Extracting and Clustering the Evaluation Objects of the Chinese Product Recommendation System Based on the Opinion Mining

Yingbin Xue[1, 2], Xiaoye Wang[1, 2], Yingyuan Xiao[1, 2], Yukun Li[1, 2], Wenguang Zheng[1, 2]

[1]Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Tianjin University of Technology, 300191, Tianjin, China)

[2]Key Laboratory of Computer Vision and System, Tianjin University of Technology, Ministry of Education, 300191, Tianjin, China

Keywords: Product recommendation, opinion mining, extraction of theme words, clustering

Abstract: The research of product recommendation system mainly focuses on the user s behavior or the commodities. contents, but rarely focuses on the commodies. reviews. This paper extracts useful information hidden in the commodies. reviews by opinion mining technology. It is more targeted that recommending product to users according to the user's favorite property. The main process of opinion mining is the extraction of topic words and the polarity judgement of polar words. Because the time complexity of the topic extracting algorithm is high, this paper extracts the explicit evaluation object and evaluation words by using the method of matching noun phrase and then setting up a semantic mapping set of evaluation objects and evaluation words to determine the implicit evaluation object. In this paper, k-means and BIRCH are combined to cluster the evaluation objects. K-Means algorithm is used for pre-clustering for the BIRCH algorithm to solve local optimum. And the advantage of BIRCH is it can get the number of clusters by self-learning. And delete the clusters contained few contents to pruning evaluation objects. It can reduce the time complexity and guarantees the clustering effect.

1. Introduction

With the rapid development of Internet, user-supplied reviews are solicited ubiquitously by online retailers [3]. The review is a true evaluation of customers who purchased this product. The customer’s review plays an important role in deciding the purchasing behavior for online shopping as a customer prefers to get the opinion of other customers by observing their opinion through online products’ reviews. Which reflect the customers’ sentiments and have a substantial significance for the products being sold online website? However, if the sale of a product is huge, there will be more reviews. It may be quite time-consuming for users to view them one by one, and not all the reviews are helpful. The solution for this situation is users need to choose the rating (good, neutral, bad) of the product when they submit reviews, so that all users can know the overall evaluation of the products. But the descriptions of product’s attributes are still not detailed. Customers need to further read to get their needed attribute information, so fussy viewing process is likely to reduce the user's desire to buy the product.

It is an urgent need for a recommendation system based on products’ reviews, which can extract useful information from reviews information. For example, e-commerce platform can display the products’ attributes and the modifiers used to describe it. The user also can set their favorite product’s attributes. The system will recommend products to user based on their favorite to meet their demands. The users no longer need to view all the reviews.

In this paper, we use the opinion mining technology to analyze the reviews texts. The concept of opinion mining was first proposed by Kim and Hovy in 2004 [4]. They decompose opinion mining technology into four processes: extract thematic words, identify opinion holders, determine the scope of statements, and determine polarity of polar word. However, extracting thematic words and determining the polarity of polar words are the main processes of opinion mining. This paper mainly researches the extraction and clustering of thematic words.

2. Related Works

Extracting thematic words is a process of extracting the subject was evaluated and the evaluation words. This paper called them evaluation object and evaluation terms. The current extraction technology includes two kinds of technology, they are supervised learning and unsupervised learning respectively.

2.1 The Supervised Learning Extraction

Kobayashi and Matsumoto et al. [5] consider an opinion as a chunk of information consisting of these three slots: <Subject, Attribute, Value>. The attribute slot specifies which aspect of a subject is focused on. Attributes of a subject of evaluation include its qualitative and quantitative properties, its constituents, and services associated with it. The value slot specifies the quantity or quality of the corresponding aspect. They matched thematic words according to these three slots. Chen Qizhe, Yao Tiantang and others [1] used the TF / IDF algorithm to extract the evaluation objects. They computed the frequency of co-occurrence of each evaluation object and evaluation terms and the distance between them and the character of words in the reviews to confirm evaluation objects. But these supervised learning algorithms have human participation, lead to the coverage rate is not enough.

