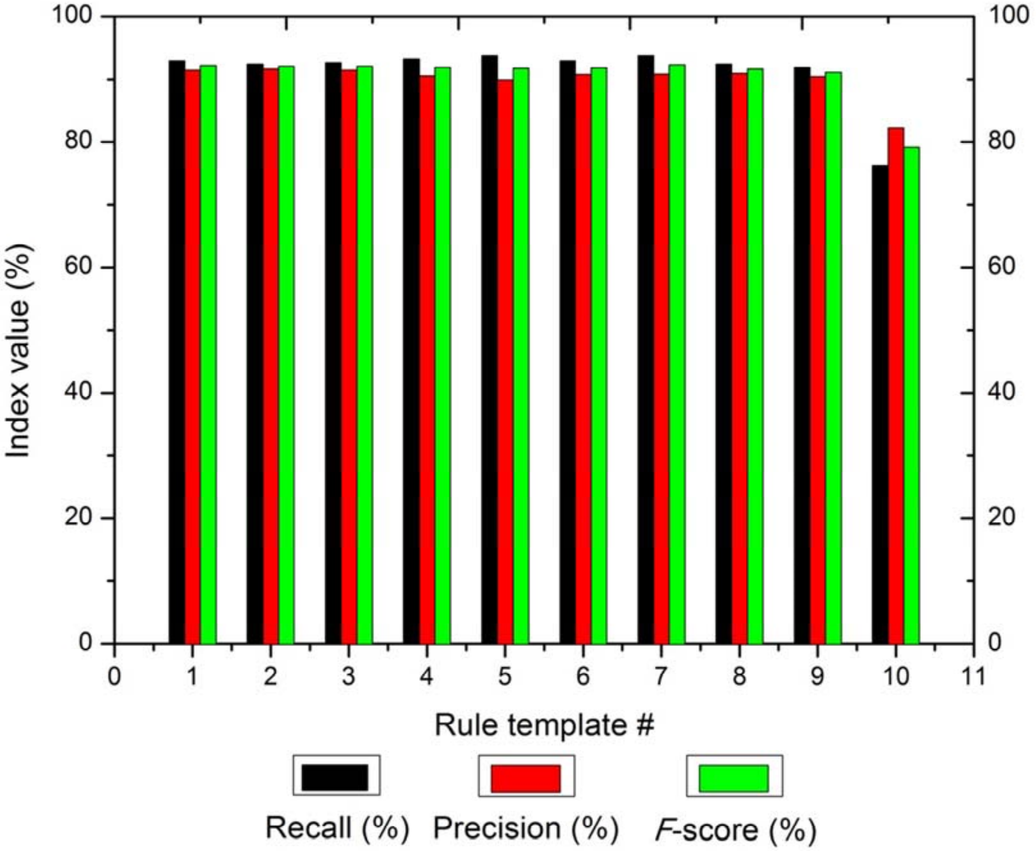


Figure 16. Precisions, recall, and F-score of product feature extraction based on different rule templates

We found that the precision, recall, and *F*-score corresponding to the 7th rule template are 90.86%, 93.8%, and 92.31%, respectively. These are comprehensive optimal comparing with those of product feature extraction processes based on the other 9 rule templates. Thus, the rule template for CRF in this work will be established according to the 7th rule template.

7.3 Widths of Searching Window

Consumer reviews are always irregular expression because the purpose of consumer commenting on products at network platform is to exchange and share information. Especially for Chinese language, its complex syntax, grammar and diversified expressions make it more serious. Therefore, a proper search range is very important in order to find the valid phrases which are correlated with the current object.

With these considerations, three widths of searching window which had been described in **Figure 11** are designed such as 3, 5, and 7 respectively to extract the potential parent nodes for current product features. We also employ precision, recall and *F*-score to measure the effectiveness of finding potential parent node at different widths of searching window, and the results of them are presented in **Figure 17**. It can be seen that the comprehensive result is the optimal when the width of searching window is 5 although the recall of it increases continuously along with the increasing of width. The precision and *F*-score will be decreased once expanding the width of searching window when the potential parent node cannot be found at given range. The reason is that the phrases that are found at expanding range maybe satisfy with the constraint conditions defined at our searching algorithms such as POS or rules, it may not correlate with the current product feature at all. Thus, it decreases the precision and the *F*-score in the end.

Considering the expression habits of Chinese language and the irregularity of consumer review, the potential parent nodes of the current product features are always omitted or implicated. Therefore, they cannot be found directly under these conditions. In order to deal with this issue, a workflow of identifying the potential parent nodes for this kind of product features is presented in **Figure 18**. It is to infer the potential parent node for the current product feature according to the existing searching results namely the potential parent nodes for the same product feature at the front of consumer reviews. If the infer results are null, then the design manual for the target product which records the correlations between components/parts and their attributes is used as evidences to identify its parent node. It avoids to searching at wider range aimlessly and keep the effectiveness of searching process as well.

