International Journal of u- and e- Service, Science and Technology
Vol.9, No. 2 (2016), pp.219-228
http://dx.doi.org/10.14257/ijunesst.2016.9.2.22

A Bootstrapping Based Evaluation Object Extraction in Comments on Chinese Commodities

Geng Yushui, Zhang Lishuo and Sun Tao

School of Information, Qilu University of Technology
Jinan250353, China
gys@spu.edu.cn

Abstract

Currently product reviews information has become the focus of increasing concern to these comments often containing more information on consumer evaluation of various attributes of goods, and therefore the evaluation of mining comment object has become a hot topic. This paper presents a method called evaluation object extraction based on Bootstrapping in Chinese commodity comments, this method first comment summed up a lot of information on the properties of 10 kinds of speech rule templates, and then determined the integrity of the candidate meets these templates phrases, and these have complete of candidates as a seed, extended new property template by Bootstrapping method, and finally removed in line with all the attributes of a template as a final evaluation of the object phrase, and through experiments to calculate the accuracy, harmonic mean of recall and prove that the proposed method improves extraction accuracy properties of the object.

Keywords*: Chinese Product Reviews; Bootstrapping; Evaluation Object; Candidates*

1. Introduction

With the development of network technology and the popularity of e-commerce, more and more consumers began to express their comments on related products on e-commerce related websites, and consumer habits are beginning to see before buying goods on e-commerce sites others related comments. However, due to the different users' personal preferences and personal concerns, their comments are often very different, and users often express the views of the product multiple properties in a comment, additional comments are generally large numbers, if you would like to manually find themselves concerned about the content of these comments from time-consuming. So you want to extract these views unstructured product reviews, especially in view of each attribute describing the product become more popular topic.

To purchase mobile phones, for example, consumers tend to reference before buying comment information on the network, such as comments Zhongguancun Science Park website: "The phone style is quite new with high resolution, I am very satisfied, but it is easy to hang very annoying ", " Lenovo's mobile phone price is high, but not as good as Apple's appearance in these high-end handsets look good, quality to be the test.", " note3 pixels high, the battery can make a very long time, that is, the phone is easy to get stuck, this makes me very anxious…". These comments have been described inside each comment multiple attributes of goods, if artificial selection is very easy to determine which is what we want, but due to the number of comments appeared are generally in large numbers, it can not meet the needs of us, so you need to find a suitable way to automatically extract the needed perspective.

In recent years, the main extraction methods in domestic and foreign research are mainly divided into supervised andunsupervised methods. Zhao et al [1]. presented at the 2006 level based on hidden Markov model product named entity recognition method,

ISSN: 2005-4246 IJUNESST

which works well with the nested sequence of multiscale problems; Arun et al [2] proposed a species-based extraction method to generate a relational database. Mannai et al [3] used Bayesian network approach to achieve extraction work. Gamon M, et al [4] used tf-idf to get attribute word, and the word property was classified into general and special attributes words. Qiu et al [5] proposed a feature extraction emotion words and a small amount of seed emotion based word segmentation method, but this method is not based on considerations property. In addition, this method, not considering automatic extraction template, is artificially defined templates, scalability is very restricted. In the feature extraction and filtration process, word frequency considered only, without considering the relationship between intimacy feature words and templates. Due to the current accuracy, coverage and portability of product reviews attribute word extraction will be further improved, it is necessary to conduct in-depth study.

2. The Paper Work

Evaluation research reviews object contains objects in the extraction and analysis of their propensity to two steps, but due to the current product reviews often contain a lot of clutter content, so the evaluation target product extracted from the clutter of information becomes the study of the more important step, this paper studies the Chinese product reviews in the evaluation object extraction problem, this paper presents a method called evaluation object extraction based on Bootstrapping in Chinese commodity comments, since Bootstrapping only a small amount in the initial stages of learning data, so that not only significantly reducing the reliance on experts, but also greatly enhance the versatility and portability of algorithms, this paper chooses this method to study the object extraction. First select some comments as experimental data sets, its word and stop words [6] removal, artificial data set selected candidate speech combination rule noun phrase, the phrase comments satisfy the rules extracted, and then judge the integrity of the candidate phrases, and the integrity phrase would be called candidates. Due to the limited definition of artificial rules, select a property and not a good phrase to show all of the properties mentioned in the comments, so this selection will be screened by the candidate noun phrases through Bootstrapping method [7] to expand its properties words, thus avoiding the limitations of artificial selection attribute word. The experimental results to calculate the accuracy, recall and harmonic mean judge performance of this method.