2.2 Unsupervised Extraction Research

Jeonghee Yi et al. [7] applied NLP techniques to develop the hybrid language model and similarity test model, they can identify topic related feature terms from online review articles, enabling sentiment analysis at finer granularity. Velislava Stoykova proposed an approach using statistically-based techniques for collocation extraction to analyze semantic content of academic subjects [10]. Popescu and Etzioni establish a system of extracting subject term based on Point wise Mutual Information [8]. This system calculating the PMI of the noun and all of the phrases of this noun, and regard the noun with the highest PMI as the evaluation object. Because the calculation of PMI value depends on the corpus too much, the algorithm is inefficient.

Because many redundant words are produced in the process of participles and extract thematic words, it is necessary to prune the evaluation object to remove redundant information. However, most of the current opinion mining research process does not prune the evaluation object. It will affect the accuracy of the follow-up research.

In order to solve the above problems, this paper first match the review texts based on the noun phrase pattern to extract the explicit evaluated object and the evaluation words. And then mine the implicit evaluated object by establishing the semantic mapping set. At the same time, we extract the adverbs word in order to determine the polarity of the evaluation word in the next work. This paper takes the K-Means clustering algorithm and BIRCH clustering algorithm to prune the evaluation object. K-Means clustering algorithm is used for pre-clustering to solve local optimum of BIRCH. And then use the BIRCH algorithm to cluster the evaluation object and delete some clusters contained few contents to prune the evaluation object. It will reduce the time complexity and ensuring the clustering effect.

3. The Extraction of the Evaluation Object

The evaluation object refers to the product attributes described in their reviews. Consider one example, “外观漂亮/appearance is beautiful". The evaluation object is " 外观/appearance ". Appearance is an attribute of laptop. And many different evaluation objects are the same attribute of the products. Such as "外观/appearance" can also be expressed as "样式/style", "外形/shape" and so on. Thus, it is necessary to clustering and pruning the evaluation objects.

3.1 Pre-Processing Reviews

Before extracting the thematic word, the reviews are firstly pre-processed. The obviously useless reviews and the reviews without adjective need be deleted. The obviously useless reviews include the following situations:

Sentence containing "?". These questions generally are not the subjective reviews on the products, such as "Do you think this computer is good?”

A series of numbers or letters. These numbers and letters generally are telephone number, QQ number or some URL, etc. These usually are the information of advertising.

The repeated reviews. It is impossible for two different users to have the same review on the same product. These duplicate reviews usually are a user’s repetitive reviews for the website’s score. These do not reflect the information of the good, and needs to be deleted.

The reviews without adjective need to be deleted. Because these reviews can’t provide any subjective information about the products’ attributes. For the example "The model ordered is r9290x, give me 8970m." Such reviews are useless for other users to understand the quality of the product's attributes. Therefore, such reviews need to be deleted.

We use the regex method to match the adjectives in the reviews after labeling the characteristic or property of reviews’ words, and delete the comment without a successful match.

3.2 Extract the Thematic Words

The thematic words mainly include the evaluation objects, evaluation words (adjectives) and evaluation adverbs. However, the evaluation objects are divided into implicit evaluation objects and explicit evaluation objects. Such as"散□很快/cooling quickly", the evaluation object is "散□/cooling", which is an explicit evaluation object. The implicit evaluation object exists in some reviews without a clear subject, but these reviews also express user’s subjective ideas about the products. Such as: "too slow", we can judge this review’s evaluation object is "速度/speed."

The reviews are divided into several clauses by punctuation to facilitate the extraction. The punctuation contains: ",", ".", "!", " ", "...... ","; ","? "," ~ " and so on. In this paper, several noun phrase patterns are constructed through the syntactic analysis to extract the thematic words. The noun phrase patterns are as follows:

n and n and n pattern, such as "外观和屏幕和系统/appearance and screen and system". In this mode, there are three objects: the appearance, the screen and the system.

n and n pattern, there are two evaluation objects.

n’ s n pattern, the evaluation object is the n after the auxiliary word "’s"

The nnn or nn or n pattern, for example: "电脑整体效果/overall effect of computer", the evaluation object is the whole of all the nouns "the overall effect of the computer".