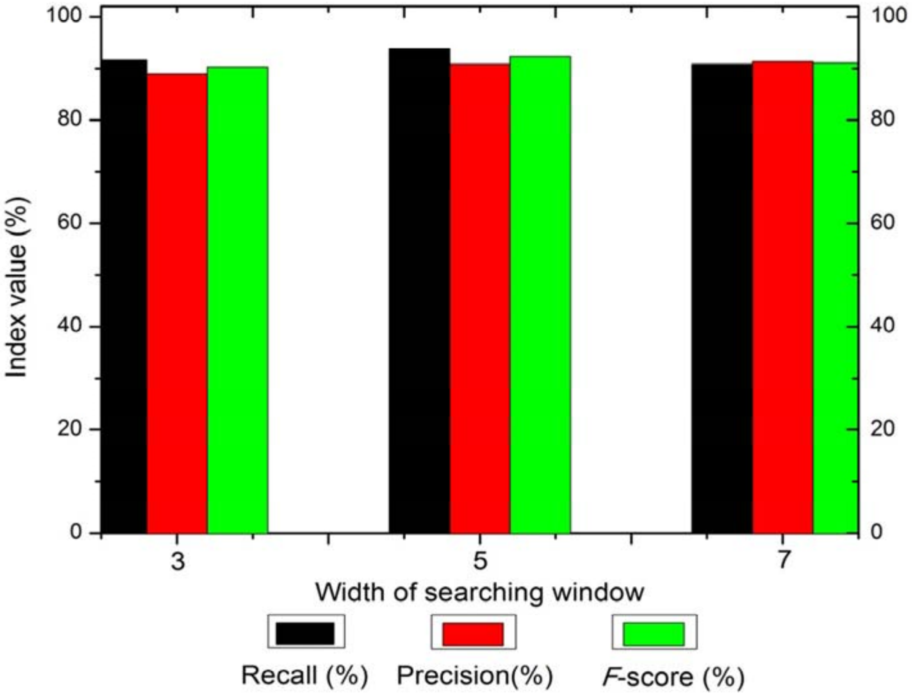
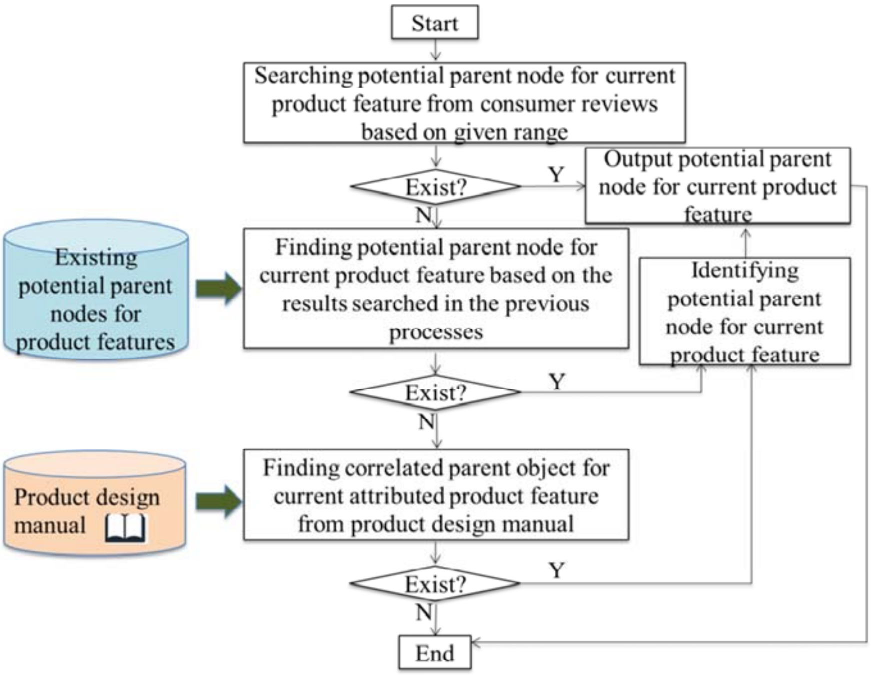


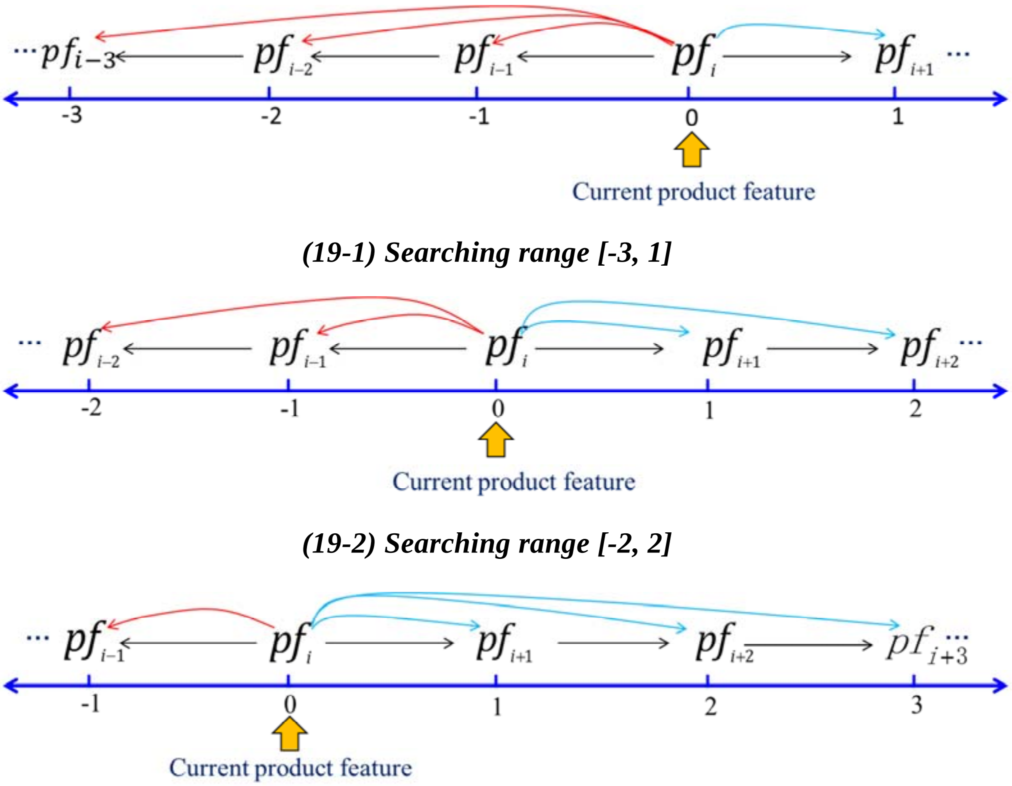
Figure 17. Precision, recall, and F-score under different widths of searching window



Start
Searching potential parent node for current product feature from consumer reviews based on given range
Exist?
Y
Output potential parent node for current product feature
N
Existing potential parent nodes for product features
Finding potential parent node for current product feature based on the results searched in the previous processes
Identifying potential parent node for current product feature
Exist?
Y
N
Product design manual
Finding correlated parent object for current attributed product feature from product design manual
Exist?
Y
N
End

Figure 18. Workflow of identifying the implicit parent nodes for some product features

Moreover, the exact coverage regions of searching window may also be different even for the same width of searching window. Using the width 5 of searching window as example, three forms of coverage regions are presented in **Figure 19**. Accordingly, the efficiencies of searching potential parent nodes are evaluated through precision, recall, and *F*-score which are presented in **Figure 20**. The form of coverage region in **Figure (19-2)** corresponds to a better result. Therefore, the practical searching range and its coverage region are set based on this result in the case study.