System frame shown in Figure 1 herein.

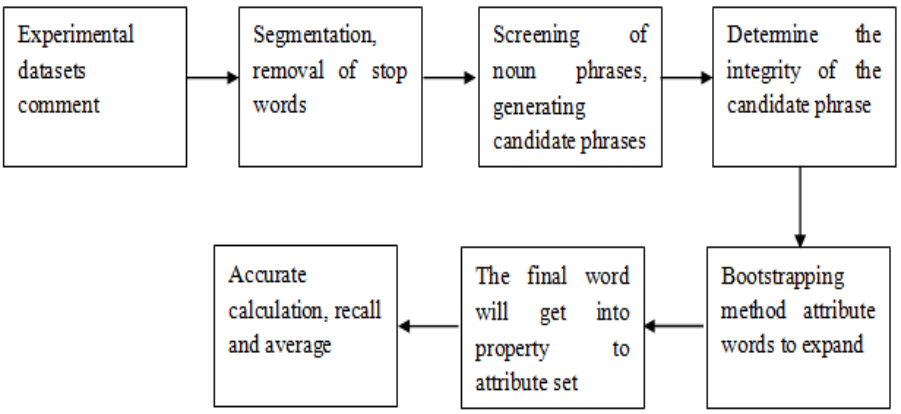


Figure 1. In this Paper, the System Framework

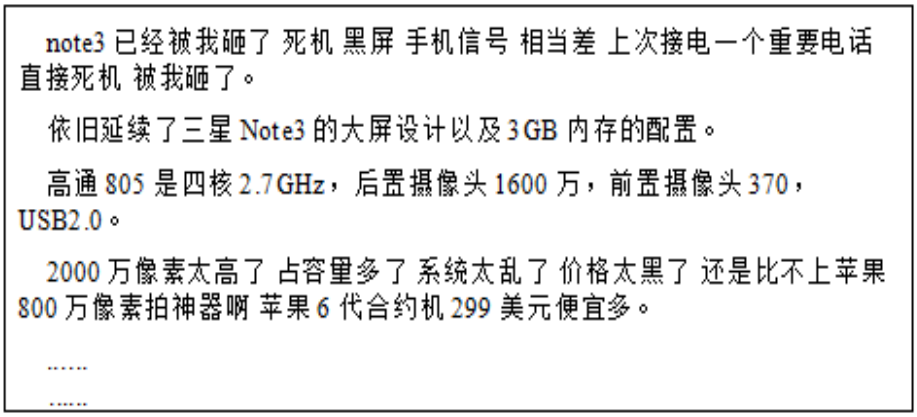
3. Candidate Evaluation Object Extraction

By analyzing a large number of comments commodity information, the evaluation found that the object of general merchandise noun or noun phrase for the article [8] also proved a noun or noun phrase as the evaluation object is feasible, so this is according to the CAS segmentation software ICTCLAS [9] for reviews conducted segmentation and mark the corresponding part of speech, speech summed up the rules to be evaluated in 10 groups, to comply with these rules of speech noun phrases called candidate phrases and phrases to determine the integrity of these candidates, to have integrity evaluation of candidate phrases called candidate object.

3.1. Screening Generate Candidate Noun Phrase Phrases

For extracted from the literature [10] provides digital text commentary 14 799 parts of the text, use ICTCLAS text segmentation, candidate evaluation object extraction rule based on the presence of speech tagging. From the perspective of the recall, the label text should be collected in the evaluation object of speech rules as much as possible. But from the perspective of complexity and feasibility analysis, speech rules should be better. In this paper, based on the above 2 points, comprehensive rules of part of speech frequency, eventually developed the following 10 group for commodity comment object noun phrase rules of parts of speech: n, n+n, n+n+n, n+v, v+n, a+n, n+a, x+n, n+x, x+n+v.