The steps of thematic word extraction are as follows:

For each sub-clause of the review, we use the above four kinds of noun phrase patterns to seek the evaluation objects and the adjectives and adverbs of them.

If the sub-clauses only have an evaluation object but without evaluation words, we save this object. And then if there is no evaluation object in the next sub-clauses but have evaluation words, so we match this evaluation object with this evaluation words together.

If there is no evaluation object in sub-clauses, only have the evaluation words. We check whether its foregoing sub-clauses have an evaluation object, and if there are evaluation object, we use it as the evaluation object of the evaluation words; Otherwise, the evaluation object of this sub-clauses is implicit evaluation objects. Thus, we establish the semantic mapping set of evaluation object and evaluation word to affirm the implicit evaluation object.

3.3 The Extraction of the Implicit Evaluation Object

This paper extracts the implicit evaluation object by establishing the semantic mapping set. The semantic mapping set refers to a mapping set consist of evaluation object and evaluation words. For example, the evaluation objects are "外□/appearance, 系□/system, CPU" and the evaluation words are "漂亮/beautiful, 流□/fluent". So, the semantic mappings set can be <外□/appearance 漂亮/beautiful >, <外□/appearance 流□/fluent >, <系□/System 漂亮/beautiful >, <系□/System 流□/fluent>, <CPU 漂亮/beautiful >, <CPU 流□/fluent>.

There are three steps of establishing the semantic mapping set:

Counting the frequency of each evaluation object and evaluation words which appear in all of the review text. Then keep the top 50 evaluation objects and top 100 evaluation words. They are defined as the high-frequency evaluation objects and evaluation terms.

The co-occurrences frequency of each pair of high-frequency evaluation objects and high-frequency evaluation words in all of the reviews text is calculated.

For each high-frequency evaluation object, we select top 50 co-occurrences frequency evaluation words to set up the semantic mapping sets.

4. Clustering the evaluatiom objects

4.1 The Representation of Evaluation Objects

In this paper, all evaluation objects are transformed into the form of vector for easy calculation. Top 20 high-frequency evaluation words are selected from the evaluation words. Calculate the co-occurrence frequency of the evaluation objects and each evaluation words of the top 20 high-frequency evaluation words. Thus, the evaluation object becomes a 20-dimensional vector. The element value in the vector is the co-occurrence frequency of the evaluation object and the corresponding evaluation words. For example, in this paper, the 20 evaluation words selected for the laptop are: "不□/good", "高/high", "快/fast", "□意/satisfied", "大/big", "好用/easy to use", "慢/slow", "喜□/like", "一般/ordinary", "□力/to force", "差/bad", "便宜/cheap", "厚/thick", "小/small", "漂亮/beautiful", "□/strong", "□惠/inexpensive", "重/heavy", "麻□/trouble" and "垃圾/garbage”. For the evaluation object "computer" will be expressed as a vector of [549, 200, 217, 141, 79, 127, 45, 57, 60, 42, 59, 48, 32, 125], each element in the vector represents the co-occurrence frequency of the "computer" and the 20 evaluation words in all the reviews.

4.2 The BIRCH Algorithm

In this paper, BIRCH algorithm is used to cluster evaluation objects. The core problem in BIRCH algorithm is the construction of CF tree. Clustering feature tree also called CF tree, the structure shown in Figure 1.

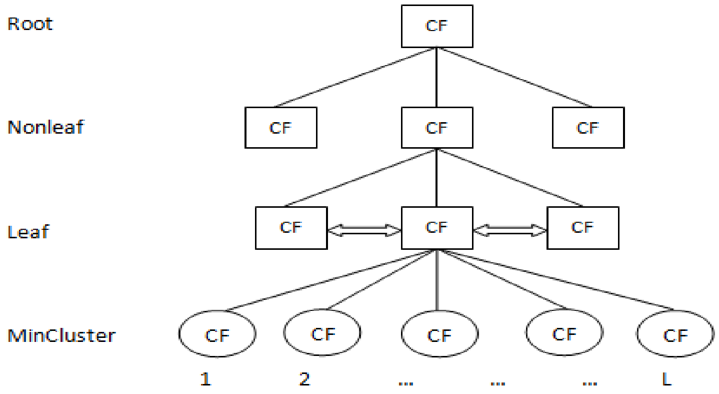

Figure 1. Clustering feature tree

The process of building CF tree is actually the process of inserting each data point into the tree. The process involves the following three parameters: balance factor B of the internal node, balance Factor L of the leaf node and threshold of the cluster diameter T.