··· pf_{i-3} pf_{i-2} pf_{i-1} pf_i pf_{i+1} ···
-3 -2 -1 0 1
Current product feature
··· pf_{i-2} pf_{i-1} pf_i pf_{i+1} pf_{i+2} ···
-2 -1 0 1 2
Current product feature
··· pf_{i-1} pf_i pf_{i+1} pf_{i+2} pf_{i+3} ···
-1 0 1 2 3
Current product feature

(19-1) Searching range [-3, 1]

(19-2) Searching range [-2, 2]

(19-3) Searching range [-1, 3]

Figure 19. Different coverage forms of searching window

These experiments and their results provide the evidences for our research works at word segmentation, product feature extraction, and product feature structure tree constructing. They are very significant for keeping the validity of our proposed methods.

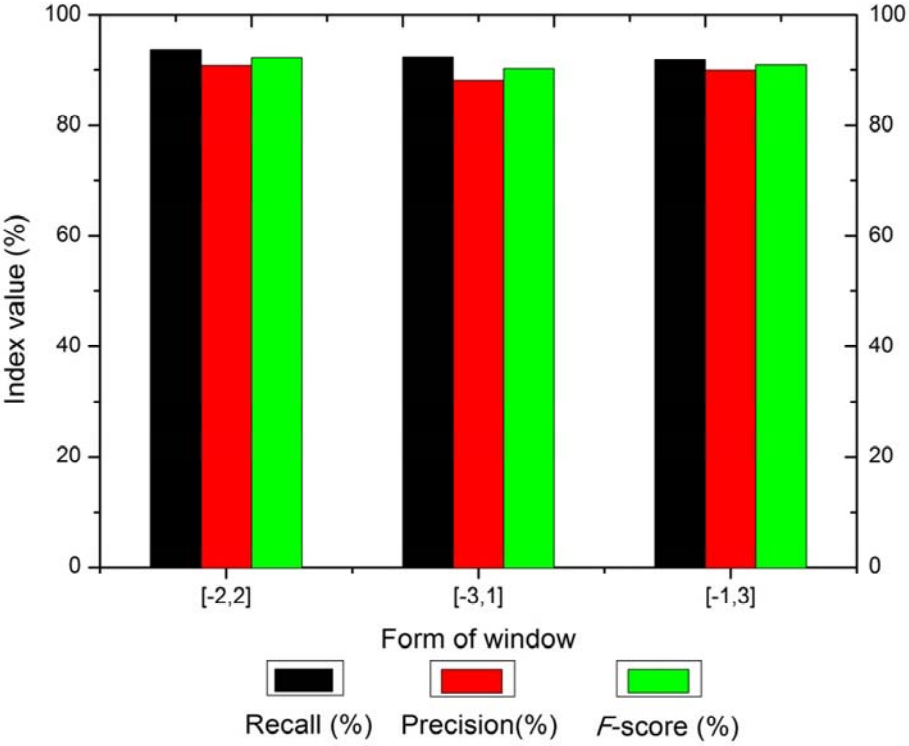


Figure 20. Efficiency of searching potential parent nodes under different forms of coverage regions

8. Case Study

Consumer reviews contain rich information regarding consumer requirements and preferences. Mining valuable information effectively from consumer reviews can provide evidences for designers, manufacturers, or retailers to implement product improvement or make market strategy. With the rapid expansion of e-commerce businesses based on network platforms and clients, more and more companies have realized the importance of this kinds of utilities. Aiming at Chinese consumer reviews, this section, using intelligent mobile phone xx-F2 as example, is to elaborate the implementations of the principles and methods mentioned above, and the applications based on product features.

By using web crawler tools *Goseeker* and *Train collector*, we retrieved 5,806 Chinese consumer reviews from e-commerce platform ***taobao.com*** (2,591), ***suning.com*** (1,243), and ***zhongguancun.com*** (1,972) which are used as analysis corpuses. According to the technique framework presented in **Figure 1**, we employ software *ictclas*, which is developed by Chinese Science Academic, as word segmentation tool to divide consumer reviews into discrete phrases and label their POS. At the same time, we employ software *ltp*, which is developed by

Harbin Institute of Technology of China to achieve syntactic parsing. And 82,724 raw phrases are obtained. After preprocessing for these raw phrases such as stop words, typos, and meaningless phrases, a two-stages optimization word segmentation process is performed presented at **Section 3.2** to make the results of word segmentation more suitable for our research tasks, and the key parameters of these optimization phases are set presented in **Table 3**. Finally, 50,785 valid phrases are obtained. These phrases are used as the data resources (corpus) for product feature extraction.

Table 3. Parameter setting of two-stages optimizing word segmentation

No	Parameters	Explains	Setting
1	n	The length of reconstructed string	n=5
2	f	The parameter of frequency filtering	f>2
3	c	The parameter of cohesive filtering	c>0.2
4	r	The parameter of left and right entropy filtering	r>0.8
5	q	The number of relation Fi-sematic-Fj occurring at the reviews	q>3

In order to extract product features from these results of word segmentation effectively based on CRF, 9,081 phrases obtained from 1,000 consumer reviews are used as train set. We invited 2 engineers from mobile phone development department and 1 linguist from the literature of our school to annotate these phrases manually including feature, type, and opinion. It took two days, 8 hours per day to implement this task. At the same time, rule template is developed according to the analysis results of experiment at **Section 7.2**.

Based on train set and rule template, the *model* of CRF is trained through a machine learning process. And then, product features for xx-F2 product are extracted from 50,785 valid phrases of 5,806 consumer reviews based on CRF. Finally, 80 product features are obtained after merging synonym, homoionym, and alternative names.

