Research on comment target extracting in Chinese online shopping platform

Zhishuo Liu, Qianhui Shen, Jingmiao Ma and Ziqi Dong
School of Traffic and Transportation, Beijing Jiaotong University, Beijing, China

Received 8 September 2018
Revised 15 October 2018
Accepted 16 October 2018

Abstract

Purpose – This paper aims to extract the comment targets in Chinese online shopping platform.

Design/methodology/approach – The authors first collect the comment texts, word segmentation, part-of-speech (POS) tagging and extracted feature words twice. Then they cluster the evaluation sentence and find the association rules between the evaluation words and the evaluation object. At the same time, they establish the association rule table. Finally, the authors can mine the evaluation object of comment sentence according to the evaluation word and the association rule table. At last, they obtain comment data from Taobao and demonstrate that the method proposed in this paper is effective by experiment.

Findings – The extracting comment target method the authors proposed in this paper is effective.

Research limitations/implications – First, the study object of extracting implicit features is review clauses, and not considering the context information, which may affect the accuracy of the feature excavation to a certain degree. Second, when extracting feature words, the low-frequency feature words are not considered, but some low-frequency feature words also contain effective information.

Practical implications – Because of the mass online reviews data, reading every comment one by one is impossible. Therefore, it is important that research on handling product comments and present useful or interest comments for clients.

Originality/value – The extracting comment target method the authors proposed in this paper is effective.

Keywords Opinion mining, Association rule, Implicit features, SA-FCM, SA-PSO

Paper type Research paper

1. Introduction

People can buy goods without leaving home via website. However, people cannot see the product entity, there always is no guarantee that the quality of the items on line with people's expectations. For the most part, people will glance over the product review before they confirm an order. The product review not only helps people doing purchase decision, but also helps merchants understand the customer's attitude to the product. Therefore, the merchants can improve their goods and services niche targeting (Chen *et al.*2015). However, because of the mass online reviews data, reading every comment one by one is impossible.

This work is supported by the National Key Research and Development Plan of China (Grant No: 2017YFB1400100).

International Journal of Crowd Science
Vol. 2 No. 3, 2018
pp. 247-258
Emerald Publishing Limited
2398-7294
DOI 10.1108/IJCS-09-2018-0019

So research on handling product comments and present useful or interested comments for clients is significant.

One comment sentence represents one opinion. At present, the research about opinion mining mainly focus on two aspects, one is mining the emotional orientation (Pak and Paroubek, 2015; Yi *et al.*, 2003; Brody and Elhadad, 2010; Pang *et al.*, 2016). It is mainly excavating customer's attitude toward the product, generally includes negative, positive and neutral (Pang *et al.*, 2002). Positive comments are good for the sale of products, while negative comments can inhibit the sale of products (Luo, 2009; Herr *et al.*, 1991; Qiu *et al.*2015). Another one is the excavation of the opinion object (Schouten and Frasincar, 2014; Popescu and Etzioni, 2005; Jakob and Gurevych, 2010; Scaffidi *et al.*, 2007; Wang *et al.*, 2013). An online shopping product always has many properties. Such as a T-shirt has the characteristics of comfort, quality, color, logistics, etc. A telephone has the characteristics of cost, appearance, service etc. In addition, different people values different characteristics. Comment sentence's opinion objects always are one or several characteristics of the products. Mining the comment object in the comment sentence enables people to quickly find out the related reviews about the feature they concerned in a certain commodity from the massive review data. Therefore, that, the customers time will be saved.

In this paper, we concentrate on the second one, which is to excavate the opinion targets in online commodities comments. Opinion objects always include explicit features and implicit features (Qiu *et al.*2015). An explicit feature is that a word or phrase which describes a property or characteristic of the product appeared in the comment text directly. Such as comment "这款手机价格便宜。"(this kind of mobile phone's price is cheap), the opinion object is "价格"(price). This word display in the opinion target directly. Hence, we called the comment text explicit comment sentence and "价格"(price) is the explicit feature. An implicit feature is that the opinion objective does not appear in the comment texts and we can deduce the evaluation objective from the context or idioms. For example, comment "这款手机便宜。" (this kind of mobile phone is cheap), we can deduce the comment objective is "价格"(price). Hence, "price" is the implicit feature and this comment is called implicit comment sentence.