The insertion order of nodes has a great influence on the clustering results of BIRCH algorithm [6]. For two data that belong to the same cluster, the different insertion order will lead to be split into two different leaf nodes. Therefore, by pre-clustering the initial dataset, we give a rough insertion order for the given dataset, and then insert the data into the CF tree according to this order. The pre-clustering algorithm is the K-Means. It is simple, efficient and suitable for pre-processing the large-scale data. Because it does not increase the complexity of the BIRCH algorithm, the complexity is low.

4.3 Parameter Determination of BIRCH Algorithm

In the BIRCH algorithm, there are 3 parameters T, B and L need to be determination. T is the biggest diameter threshold of the leaf node. B is the number of child nodes of the non-leaf node. And L is the number of CF vector in the leaf nodes mostly. Select 200 data from the data set, calculate the distance between every two data, and then calculate the mean value EX and the variance DX of these distances, then the threshold of the cluster' s diameter T is EX + 0.25 × DX [9]. B and L are obtained by experiments. We adopt the best values of B and L by repeatedly making experiments.

4.4 Pruning the Evaluation Object

In this paper, the evaluation object fall in the same leaf node will form a cluster. If a cluster has too little evaluation objects, it means the evaluation objects of this cluster are likely to be the redundant attributes, which are irrelevant to the product. Thus, the evaluation object in the cluster will be prune. This method doesn't depend on the corpus. And the BIRCH algorithm scans the database only one time, which also has great advantages in the time complexity.

5. Experimental Results Analysis

The data used in the experiment comes from 284 best-selling laptops' reviews of Jingdong's website [2]. The reviews hold about 200,000 pieces of data totally. We first divide the data into two parts, and select 2,000 pieces of data to evaluate the validity of our method, and the rests pieces of data are used to construct the product recommendation system.

5.1 Analysis of the Evaluation Object Pruning

In order to validate the effectiveness of the evaluation object pruning, we compare the BIRCH algorithm with the p-support method of Hu& Liu [9] and PMI algorithm in three parameters, they are precision, the recall rate and the value of F. The three parameters are calculated as follows:

$$precision(p) = \frac{PC}{AP} \tag{1}$$

$$recall(r) = \frac{PC}{SP} \tag{2}$$

$$F = \frac{2rp}{p+r} \tag{3}$$

Where PC is the number of reviews pruned correctly, AP is the number of all reviews pruned. SP is the number of reviews should be pruned.

5.1.1 Compare the Pruning Result

Set the cluster' s diameter threshold T of the BIRCH algorithm to 3.8, and the number of child nodes B in non-leaf nodes to 10 and the number of child nodes in leaf nodes to 15, and set the value of p-support to 5, and the pruning threshold in the PMI algorithm is set to 0.3. The 2,000 pieces of data are divided into 10 set. These three methods were tested in different segments of the data to respectively calculate the precision, recall and the value of F, and they are shown in Figure 2-4.

Figure 2-4 illustrate the comparison results of three algorithms. BIRCH algorithm is obviously better than the PMI algorithm and p-support algorithm. Because PMI algorithm is very dependent on the corpus and threshold, it needs a large scale of corpus. The improper corpus will directly affect the pruning results. And the attribute features are manually established, which is easy to cover incomplete, resulting in low efficiency.