Product feature extraction is a very crucial step for ensuring the effectiveness of the next comprehensive analysis and application based on product features. In order to verify the validity of our proposed methods, we design a five-fold intersection experiments by using 5,000 phrases from the results of word segmentation. These 5,000 phrases are divided into 5 subsets which are labeled 1, 2, 3, 4, and 5 respectively, and each subset contains 1,000 phrases. The effectiveness of product feature extraction based on five-fold intersection experiments are measured through indexes precision, recall, F-score. At the same time, we calculated the precision, recall, and F-score of product feature extraction based on the methods proposed by Jakob's work, which is the closest with our works at the aspect of product feature extraction, by using the same phrase set. Finally, we compare the results obtained based on our methods with those obtained based on the methods of Jakob's work (Jakob & Gurevych, 2010) which are presented in **Table 4**. Obviously, the precision, recall, and F-score of our methods are all better than those of Jakon's work. It denotes that our methods of extracting product features from consumer reviews are valid, especially for Chinese consumer reviews.

Table 4. Experiment result comparision between our methods and Jakob's work

Precision(%)		Recall(%)		*F*-score(%)	
Our methods	Jakob's work	Our methods	Jakob's work	Our methods	Jakob's work
93.80	86.47	90.86	78.70	92.31	79.63

Based on the product features extracted above and the results of word segmentation, the frequency of each product feature is calculated, so does the sentiment score of it. And the potential parent nodes of the attributed product features are identified based on the **Algorithm 1** presented in **Figure 12** and the workflow presented in **Figure 18**. As a result, the attributed product features are added into the basic product structure of product xx-F2. Thus, product feature structure tree for xx-F2 is established which is illustrated in **Figure 21**. The unit of product feature structure tree is a four tuple: < F_i, frequency, score, F_j > where F_i is parent node and F_j is child node. Frequency denotes the times of product feature F_j appearing at consumer reviews. Score denotes the sentiment evaluation of consumer to product feature F_j. Based on the product feature structure tree and the data on it including frequency and sentiment score, the influence or interaction relations between the parent nodes of product feature structure tree and its child nodes can be inferred conveniently.

In this work, a Bayes theory based application is investigated based on product feature structure tree that is to infer the factors (namely child nodes) of leading to the negative valuations or low sentiment scores of their parent nodes. The mathematic description of this inferring process is as following:

For a Bayes network which is concerned on a set of variables $\boldsymbol{X} = \{\boldsymbol{X_1}, \boldsymbol{X_2}, \cdots, \boldsymbol{X_n}\}$, it contains two aspects: ① network structure **S** in which variables **X** are conditional independency, and ② local probability distribution $\boldsymbol{P}$ which connects with each variable. Let variable $\boldsymbol{X_i}$ corresponds to a node of Bayes network, and $\boldsymbol{X_j}$ is the parent node of variable $\boldsymbol{X_i}$, then the probability of child node leading to the low sentiment score of its parent node can be generalized as following.

$$\mathbf{P}(\boldsymbol{X_i} = \boldsymbol{L} | \boldsymbol{X_j} = \boldsymbol{N}) = \frac{\boldsymbol{P}(\boldsymbol{X_j}=\boldsymbol{N}|\boldsymbol{X_i}=\boldsymbol{L})\boldsymbol{P}(\boldsymbol{X_i}=\boldsymbol{L})}{\boldsymbol{P}(\boldsymbol{X_j}=\boldsymbol{N})} \tag{15}$$

Where $\boldsymbol{P}(\boldsymbol{X_i} = \boldsymbol{L})$ is the ratio of unsatisfied consumer reviews (L) for product feature $\boldsymbol{X_i}$ relative to all consumer reviews. It is calculated as following.

$$\boldsymbol{P(X_i = L) = \frac{\sum_{s=1}^{S} m(s,X_i,L)}{\sum_{s=1}^{S} n(s,X_i)}} \tag{16}$$

Where $\boldsymbol{n(s,X_i)}$ denotes the times of product feature $\boldsymbol{X_i}$ appearing on the sth consumer review. $\boldsymbol{m(s,\ X_i,L)}$ denotes the times of product feature $\boldsymbol{X_i}$ appearing on the sth consumer review that has negative evaluating on the product feature $\boldsymbol{X_i}$.