In this paper, we proposed an opinion object extraction method in online shopping platform comments based on association rules. In Section 2, we summarize the research of predecessors. In Section 3, the comment sentences are segmented and part-of-speech tagged. In addition, we use IF-IDF select feature words. Then we used the particle swarm algorithm based on simulated annealing (SA-PSO) to select feature words once again, and obtain the feature words set. In Section 4, we used an improved FCM algorithm based on SA to cluster the explicit comment sentences (SA-FCM). In Section 5, we mine association rules among explicit features, opinion words and categories. In addition, establish an association rules table. According to the opinion words and association rules table, the evaluation objects in the comment features can be distinguished. At last, an experiment in Section 6 show that the comment objects extraction method we proposed in this paper is meaningful. In addition, Section 7 is the conclusion.

2. Literature review

In the research on the extraction of comment objects, many scholars believe in that the evaluation object is always a noun or noun phrase, and the evaluation word is always an adjective (Liu *et al.*, 2016; Zhang *et al.*, 2010; Liu *et al.*, 2015). Some scholars research on extracting the evaluation words and the evaluation object at the same time (Liu *et al.*, 2016; Chen *et al.*, 2016, Liao *et al.*, 2017). Liu *et al.* (2010) research on comment target extraction and corresponding sentiment classification. First, Word segmentation, part-of-speech annotation and syntactic analysis are carried out for the given text library. Second, they extract the noun and noun phrase as the comment targets. Then, word frequency filter, PMI filter and noun pruning algorithm are used to filtrate the comment targets. At last, they use the undirected method to judge the orientation of the evaluation object. The algorithm proposed in the paper is simple but the authors ignore the implicit features. Wang *et al.* (2013) proposed a hybrid association rule mining method for implicit features extraction in Chinese comments. Jiang *et al.* (2014) used a modified collocation extraction algorithm mining the basis association rule, then he used a semi- supervised LDA topic model to extract the new rules to extend the basis association rules library, so as to improve the recognition effect of implicit features. Zhang and Zhu (2013) adopted a method based on the co-occurrence association rules, which mainly consists of four steps:

(1) Determining the co-occurrence matrix.

(2) Determine the lexical correction matrix.

(3) Acquisition of candidate feature sets.

(4) Extract implicit features.

Hai *et al.* (2011) also adopted the co-occurrence association rule mining method to distinguish implicit features. The first stage was generating an association rule table between opinion word and explicit features, and the second stage is the usage of the association rules. Xu *et al.* (2015) proposed an implicit feature extraction method based on SVM classifier.

The algorithm we proposed in this paper, not only can mining explicit features but also can mining implicit features in the comment texts.

3. Data preprocessing

The computer cannot read natural language directly. Therefore, before extracting the comment objects, it must to process natural language into a form that computers can understand.

3.1 Representation of documents

When we collecting the comment texts, word segmentation and part-of-speech (POS) tagging are the first steps. For example, the word segmentation and POS tagging of the comment sentence "手机价格真便宜。" (the mobile phone's price is so cheap."手机/n 价格/n 真/d 便宜/a 。/wp" (the mobile phone/n price/n is so/d cheap/a) . In addition, *n*, v, d, a respectively denote nouns, verbs, adverbs, adjectives. Then, remove the words which is insignificant in the text, such as "because" and "er". Such words cannot provide useful information for text content and it will increase the algorithm time cost if maintained.

To distinguish the importance between different feature words, it is necessary to calculate the weights of the feature words. In this paper, TF-IDF method (Joeran *et al.*, 2017) is used to calculate the weight. In addition, the formula is as follows:

$$\mathrm{TF}-\mathrm{IDF}=\mathrm{TF}\times\mathrm{IDF}=\frac{a}{t}\times\log\frac{b}{c+1} \tag{1}$$

In formula 1, a is the frequency of the feature in the document set; t is the total number of times of all the features in the document set, b is the number of document in the document set, c is the number of documents that contains the feature.