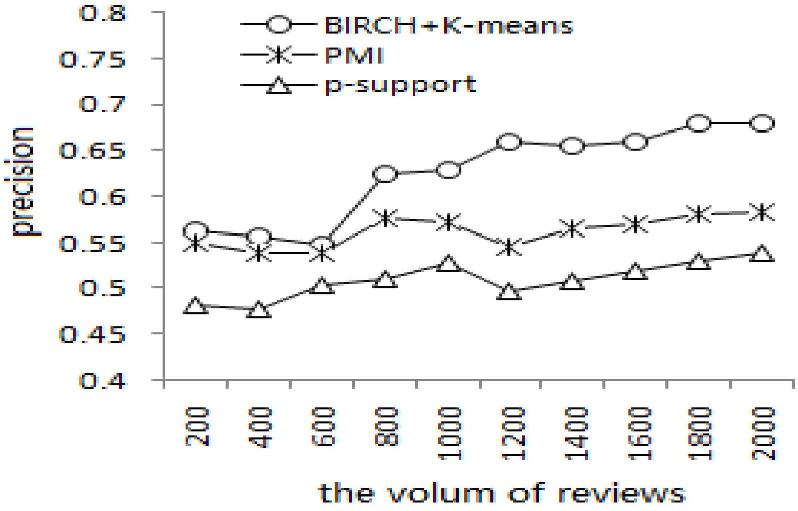

Figure 2. The precision comparison of three methods

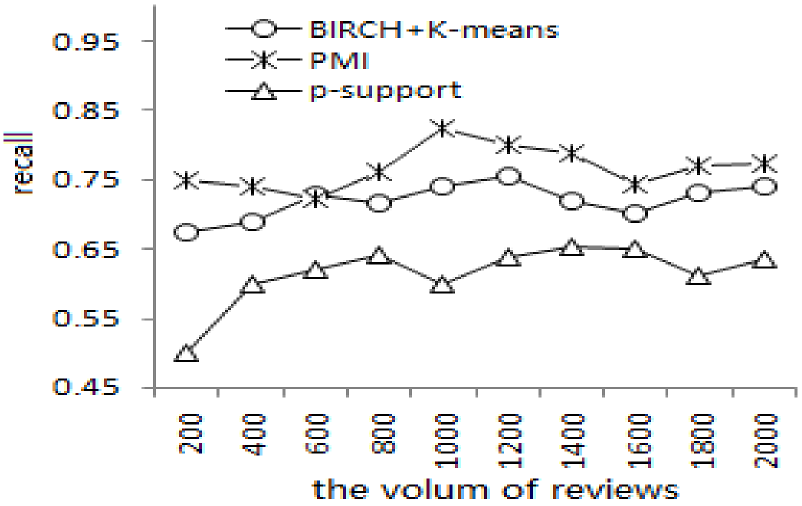

Figure 3. The recall comparison of three methods

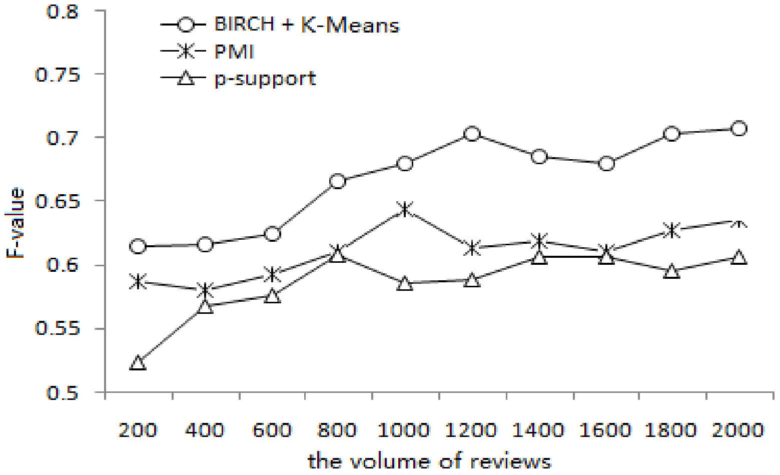

Figure 4. The F-value comparison of three methods

5.1.2 Compare the Time Complexity.

Figure 5 shows the comparison among the running time of three methods. Since the PMI algorithm needs to obtain the number of web searches of the evaluation object, it is the most time consuming, the BIRCH algorithm has the lowest time complexity.