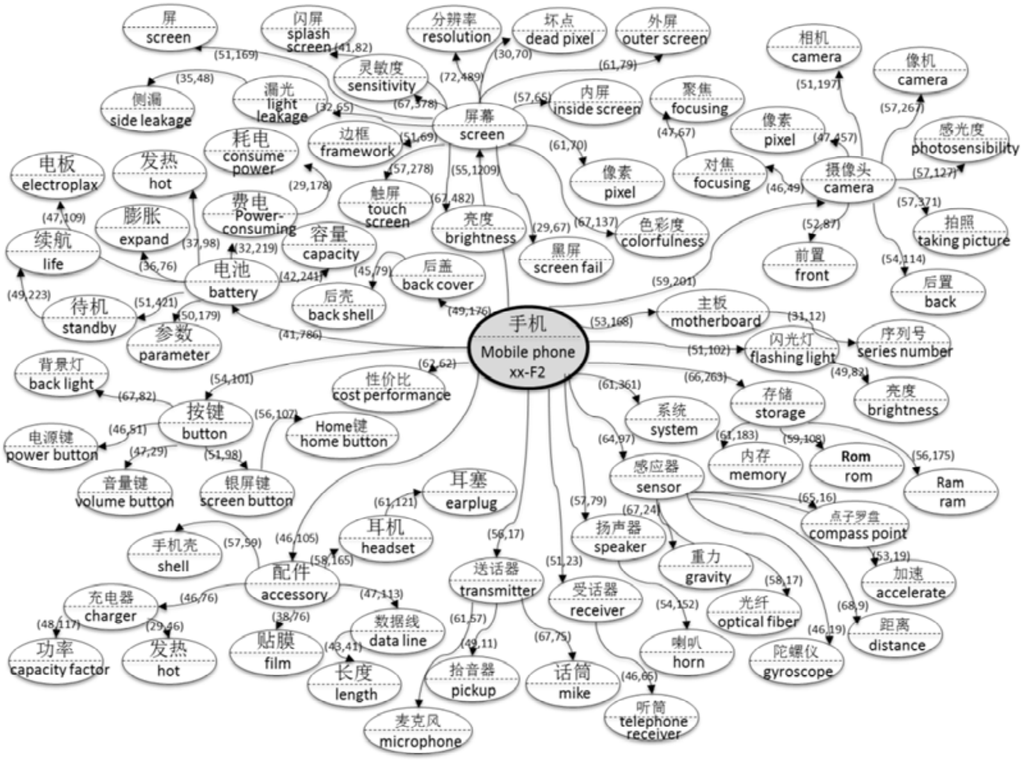


Figure 21. Product feature structure tree for product xx-F2

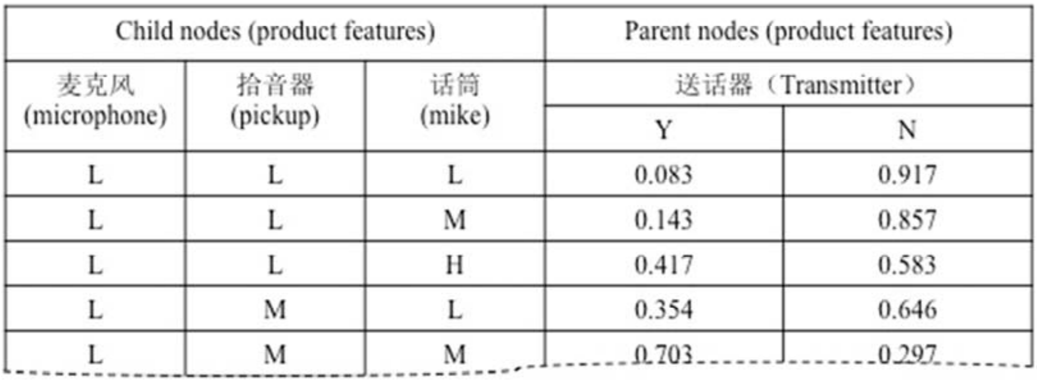
And, $\boldsymbol{P(X_j = N|X_i = L)}$ denotes the probability that a child node (product feature) $\boldsymbol{X_i}$ being evaluated as negative (described as L) leads to its parent node (product feature) $\boldsymbol{X_j}$ being evaluated as a poor feature (described as N) by consumers. Thus, $\boldsymbol{P(X_j = N|X_i = L)}$ can be generalized as following.

$$\boldsymbol{P(X_j = N|X_i = L) = \sum P(X_j = N|X_i,\cdots,X_k)P(X_i = \Gamma)_{\Gamma=L,M,H},\cdots,P(X_k = \Gamma)_{\Gamma=L,M,H}} \tag{17}$$

Where $\boldsymbol{P(X_j = N|X_i,\cdots,X_k)P(X_i = \Gamma)}$ denotes the probability that the child nodes (product features) $\boldsymbol{\{X_i,\cdots,X_k\}}$ being evaluated as positive (described as H), negative (described as L), and neutral (described as M) lead to its parent node (product feature) $\boldsymbol{X_j}$ being evaluated as a poor feature (described as N) by consumers. The probability sum of these child nodes being evaluated as positive (H), negative (L), and neutral (M) respectively denotes the probability of parent node (product feature) $\boldsymbol{X_j}$ being evaluated as a poor feature (described as N) when the child node (product feature) $\boldsymbol{X_i}$ is evaluated as negative (described as L) namely $\boldsymbol{P(X_j = N|X_i = L)}$.

Using substructure 送话器(transmitter), which has a relative low sentiment score according to our statistical results and consists of three child nodes such as 麦克风(microphone), 拾音器(pickup) and 话筒(mike), as example, the correlation matrix between the child node evaluations and the parent node evaluations from consumers is established by experts based on their observations on 3,000 consumer reviews which is presented **Table 5**. On the basis of this, the influences between parent nodes and their child nodes can be calculated based on formulas (15)-(17). The results shown that the probabilities of a relative low sentiment scores of 送话器(transmitter) causing by its child nodes such as 麦克风(microphone), 拾音器(pickup) and 话筒(mike) are 0.415, 0.327, and 0.258, respectively. It can be seen that this relative low sentiment score of 送话器(transmitter) is the most likely caused by 麦克风(microphone). Thus, designers or manufacturers should improve the 麦克风(microphone) for the future in order to increase the satisfactions of consumers for their products, and gain profit margins under fierce market competition in the end.

Table 5. Observation results of influence among product features



Child nodes (product features)			Parent nodes (product features)	
麦克风 (microphone)	拾音器 (pickup)	话筒 (mike)	送话器（Transmitter）	
			Y	N
L	L	L	0.083	0.917
L	L	M	0.143	0.857
L	L	H	0.417	0.583
L	M	L	0.354	0.646
L	M	M	0.703	0.297

Similarly, the influences among other nodes on the product feature structure tree can also be analyzed in this way. It will provide valuable evidences for the designers, manufacturers, or retailers.

9. Discussions

The main goal of Online reviews from consumers is to exchange or share information among them. The languages from consumers are characterized as oral, haphazard, and irregular syntax. And some new words or terms are also created or introduced continuously, specifically for young people. Therefore, it is necessary to adopt two-stages optimization method for word segmentation. This process can deal with the error results of direct word segmentation first, and find some new words or terms. For example, “分辨率(pixel)”, in fact, is a kind of attribute descriptions of intelligent electronic products. So “分”, “分辨”, and “辨率” are all error results of word segmentation but these results exactly exist in practice. Obviously, it is necessary to delete these error phrases from the results of original word segmentation process in order to keep the accuracy of our research and analysis works. Based on the results of original word segmentation, the correct form namely “分辨率(pixel)” can only be generated through word reorganization. However, new error forms can also be generated such as “*分”, “分辨”, and “率*”, etc. Through three filter algorithms such as frequency filtering, cohesive filtering, and left & right entropy filtering, most of these error results can be deleted from the results of original word segmentation. In addition, some new terms or phrases can also be found such as “云存储(cloud storage)” and “语音识别(speech recognition)”, etc. All these new words and terms, along with the correct results of word segmentation, are input into user dictionary again which is used to guide word segmentation at practice. And then, the process of word segmentation will be restarted based on this extended user dictionary. As a result, the correct rate of word segmentation is increased remarkably. For example, we used 1,000 consumer reviews as experiment corpus, and invited two development engineers of intelligent mobile phone and one linguist to divide reviews into phrases and annotate their POSs manually. The results are used as reference to evaluate the efficiency of word segmentation methods. And then, two kinds of word segmentation processes such as word segmentation based on *ictclas* tool directly and our proposed two-stages optimizing methods. Comparing with the reference results obtained from experts, the results generated from our two-stages optimizing method are more accuracy than those of *ictclas* tool directly which had been explained in **Figure 14** and **Figure 15**. Therefore, two-stages optimizing word segmentation method for Chinese consumer reviews is valid and necessary. It ensures to provide high quality data for the next product feature extraction analysis and application.