Suppose z_i is the ith comment sentence, and each comment sentence consists of several feature words. So the comment sentence can be respected by.

$z_i=\left\{(t_1,\ w_1),(t_2,w_2),\ldots,(t_k,w_k),\ldots,(t_m,w_m)\right\}$, t_k is the feature words, w_k is the TF-IDF value and w_k is the weight vector of t_k.

According to the TF-IDF, we can select n features with the maximum TF-IDF value as the candidate features set. In addition, the vector representation of each feature word can be calculated.

3.2 Feature selection

After using the TF-IDF to select feature words, the dimension of the feature words' vector space is relatively high. So the dimension should be selected again to reduce the running time of the algorithm.

Particles swarm optimization (PSO) algorithm is an intelligent optimization algorithm which is simulated the birds foraging behavior. In addition, traditional PSO algorithm has existed a problem of premature convergence. So import the SA algorithm to PSO, which can accept a poor solution with a certain probability when the particles update the current solution. Thereby avoiding the particles search in a local scope, and make the particles gradually find the global optimal solution. In this paper, we use the SA-PSO algorithm proposed by Zhang Qiliang and Chen Yongsheng (Zhang and Chen, 2015) to select feature words for the second time.

Using binary encoding of the feature words. The value can only be 1 or 0, when the value equals to 1, it declares the feature arises in the sentence. If not, the value equals to 0.

And the fitness was calculated by formula 2:

$$\mathrm{f}(x_i, x_0) = \sum_{i=1}^{n} (x_i - x_0)^2 \tag{2}$$

where $f(x_i, x_0)$denotes the sum of all feature words and their initial position distance. x_i presents the position coordinates of the ith feature, x_0 denotes a given initial position coordinate. n is the number of the features.

4. Cluster analysis

In this paper, we first cluster the explicit comment sentences, then excavate the correlation between the opinion words and categories. Because of a huge number of online reviews and the high dimension of the review words, in this paper we use the Fuzzy C-means (FCM) clustering algorithm which applied to high-dimensional data to cluster the comment sentences. At the same time, we combined FCM clustering algorithm with SA algorithm to avoid the algorithm into local optimum.

4.1 Fuzzy C-means algorithm

Fuzzy C-means (FCM) (Bezdek, 1981) method is a kind of algorithm based on the objective function. When classifying the sample, every sample is not belong to one category certainly, but at a certain probability (membership) to determine the degree of each sample belongs to each category.

The objective function of FCM algorithm is:

$$\mathrm{J(U, V)} = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^{m} d_{ik}^{2} \tag{3}$$

Where:

c = is the number of categories;

$\mathrm{V} = (v_1, v_2, \ldots, v_c)$ donates the cluster center for each category;

$v_i = (v_{i1}, v_{i2}, \ldots, v_{ip})(i = 1, 2, \ldots, c)$ donates the cluster center of i th category.

$\mathrm{U} = (u_{ik})_{c \times n}$ is the membership matrix;

$d_{ik} = z_k - v_i$ is the Euclidean distance of the k th sentence to the cluster center of the i th class.

$\mathrm{J(U,V)}$ represents the sum of the weighted square distances between all sentences and clustering center. The clustering criterion of FCM clustering algorithm is that, make the objective function get its minimum value under some conditions by iterating over the membership matrix and clustering center.

When objective function obtains the minimum value, the FCM algorithm will output c clustering center and one membership matrix. Then, judge the category of each comment sentences according to the principle of maximum membership.

4.2 Simulates annealing-fuzzy C-means algorithm

FCM algorithm is likely to fall into local optimum during the iteration. This paper import SA (SA) algorithm into FCM algorithm, and let the FCM algorithm escape the local optimal. In the iteration process, the new solution is accepted with a certain probability but not always accept the new solution.

5. Extract the comment target based on association rules

In this chapter, we first look for the association rules among explicit features, opinion words, in each category. Build an association rules table, and regard it as a basis rules when mining the evaluation objects. Then we can excavate the implicit features contrasting the association rule table and the implicit comment words.