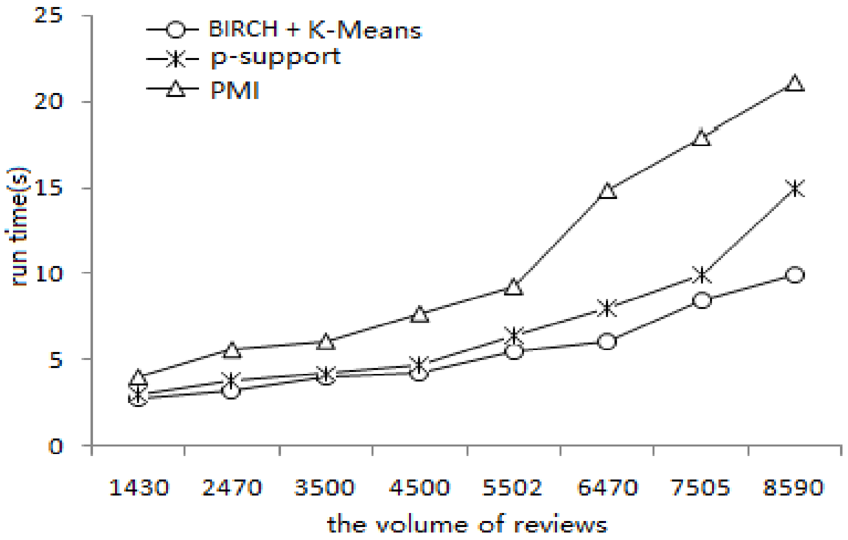

Figure 5. The running time comparison of three methods

5.2 Comparison of Clustering Result for the Evaluation Objects

In the process of clustering all the evaluation objects, there are two kind of method. One is setting up the mapping set between the attributes of the product and the common expressions of these attributes. Then search the mapping set to determine the final attribute of each evaluation object. Another method is clustering, the most widely used algorithm is K-Means clustering. In order to verify the validity of BIRCH clustering algorithm, we compared it with the above two methods and made analysis. Using the data from 4.1.1 for these three methods. The experiment mainly compares the accuracy of the three methods.

$$accuracy = \frac{CC}{AR} \tag{4}$$

Where, CC is the number of review clustering correctly, and AR is the number of all review. The mapping sets of attributes and the common expression are shown in Tabel 1

Table 1. Mapping Set of Attributes

Attributes	Common expression of attribute
Computer	电脑/computer,货/goods,总体/ overall,东西/ product,机子/ computer,机器 /machine,笔记本/notebook,整体/overall,产品/product,效果/effect,本/notebook, 本子/notebook， 商品/goods
Cost performance	性价比/costperformance,价格/price,价钱/price,价/price
Delivery speed	送货速度/delivery speed,物流/logistics, 快递/express
Appearance	外观/appearance,外形/style,外表/look
Cooling	散热/cooling,热量/heat
Sound	音质/sound quality,噪音/noise
Usage	使用/use, 用起来/use
Service	服务/service,京东/Jingdong, JD,售后/sale,售后服务/after-sales service

In the K-means clustering algorithm, the distance of the evaluation object was measured by Euclidean distance. And the experiment shows that when the objects are clustered into 13 clusters, the clustering result was best. Therefore, K was set to 13.

Contrastively analyzing the results of these three methods, Figure 6 is line chart of the accuracy of clustering attributes by three methods.

The clustering effect of the BIRCH is the best, which can be seen from Figure 6. Because the mapping set is manually established, which inevitably is incomplete. Thus, it is most inefficient in three methods. However, the BIRCH+K-means clustering is equivalent to combine the re-clustering based on K- means algorithm on basis of clustering themselves. Thus, the effect is better than the effect of K-means clustering. In addition, the BIRCH+K-means algorithm has more obvious advantage than the other two algorithms as the amount of data increases. The BIRCH+K-means algorithm is appropriate for clustering large scales sets of data.