Product feature extraction is a complex task especially for Chinese consumer reviews, and also a crucial stage that will influence the effectiveness of applications based on product features directly. Therefore, product feature extraction in this work adopted supervised product feature extraction strategy due to its high precision. Thus, the core work is to design a reasonable rule template. Besides the elements of existing traditional rule templates, the rule template developed in this work added two kinds of elements such as governing word and opinion word to support product feature extraction and sentiment identification. By doing these, some implicit product features or sentiment expresses can be detected by combining these new adding elements with the existing elements of existing rule templates which were presented in **Figure 8**. For example, “杠杠的(ganggangde means very good)” is a recent popular express which describes a kind of positive evaluation. It is an opinion word but it isn’t contained at user dictionary exactly. Thereupon, we added it into extended user dictionary, and annotated it as opinion word manually at train set. And then, the implicit product features concerned with it can be extracted conveniently, and their sentiment score can be calculated accurately. In addition, "战斗机(fighter)" is another popular express recently. In essence, it is a noun phrase. But it is always used as an adjective phrase to modify a product feature around it and express a positive sentiment. Likeness, this phrase is also not contained at user dictionary. Therefore, it is high significant for product feature extraction from Chinese consumer reviews to find new words especially for opinion words to extend existing user dictionary through two-stages optimizing word segmentation process, and annotate the opinion attributes of phrases at train set and rule template. After doing this, the implicit product features and their sentiment evaluation can be processed accurately. These were verified in **Figure 16** which presents the efficiency of product feature extraction based on 10 different rule templates, and the 7th rule template which was proposed in this work has better results than those of rule templates.

In addition, product features are always internal correlated with each other. For example, "摄像头(camera)" and "像素(pixel)" are two product features, and may appear at different consumer reviews discretely. However, product feature "像素(pixel)" is one of the attributes of product feature "摄像头(camera)" in essence. Therefore, the internal correlation among them is an inevitable existence. Unfortunately, the existing researches don't explore this fact. This paper discussed this issue. Product feature structure tree is the representation form of the internal correlations among product features. It integrates product features which distributes at consumer reviews concretely into a whole object, and makes the comprehensive applications based on product features feasible. However, the numbers between parent nodes (product features) and its child nodes (product features), according to our observations, don't satisfy with cumulative calculation law both frequency and sentiment score e.g. between "送话器(transmitter)" and {"麦克风(microphone)", "拾音器(pickup)", and "话筒(mike)"}. The reason is that many consumers provide a snippet text description for products only for the goal of completing evaluation task required by platform or system. As a result, many product features at consumer reviews are not evaluated by consumers at all. Therefore, the influences among product features cannot be reflected by the numbers on product feature structure tree directly. For this reason, a method of inferring the influences among product features based on product feature structure tree is proposed by using Bayes theory. This method uses the sentiment scores of product features as evidences to identify the product features that need to be analyzed in depth because of its low or negative evaluations from consumers. At the same time, it makes full use of the practical evaluation results of each review from consumers. Therefore, the inferring results are more convince. For example, product feature "送话器(transmitter)" is determined as the object that need to be inferred the elements leading to its low or negative sentiment score. According to the data on product feature structure tree, child node (product feature) ”拾音器(pickup)” may be the potential element because of its lowest sentiment score. However, the inferring result from our proposed method based on Bayes theory is that child node “麦克风(microphone)” has the maximal possible of leading to low sentiment score of its parent node or negative evaluation. It is in accordance with fact. Even if the sentiment scores of “麦克风(microphone)” are not the lowest while the frequency of product feature received negative evaluation are very high which means a large amount of consumers pay attention on this product feature and give negative evaluation on this product feature. Therefore, this leads to a lower sentiment score of its parent nodes. From the perspective of probability theory and mathematical statistics, a minority events always have no statistic means in general. Therefore, product feature structure tree makes the research and analysis on the internal relations among product features feasible, and the inferring method based on Bayes theory is a valid method to keep the applications more reasonable.

10. Conclusions

A large amount of product reviews provide valuable consumer feedback. In the past decade, many researchers in computer science and information management have paid much attention to extract product features from consumer reviews, and analyze the opinion direction of consumer for product features. This paper, aiming at Chinese consumer reviews, investigates the issues of product feature extraction and the applications at product feature level. It is high significant because of emerging a huge e-commerce market in China.

In this work, a technique overview of extracting product features from Chinese consumer reviews is proposed in which two-stages optimizing word segmentation, product feature extraction based on CRF, and product feature structure tree establishing are investigated. Two-stages optimizing word segmentation process mainly consists of phrase reconstructing, frequency filtering, cohesiveness filtering, and left & right entropy filtering. It increases the correct rate of word segmentation through new phrase finding to expand user dictionary and the second word segmentation process. Likeness, an expanded rule template is proposed in which governing word and opinion word annotating are added to detect the implicit product features and infrequent opinion words. It increases both the efficiency of product feature extraction from Chinese consumer reviews and the accurateness of sentiment evaluation for product features. At the same time, two quantitative characters are defined to describe the preference extent of consumers for a product feature. Furthermore, product feature structure tree is established based on the inevitable internal correlations among product features. An algorithm is proposed to find the potential parent nodes for current product features from the results of word segmentation and different similarity functions are employed to evaluate the similarity between the potential parent nodes and the nodes of basic product structure in order to add the attribute product features into basic product structure. On the basis of these, an inferring application based on product feature structure tree is explored to identify the potential factors that lead to the low sentiment score of its parent node by using Bayes theory. This is high significant for designers, manufacturers, or retailers to implement product update, quality improvement, and market strategy, etc. Moreover, categories of comparative experiments and profound analysis are conducted on 5,806 real consumer reviews. The results generated from them provide the evidences for our research works. Finally, the case study verified the effectiveness of our proposed methods and applications.