5.1 Association rules

To establish an association relationship between opinion words and opinion objects, the first thing is to mining the frequent item sets between opinion words and evaluation objects. Then establish the association rules between opinion words and targets according to the frequent item sets. At last, build a classifier using the association rule table.

Association rule (Agrawal *et al.*, 1993) is a kind of important correlation between data. Set $\mathrm{I} = \{i_1, i_2, \ldots, i_n\}$ represents a set of feature words, and $\mathrm{T} = \{t_1, t_2, \ldots, t_n\}$ represents a set of text, t_i is a text, and t_i consisting of multiple feature words, so t_i is a set of feature words and $t_i \subseteq I$. Then the association rules can be described as follows:

$$\mathrm{X} \rightarrow \mathrm{Y}, \text{ and } \mathrm{X} \subset \mathrm{I}, \ \mathrm{Y} \subset \mathrm{I}, \ \mathrm{X} \cap \mathrm{Y} = \varnothing \tag{4}$$

There are several important definitions in the association rules:

Definition1: Support.

$$\mathrm{Support}(\mathrm{X} \rightarrow \mathrm{Y}) = \mathrm{Support} \ \ \mathrm{X} \cap \mathrm{Y} = \mathrm{P(XY)} = \frac{n(X \cup Y)}{n} \tag{5}$$

Support reflects the probability of the feature words set {X, Y} appeared in text set T. In formula 5, n denotes the number of texts.

Definition2: Confidence.

$$\mathrm{Confidence}(\mathrm{X} \rightarrow \mathrm{Y}) = \mathrm{P}(Y|X) = \frac{n(X \cup Y)}{n(X)} \tag{6}$$

Confidence reflects that in the item set containing X, the possibility of containing Y.

Definition3: Frequent item sets.

Usually, you need to set a minimum support $\mathrm{min_sup}$ and a minimum confidence $\mathrm{min_conf}$ to limit the rules. If $\mathrm{Support(X)} \geq \mathrm{min_sup}$, X was called a frequent item set, otherwise it is an infrequent item set.

Definition4: Strong association rules.

If $\text{Support}(X \rightarrow Y) \geq \min_s up$ and $\text{Confidence}(X \rightarrow Y) \geq \min_c onf$, the association rule $X \rightarrow Y$ was called strong association rule.

The support, confidence, and frequent item sets are the prerequisites and conditions for association rule mining. Strong association rules are the result of mining association rules.

5.2 Mining association rules based on apriori algorithm

The Apriori algorithm (Agrawal and Srikant, 1994) uses a layer-by-layer progressive and iterative algorithms to look up k frequent item sets that satisfy the minimum support, and looks up the k + 1 frequent item set layer by layer until it can't find the k + 1 frequent items. Apriori algorithm is simple and easy to implement, so in this paper we mining the association rules by using Apriori algorithm.

After the first step of excavation, the number of association rules that satisfying the given constraints may be relatively large, and it is necessary to pruned the rules so as to improve the rules ability of distinguish the categories. This paper selects the confidence threshold method to prune the rules. That is, set a minimum confidence level. When the confidence level of a certain rule is smaller than the minimum confidence level, the rule is deleted; otherwise it is retained. The algorithm is as follows (Wu and Kumar, 2013):

```
Algorithm Apriori
  $I_1 \leftarrow initialT(0,1)$;//candidate1 item set, $T(0,1)$is feature set that was
  made up of 0 and 1;
  $F_1 \leftarrow \{i \in I_1 | i.supCount/n \geq min\_sup\}$;//frequent 1 item set, and $^{supCount}$ is
  the item set
  number of $X$ and $y$ co-occurance; $^{min\_sup}$is the minimum support;
  $R_1 \leftarrow \{f | f \in F_1, f.supCount/f.condCount \geq min\_conf\}$//rule set which satisfy
  the condition, $condCount$ is the item set number of $X$, $min\_conf$is the
  minimum confidence;
      for (k = 2;$F_{k-1} \neq \phi$,k++) do
          $I_k$←candidate-gen($F_{k-1}$);//the function togenerate$C_k$
          for each transaction $t \in T$ do
              for each candidate $i \in I_k$ do
                  if $i.condset$ is contained in $t$ then
                      $i.condCount++$;
                      if $t.class = i.class$ then
                      $i.supCount++$;
          $F_k \leftarrow \{i \in I_k | i.supCount/n \geq min\_sup\}$;//    $n$ is the total item in $T$;
          $R_k \leftarrow \{f | f \in F_k, f.supCount/f.condCount \geq min\_conf\}$
          return $R \leftarrow \cup_k R_k$
          The pseudo-code of candidate-gen ($F_k$) is as follows:
          function candidate-gen ($F_{k-1}$)
  $I_k \leftarrow \varphi$;
      for all$f_1, f_2 \in F_{k-1}, f_1 = \{i_1, ..., i_{k-2}, i_{k-1}\}, f_2 = \{i_1, ..., i_{k-2}, i_{k-1}'\}$ and
      $i_{k-1} < i_{k-1}'$ do
          $i \leftarrow \{i_1, .., i_{k-2}, i_{k-1}'\}$;
          $I_k \leftarrow I_k \cup \{i\}$;
          for each $(k-1)$-subset $s$ of $i$ do
              if ($s \notin F_{k-1}$) then
                  delete $i$   from $I_k$;
      return $I_k$
```