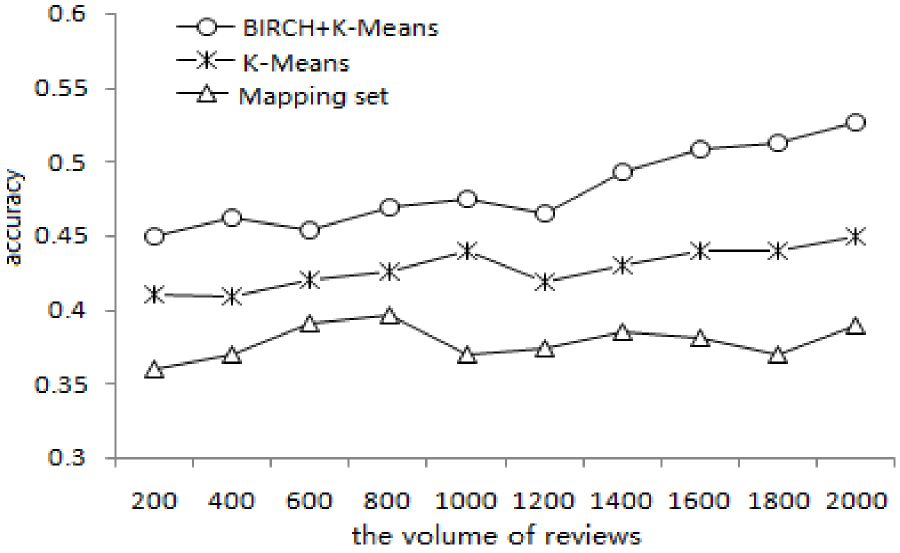

Figure 6. The comparison of clustering accuracy by three methods

6. Conclusion

In this paper, we extract the explicit evaluation object and evaluation words by matching the evaluation texts based on the noun phrase pattern. And then establishes the implicit evaluation by establishing the semantic mapping set of the evaluation object and the evaluation terms. Combining the BIRCH clustering with K-Means clustering algorithm to re-cluster the evaluation object with different opinions. Because the BIRCH algorithm is a self-learning clustering algorithm, it can be used to pruning the evaluation object to delete some clusters contained few contents (ie, redundant evaluation objects). The experiment indicated the BIRCH+K-Means algorithm has the high effectiveness on precision, recall and F-value compare with other two algorithms. And the mentioned algorithm has the lowest time complexity especially for large scales sets of data.

References

[1] Chen, Q. et al. 2009. A Study of ReIation Extraction between Topics and Sentinents for Chinese 0pinioned Sentences. Proc. of CCIR ’2009 (2009), 504–512.

[2] Jingdong best-selling 284 notebook commentary data:

[3] Kim, S.-M. et al. 2006. Automatically assessing review helpfulness. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing - EMNLP ’06 (2006).

[4] Kim, S.-M. and Hovy, E. 2006. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text. (2006). DOI: https://doi.org/10.3115/1654641.1654642.

[5] Kobayashi, N. et al. 2005. Collecting Evaluative Expressions for Opinion Extraction. Journal of Natural Language Processing. (2005). DOI: https://doi.org/10.5715/jnlp.12.3_203.

[6] L, K. et al. 2011. Optimization for BIRCH pre-clustering algorithm Computational Intelligence and Informatics. 2011 IEEE 12th International Symposium (2011), 475–480.

[7] Nasukawa, T. and Yi, J. 2003. Sentiment Analysis: Capturing Favorability Using Natural Language Processing. Proceedings of the 2nd international conference on Knowledge capture. (2003). DOI: https://doi.org/10.1145/945645.945658.

[8] Popescu, A.M. and Etzioni, O. 2007. Extracting product features and opinions from reviews. Natural Language Processing and Text Mining.

[9] Shao, F. et al. 2004. BRICH Clustering Algorithm with Mult_Threshold. Computer Engineering and Applications. 40 (12), (2004), 174–176,195.

[10] Stoykova, V. 2017. Extracting Academic Subjects Semantic Relations Using Collocations. 4, 1 (2017), 3–6.