Potential research work can be extended in many directions such as product quality and risk management and the dynamic evolution characteristics of the influences among product features, etc. These are also our future research directions.

Acknowledgements

This project was supported by the Natural Science Foundation of China (NSFC) under contract 51405462 and 51175486. Zhejiang Province Fund of Natural Science, China under contract LY16G010006 and LQ15G010005.

Reference

Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of 20th international conference on very large data bases(VLDB '94)*, 487-499.

Bahu, S. M. & Das, S. N. (2015). An Unsupervised Approach for Feature Based Sentiment Analysis of Product Reviews. *International Journal of Scientific Research Engineering & Technology*, *4*(5), 484-489.

Chang, Y. C., Chu, C. H., Chen, C. C. & Hsu, W. L. (2016). Linguistic Template Extraction for Recognizing Reader-Emotion. *International Journal of Computational Linguistics and Chinese Language Processing*, *21*(1), 29-50.

Chen, L., Qi, L. L. & Wang, F. (2012). Comparison of feature-level learning methods for mining online consumer reviews. *Expert System with Applications*, *39*(10), 9588-9601. doi: 10.1016/j.eswa.2012.02.158

Choi, Y. & Cardie, C. (2009). Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing(EMNLP '09)*, *2*, 590-598.

Dai, H. J., Tsai, R. T. H. & Hsu, W. L. (2014). Joint Learning of Entity Linking Constraints Using a Markov-Logic Network. *International Journal of Computational Linguistics and Chinese Language Processing*, *19*(1), 11-32.

Dave, K., Lawrence, S. & Pennock, D. M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product review. In *Proceedings of the 12th* *international conference on World Wide Web (WWW 2003)*, 519-528. doi: 10.1145/775152.775226

Dellarocas, C. (2003). The digitization of word of mouth: promise and challenges of online feedback mechanisms. *Management Science*, *49*(10), 1407-1424. doi: 10.1287/mnsc.49.10.1407.17308

Duan, W., Gu, B. & Whinston, A. B. (2008). Do online reviews matter? An empirical investigation of panel data. *Decision Support System*, *45*(4), 1007-1016. doi: 10.1016/j.dss.2008.04.001

Etzioni, O., Cafarella, M., Doweny, D., Popescu, A.-M., Shaked, T., Soderland, S.…Yates, A. (2005). Unsupervised name-entity extraction from the Web: an experimental study. *Artificial Intelligence*, *165*(1), 91-134. doi: 10.1016/j.artint.2005.03.001

Forman, C., Ghose, A. & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: the role of reviewer identity discloser in electronic markets. *Information System Research*, *19*(3), 291-313. doi: 10.1287/isre.1080.0193

Godes, D. & Mayzlin, D. (2004). Using online conversations to study word of mouth communication. *Marketing Science, 23*(4), 545-560. doi:10.1287/mksc.1040.0071

Htay, S. S. & Lynn, K. T. (2013). Extracting Product Features and Opinion Words Using Pattern Knowledge in Customer Reviews. *The Scientific World Journal,* Article ID 394758. doi: 10.1155/2013/394758

Hu, M. & Liu, B. (2004a). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining*, 168-177. doi: 10.1145/1014052.1014073

Hu, M. & Liu, B. (2004b). Mining opinion features in customer reviews. In *Proceedings of the 19th national conference on Artifical intelligence*, 755-760.

Hu, Z. K., Zheng, X. L., Wu, Y. F. & Chen, D.-r. (2013). Product recommendation algorithm based on users' reviews mining. *Journal of Zhejiang University (Engineering Science)*, *47*(8), 1475-1485. [In Chinese]

Jakob, N. & Gurevych, I., (2010). Extracting opinion targets in a single- and cross- domain setting with conditional random fields. In *Proceedings of the the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP' 10)*, 1035-1045.

Jiang, M. T.-J., Shih, C.-W., Yang, T.-H., Kuo, C.-H., Tsai, R. T.-H. & Hsu, W.-L. (2012). Enhancement of Feature Engineering for Conditional Random Field Learning in Chinese Word Segmentation Using Unlabeled Data. *International Journal of Computational Linguistics & Chinese Language Processing*, *17*(3), 45-86.

Jindal, N. & Liu, B. (2006). Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval*, 244-251. doi: 10.1145/1148170.1148215

Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K. & Fukushima, T. (2004). Collecting evaluative expressions for opinion extraction. In *Proceedings of the first international joint conference on natural language processing (IJCNLP-04)*, 596-605.

Kobayashi, N., Iida, R., Inui, K. & Matsumotto, Y. (2005). Opinion extraction using a learning-based anaphora resolution technique. In *Proceedings of the second international joint conference on natural language processing (IJCNLP-04)*, 173-178.

Lafferty, J. D., McCallum, A. & Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning(ICML 01)*, 282-289.

Li, F. T., Han, C., Huang, M. L., Zhu, X., Xia, Y.-J., Zhang, S. & Yu, H. (2010). Structure aware review mining and summarization. In *Proceedings of the 23rd International Conference on computational Linguistics*, 653-661.