6. Experiment evaluation

6.1 Data preprocessing

Grab the comment data of HUAWEI mobile phone (model MATE S) from the Taobao platform using the data grabber tool Octopus. There are 2265 valid comments are crawled. Each comment consists of multiple clauses and contains one or more evaluation objects. First of all, the language technology platform LTP is used to split the comment sentences, and obtained 9362 clauses. Then, the comment clauses are segmented, tagged, and removed the insignificance words. Then use the TF-IDF method mentioned in the section 2 to select the feature for the first time. Finally, using SA-based particle swarm optimization algorithm to select the feature words for the second time to obtain the feature word set.

6.2 Experiment procedure

First, mark the comment objects of each comment and divide the comment objects into nine categories manually. They are customer service, performance, appearance, battery, pixel, sound quality, price, logistics and quality. These categories are different from each other and cover almost all comment objects. In addition, divided the comment set into explicit comment sentences set and implicit comment sentences set. Then cluster the explicit comment into nine classes using the SA-FCM algorithm proposed in Section 3. Then every class of texts is put into one document set. Each document set corresponds to a comment object. After that, extract the association rules between opinion words and objects. At last, excavate the implicit features of the implicit comment sentences according to the association rules.

6.2.1 Clustering the explicit comments. The SA-FCM method is used to clustering the explicit sentences. Because the comment objects are divided into 9 categories by hand, the number of clustering class is set to nine. In this paper, the program of SA-FCM algorithm is written in MATLAB. After the iteration of the algorithm, we finally obtained the membership matrix of the explicit sentences. Part of the degree of membership is shown in Table I, where the row represents the category and the column represents the features.

According to the principle of maximum membership: the sentences' category is the highest degree of membership in the evaluation matrix. For example, in the first sentence, the 4th degree of membership is the largest. Therefore, sentence 1 belongs to the 4th category.

After obtaining the category of every explicit sentence, put the same category sentences into the same folder, then we will get nine folders. Marking each comment sentence according to its category, explicit features, and opinion words. Such as $z_i =< o_i, q_i, c_j >$, where z_i is a comment sentence,o_i is the explicit features in the sentence, and q_i is opinion

1	2	3	4
0.0299	0.0111	0.0045	0.0072
0.1378	0.0756	0.8309	0.0218
0.247	0.5093	0.0734	0.0438
0.2541	0.0466	0.0155	0.0186
0.1383	0.0943	0.0171	0.0343
0.0896	0.0179	0.0092	0.0091
0.0369	0.0204	0.0059	0.0145
0.0373	0.1528	0.0156	0.7846
0.0292	0.072	0.0279	0.0661

Table I. Partially explicit sentence membership matrix

words, c_j is the category, such as the explicit comment “样式非常美观” (the style is very beautiful), the result of the labeling is “<样式, 美观, 外形>” (style, beautiful, appearance). At the same time, put synonymous explicit features into an aggregate $O_i(i = 1, 2, \ldots 9)$ as the final feature set.

6.2.2 Extract the association rules in explicit sentence. Mining of association rules for these nine document sets. The feature words are different in different document set. Then let the minimum support equals to 0.01 and the minimum confidence equals to 0.3. Then mining the association rules by the algorithm mentioned in section 4. So the association rules were shown in Table II.

In Table I, “服务→不错(service → pretty good) (5.9289 per cent, 40.3587 per cent)” indicates that the feature word “service” recommended feature word “pretty good” support was 5.9899 per cent, and the confidence was 40.3587 per cent. At the same time, the feature words “service”, “attitude”, and “customer service” all refer to the “customer service” category. In this category, use <service, pretty good> as the association rule for that category. In total, the association rules are shown in Table III.

6.2.3 Extract the implicit features. After mining the association rules of each category, the implicit features in the implicit comments can be extracted according to the comment words of the sentences and the association rules table. For example, the sentence “手机很好看” (mobile phone looks good), it is easy to found the association rule <appearance, good-looking> in the association rule table. Therefore, we can determine the implicit feature is appearance.

6.3 Experimental results

We always use precision and recall to evaluate the performance of the extraction method. Precision is the ratio of the correct number of documents retrieved to the total number of the documents retrieved. The recall is the ratio of the correct number of documents retrieved to the correct number of documents in the document library.

Precision = Number of correct information extracted/Number of information retrieved;
Recall = Number of correct information extracted/Number of information in sample.

However, the precision and recall are contradictory in some cases. The increase of the precision may lead to a decrease of the recall. In addition, the increase of the recall may also lead to a decrease of the precision. Therefore, it is usually necessary to introduce an *F*-value to measure the feature extraction effect. The formula to calculate the *F*-value is as follows:

$$\mathrm{F} - \mathrm{value} = 2 \times \mathrm{presicion} \times \mathrm{recall}/(\mathrm{presicion} + \mathrm{recall}) \tag{7}$$

In this paper, the accuracy of the manual annotation is set to 100 per cent. The precision, recall and F-value of the feature extraction algorithm proposed in this paper are shown in Table IV.

Table II.
(Portion) association rules

Rule	(Support, Confidence)
服务 → 不错 (service → pretty good)	(5.9289%, 40.3587%)
服务 → 花粉 (service → pollen)	(5.7971%, 39.4619%)
正品 → 不错 (genuine → pretty good)	(4.7431%,33.8028%)
态度 → 不错 (attitude → pretty good)	(3.6232%, 60.4396%)
速度 → 不错 (speed → pretty good)	(3.6891%, 46.6667%)

Category	Opinion targets	Opinion words
服务 (customer service)	客服、花粉、态度、服务 (costumer service, pollen, attitude, service)	不错、耐心、热情、周到、敬业、认真、亲切、细心、友善 (pretty good, Patient, warm, thoughtful, professional, earnest, kind, careful, friendly
性能 (perform-ance)	性能、反应、运行 (Performance, reaction, operation)	卡、卡顿、顺畅、稳定、流畅 (carton, carton, smooth, steady, fluency)
外观 (appeara-nce)	外观、外形、样式、做工 (facade, shape, style, work)	好看、美观、漂亮、精致、精美、一般、大气 (pretty, artistic, beautiful, delicate, exquisite, general, elegant)
电池 (battery)	电池、续航、电量 (Battery, endurance, energy)	够用、耗电、耐用 (Use-enough, power consumption, durable)
像素 (pixel)	像素、照相、拍照、分辨率、画质 (pixel, photo, photo, resolution, picture quality)	清晰、清楚 (Clearness, clear)
音质 (sound quality)	音质、声音、音效 (Sound quality, sound, sound effects)	外放、柔和、低音、细腻 (overtones, soft, bass, delicate)
价格(price)	价格、价位、性价比 (cost, price, cost performance)	优惠、低、便宜、划算、实惠、贵、物有所值、物美价廉、高 (Preferential, low, cheap, cost-effective, affordable, expensive, good value, cheap and fine, high)
物流 (logistics)	物流、快递 (Logistics, express)	给力、快 (force, fast)
质量 (quality)	质量、品牌、品质 (Quality, brand, character)	正品、原装、真机 (Authentic, original, authentic)

Table III.
Association rule table

	SA-FCM + association rules (%)	SA-FCM (%)
precision	83.26	75.26
recall	68.37	68.72
F-value	75.08	71.84

Table IV.
Precision and recall

From Table IV, we can see that after mining association rules, the precision and *F*-value improved a lot. At the same time, the association rule table established in this paper can not only be used to extract the implicit feature, but also can be used to extract the explicit feature according to the opinion words. Therefore, the algorithm proposed in this paper has a certain practical value.