Li, S., Ye, Q., Li, Y. J. & Law, R. (2009). Mining features of products from Chinese customer online reviews. *Journal of Management Sciences in China*, *12*(2), 142-152. [In Chinese]

Li, Z. H. (2013). *Research on Key Technologies of Chinese Dependency Parsing* (Doctoral dissertation, Harbin Institute of Technology). Retrieved from http://hlt.suda.edu.cn/~zhli/papers/zhenghua-2013-phd-thesis.pdf. [In Chinese]

Liu, B., Hu, M. & Cheng, J. (2005). Opinion observer: Analyzing and comparing opinions on the Web. In *Proceedings of 2005 World Wide Web conference(WWW 05)*, 342-351. doi: 10.1145/1060745.1060797

Liu, D. Y. & Wang, L. F. (2013). Keywords extraction algorithm based on semantic dictionary and lexical chain. *Journal of Zhejiang University of Technology*, *41*(5), 545-551. [In Chinese]

Liu, L. Z., Song, W., Wang, H. S., Li, C. C. & Lu, J. L. (2014). A Novel Feature-based Method for Sentiment Analysis of Chinese Product Reviews. *China Communications*, *11*(3), 154-164. doi: 10.1109/CC.2014.6825268

Liu, T., Wu, G. & Yao, T. (2006). Opinion Searching in Multi-Product Reviews. In *Proceedings of the Sixth IEEE International Conference on Computer and Information Technology (CIT'06)*. doi: 10.1109/CIT.2006.132

Liu, T. & Ma, J. H. (2009). Theories and Methods of Chinese Automatic Syntactic Parsing. *Contemporary linguistics*, *11*(2), 100-112. [In Chinese]

Lv, P., Zhong, L., Cai, D. B. & Wu, Y. T. (2014). Effective mining product features from Chinese review based on CRF. *Computer Engineering & Science*, *36*(2), 359-366. [In Chinese]

Ma, B. Z. & Yan, Z. J. (2014). Product features extraction of online reviews based on LDA model. *Computer Integrated Manufacturing Systems*, *20*(1), 96-103. [In Chinese]

Miao, Q., Li, Q. & Zeng, D. (2010). Mining fine grained opinions by using probabilistic models and domain knowledge. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 358-363. doi: 10.1109/WI-IAT.2010.193

Ouyang, C. P., Liu, Y. B., Zhang, S. Q. & Yang, X. H. (2015). Features-level Sentiment Analysis of Movie Reviews. *Adcanced Science and Technology Letters*, *81*, 110-113. doi: 10.14257/astl.2015.81.23

Popescu, A. & Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the conference on human language technology and empirical methods in natural language processing(HLT '05)*, 339-346. doi: 10.3115/1220575.1220618

Song, H., Yan, Y. & Liu, X. Q. (2012). A grammatical dependency improved CRF learning approach for integrated product extraction. In *Proceedings of 2nd International Conference on Computer Science and Network Technology(ICCSNT)*, 1787-1794. doi: 10.1109/ICCSNT.2012.6526267

Srikant, R. & Agrawal, R. (1995). Mining generalized association rules. In *Proceedings of the 21th international conference on very large data bases(VLDB '95)*, 407-419.

Tu, X. H., Zhang, H. C., Zhou, K. F. & He, T. T. (2012). Extracting Structured Information from Chinese Wipipedia and Measuring Relatedness between Words. *Journal of Chinese Information Processing*, *26*(3), 109-114. [In Chinese]

Turney, P. D. (2002). Thumbs, up or thumbs down: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*, 417-424. doi: 10.3115/1073083.1073153

Wang, H. S., Liu, L. Z., Song, W. & Lu, J. (2014). Feature-based Sentiment Analysis Approach for Product Reviews. *Journal of Software*, *9*(2), 274-279. doi:10.4304/jsw.9.2.274-279

Wang, W. & Wang, H. W. (2016). Comparative network for product competition in feature-levels through sentiment analysis. *Journal of Management Sciences in China*, *19*(9), 109-126. [In Chinese]

Wang, W. P. & Meng, C. C. (2011). Opinion Object Extraction Based on the Syntax Analysis and Dependency Analysis. *Computer System Applications*, *20*(8), 52-57. [In Chinese]

Wang, Y., Zhou, X. G. & Sun, Y. (2012). Research on Automatic Building of Word Correlation Net Based on Statistic. *Computer & Digital Engineering*, *40*(2), 15-18. [In Chinese]

Wei, C. P., Chen, Y. M., Yang, C. S. & Yang, C. C. (2010). Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews. *Information Systems and e-Business Management*, *8*(2),149-167.

Wong, T.-L. & Lam, W. (2005). Hot item mining and summarization from multiple auction Web sites. In *Proceedings of the fifth IEEE international conference on data mining (ICDM'05)*, 797-800. doi: 10.1109/ICDM.2005.78

Wong, T.-L. & Lam, W. (2008). Learning to extract and summarize hot item features from multiple auction Web sites. *Knowledge and Information and System*, *14*(2), 143-160. doi: 10.1109/ICDM.2005.78

Xia, T. (2007). Study on Chinese Words Semantic Similarity Computation. *Computer Engineering*, *33*(6), 191-194. [In Chinese]

Yi, J. & Niblack, W. (2005). Sentiment Mining in WebFountain. In *Proceedings of the 21st International Conference on Data Engineering (ICDE 2005)*, 1073-1083. doi: 10.1109/ICDE.2005.132

Yin, C. X. & Peng, Q. K. (2009). Sentiment Analysis for Product Features in Chinese Reviews Based on Semantic Association. In *Proceedings of International Conference on Artificial Intelligence and Computational Intelligence 2009(AICI 09)*, 81-85. doi: 10.1109/AICI.2009.326

Zhang, H. P., Yu, Z. G., Xu, M. & Shi, Y. L. (2011). Feature-level sentiment analysis for Chinese product reviews. In *Proceedings of 3rd International Conference on Computer Research and Development(ICCRD 2011)*, 135-140. doi: 10.1109/ICCRD.2011.5764099

Zhang, S. & Li, F. (2015). Opinion Target and Polarity Extraction Based on Iterative Two-Stage CRF Model. *Journal of Chinese Information Processing*, *29*(1), 163-169. [In Chinese]

Zheng, M. J., Lei, Z. C., Liao, X. W. & Chen, G. L. (2013). Identify Sentiment-Objects from Chinese Sentences based on Cascaded Conditional Random Fields. *Journal of Chinese Information Processing*, *27*(3), 69-77. [In Chinese]

Zhou, X. J., Wan, X. J. & Xiao, J. G. (2013). Collective opinion target extraction in Chinese microblogs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1840-1850.

Zhuang, L., Feng, J. & Zhu, X. Y. (2006). Movie review mining and summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management(CIKM '06)*, 43-50. doi: 10.1145/1183614.1183625

Zu, L. J. & Wang, W. P. (2014). Research of Extracting Product Features from Chinese Online Reviews. *Computer System Applications*, *23*(5), 196-201. [In Chinese]