7. Conclusion

The online products reviews enable user to learn about the product more comprehensive from the purchased user's real feeling. It can help users make purchase decisions, and the businesses also can improve the product based on the user's feedback. This paper makes an opinion mining research on user online reviews and primary research studies the features in the reviews and takes the implicit features into account, to achieve a more complete extraction of review features. This paper grabs user review data from the e-commerce platform, and carries out operations such as word segmentation, and part-of-speech tagging on the captured data. Then extracts the feature words twice. An FCM clustering algorithm based on SA algorithm is proposed to cluster explicit comment sentences. Based on the classification document sets of explicit sentences, using text association rule mining algorithm to extract rules, and establish rule libraries of each category to extract implicit features. Through an experiment, we compared the precision, recall and *F*-value use the association rules and not use the association rules. In addition, find that the values of using association rules are higher than that without association rules. Therefore, we can draw a conclusion that the method proposed in this paper - use SA-FCM clustering first and then use association rule table to mine comment objects is effective.

At the same time, there are many deficiencies in this paper. For example, first, the study object of extracting implicit features is review clauses, and not considering the context information, which may affect the accuracy of the feature excavation to a certain degree. Second, when extracting feature words, the low-frequency feature words are not considered, but some low-frequency feature words also contain effective information. In later work, it would greatly improve the accuracy of excavate the implicit features if these two points are considered.

References

Agrawal, R. and Srikant, R. (1994), "Fast algorithms for mining association rules", *Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 487-499.

Agrawal, R., Imielinski, T. and Swami, A. (1993), "mining association rules between sets of item in large database", *Proceedings of The 1993 ACM SIGMOD International Conference on Management of Data - SIGMOD'93*, p. 207.

Bezdek, J.C. (1981), "Pattern recognition with fuzzy objective function algorithms", *Plenum*, Vol. 22, pp. 203-239.

Brody, S. and Elhadad, N. (2010), "An unsupervised aspect-sentiment model for online reviews", *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, *Association for Computational Linguistics*, *Los Angeles, Cafifornia*, pp. 804-812.

Hai, Z., Chang, K. and Kim, J.J. (2011), "Implicit features identification via co-occurrence association rule mining", *International Conference on Intelligent Text Processing and Computational Linguistics and Intelligent Text Processing*, pp. 393-404.

Herr, P., Kardes, F.R. and Kim, J. (1991), "Effects of word-of-mouth and product-attribute information on persuation: an accessibility- diagnosticity perspective", *Journal of Consumer Research*, Vol. 17 No. 4, pp. 454-462.

Jakob, N. and Gurevych, I. (2010), "Extracting opinion targets in a single-and cross-domain setting with conditional random fields", *Proceeding of the Conference on Empirical in Natural Language Processing*, pp. 1035-1045.

Jiang, W., Pan, H. and Ye, Q. (2014), "An improved association rule mining approach to identification of implicit product aspects", *The Open Cybernetics and Systemics Journal*, Vol. 8 No. 1, pp. 924-930.

Liao, X., Chen, X., Wei, J., *et al.*, (2017), “A multi-layer relation graph model for extracting opinion targets and opinion words”, *Acta Automatica Sinica*, Vol. 43 No. 3.

Liu, K., Xu, L. and Zhao, J. (2015), “Co-extracting opinion targets and opinion words from online reviews based on the word alignment mode”, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 27 No. 3, pp. 636-650.

Liu, Q., Huang, H. and Feng, C. (2016), “Co-extracting opinion targets and opinion-bearing words in chinese Micro-Blog texts”, *ACTA Electronica SINICA*, Vol. 44 No. 7.

Liu, H., Zhao, Y., Qian, B. and Liu, T. (2010), “Comment target extraction and sentiment classification”, *Journal of Chinese Information Processing*, Vol. 24 No. 1.

Luo, X. (2009), “Qualifying the long-term impact of negative word of mouth on cash flows and stock prices”, *Marketing Science*, Vol. 28 No. 1, pp. 148-165.

Pang, B., Lee, L. and Vaithyanathan, S. (2002), “Thumbs up? Sentiment classification using machine learning techniques”, *Proceeding of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79-86.

Pang, J.H., Li, X., Xie, H. and Rao, Y. (2016), “SBTM: topic modeling over short texts”, *DASFAA 2016 Workshop*, pp. 43-56.

Popescu, A.M. and Etzioni, O. (2005), “OPINE: extracting product and opinion from reviews”, *Proceeding of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 32-33.

Scaffidi, C., Bierhoff, K., Chang, E., Felker, M., Ng, H. and Jin, C. (2007), “Red opal: product-feature scoring from reviews”, *Proceeding of the 8th ACM Conference on Electronic Commerce*, pp. 182-191.

Schouten, K. and Frasincar, F. (2014), “Finding implicit features in consumer reviews for sentiment analysis”, *International Conference on Web EngineerinG (ICWE2015)*, pp. 130-144.

Wang, W., Xu, H. and Wan, W. (2013), “Implicit feature identification via hybrid association rule mining”, *Expert Systems with Applications*, Vol. 40 No. 9, pp. 3518-3531.

Wu, X. and Kumar, V. (2013), *The Top Ten Algorithms in Data Mining*, Tsinghua University Press.

Xu, H., Zhang, F. and Wang, W. (2015), “Implicit feature identification in chinese reviews using explicit topic mining model”, *Knowledge-Based Systems*, Vol. 76, pp. 166-175.

Yi, J., Nasukawa, T., Bunescu, R. and Niblack, W. (2003), “Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques”, in *Proceeding of the Third IEEE International Conference on DataMing (ICDM’03)*, pp. 427-434.

Zhang, L., Liu, B., Lim, S. and O’Brien-Strain, E. (2010), “Extracting and ranking product features in opinion documents”, *Proceeding of the 23rd International Conference on Computational Linguistics: Posters*, *ACL*, pp. 1462-1470.

Zhang, Y. and Zhu, W. (2013), “Extracting implicit features in online customer reviews for opinion mining”, in *Proceedings of the 22td International Conference on World Wide Web Companion*, pp. 103-104.

Further reading

Beel, J., Langer, S. and Gipp, B. (2017), “TF-IDuF: a novel term- weighting scheme for user modeling based on users’ personal document collections”, *12th iConference*.

Olson, D.L. and Delen, D. (2008), *Advanced Data Mining Techniques*, p. 138.

Pak, A. and Paroubek, P. (2010), “Twitter as a corpus for sentiment analysis and opinion minning”, *LREC*, Vol. 10 No. 2010, pp. 1320-1326.

Shi, Y. and Eberhart, R. (1998), *A Modified Particle Swarm Optimizer*, *Springer Berlin Heidelberg*, Vol. 3612, pp. 439-439.

XingJun, C., Jingjing, W., Xiangwen, L., Si-Yuan, J. and Guo-Long, C. (2016), "Extraction of opinion targets and opinion words form chinese sentences based on word alinment model", *Journal of Shandong University (Natural Science)*, Vol. 51 No. 1.

yuanlin, C., Yueting, C., Yi, L. and Yang, X. (2016), "Transaction rating credibility based on user group preference", *Journal of Tsinghua University (Sci& Technol)*, Vol. 55 No. 5, pp. 13-18.

Zhang, Q. and Chen, Y. (2012), "Partical swarm algorithm modified based on ideal of simulated annealing for knapsack problem", *Journal of Modern Electronics Technique*, Vol. 12, pp. 85-89.

Corresponding author
Zhishuo Liu can be contacted at: liuzhishuokobe@163.com