To demonstrate the applicability of these rules, the paper picked a lot of product reviews phrases, through the analysis to these comments phrases, the evaluation found that most commodity objects that meet these rules, Figure 2 is taken by some experimental datasets comment artificial selection.(Note: This article is for the Chinese comment to study, so the experimental data set is Chinese.)


note3 已经被我砸了 死机 黑屏 手机信号 相当差 上次接电一个重要电话直接死机 被我砸了。

依旧延续了三星 Note3 的大屏设计以及 3GB 内存的配置。

高通 805 是四核 2.7GHz，后置摄像头 1600 万，前置摄像头 370，USB2.0。

2000 万像素太高了 占容里多了 系统太乱了 价格太黑了 还是比不上苹果 800 万像素拍神器啊 苹果 6 代合约机 299 美元便宜多。

……

……

Figure 2. Experimental Comment Dataset (Section)

Figure 3 is an article by the Chinese Academy of segmentation software ICTCLAS experimental dataset segmentation, removal of stop words and speech after tagging section shows.

note3/x 已经/d 被/p 我/r 砸/v 了/u 死/v 机/n 黑/a 屏/n 手机/n 信号/n 相当/d 差/a 上次/t 接/v 一个/m 重要/a 电话/n 直接/a 死/v 机/n 被/p 我/r 砸/v 了/y 。/w

依旧/z 延续/v 了/u 三星/n Note3/x 的/u 大/a 屏/n 设计/v 以及/c 3GB/x 内存/n 的/u 配置/v 。/w

高/a 通/v 805/m 是/v 四/m 核/n 2.7GHz/x ，/w 后/f 置/v 摄像头/n 1600/m 万/m ，/w 前/f 置/v 摄像头/n 370/m ，/w USB2/x ./w 0/n 。/w

2000/m 万/m 像/v 素/a 太/d 高/a 了/u 占/v 容里/n 多/a 了/u 系统/n 太/d 乱/v 了/u 价格/n 太/d 黑/a 了/u 。/w

手机/n 屏幕/n 虽然/c 大/a ，/w 但是/c 很/d 耗电/v ，/w 电池/n 撑/v 的/u 不/d 长/a ，/w 并且/c 散热/v 不好/a 。/w
耗电/v 太/d 多/a ，/w 但是/c 手机/n 系统/n 很/d 棒/a ，/w 我/r 很/d 喜欢/v 。/w 只有/c 偶尔/d 会/v 出现/v 死/v 机/n 还/d 有/v 黑/a 屏/n 现象/n ，/w 这/r 是/v 怎么/r 回/v 事/n ？/w

Figure 3. Experimental Comments Dataset Speech Tagging (Part)

As can be seen by the above experiment 10 set of rules selected herein is applicable. After the experimental data set in line with the rules of these types of speech phrases are filtered out, we will find a lot of noise words in it, it can neither help us analyze the properties mentioned in the comments, but also interfere with our results, which when they need to be filtered out, leaving the final needs of candidates. At this article will do these candidate noun phrases integrity judge, we will have the integrity of the candidate noun phrases as candidate for next steps in the operation.

3.2. Candidate Object Extraction.

The resulting candidate noun phrases judge their integrity. First we define in the comments A=a1 a2 ... ai ... an, where ai represents the i-th word in the comments of A. Assuming candidate B appears in which n different positions b1, b2, ... bi ... A, in a review, bn. If there is at least one set of <i, k> ($1 \leq i < k \leq n$), making A first bi-1 bk -1 words and words are not the same, then this is called B left intact (if B is the first text in a word it must be left intact); if there is at least one group <i, k> ($1 \leq i < k \leq n$), such that a, the first bi+B words and bk+B words are not the same, then B is called the right of this complete (if B is the last word in the text, it must be right complete); if B is right both left intact intact, then this B evaluation of the integrity of the object is called.

The example in Figure 4.Which conform to the rules drawn objects to be evaluated according to the rules, as follows: “note3/x battery/n”, “battery /n radiating/v”, “note3/x battery /n radiating/v”and so on. “note3/x battery /n” appears twice, first a word behind are “radiating /v”, So, this is not the right complete evaluation object; Two occurrences of “battery /n radiating /v” in front of the word are “note3/x ”, therefore, the evaluation object is not left intact; Two occurrences of “note3/x battery /n radiating /v ”around the word are not the same, so the object is the result of this evaluation have integrity, the evaluation of the real object of this article is to be extracted. Meaning that the definition of integrity, to avoid one-sided assessment drawn objects or properties words. Just want to extract “screen /n colors /n”instead of “colors /n”, you can pass the integrity of the non-filtered to remove the “screen” retain “screen colors”.

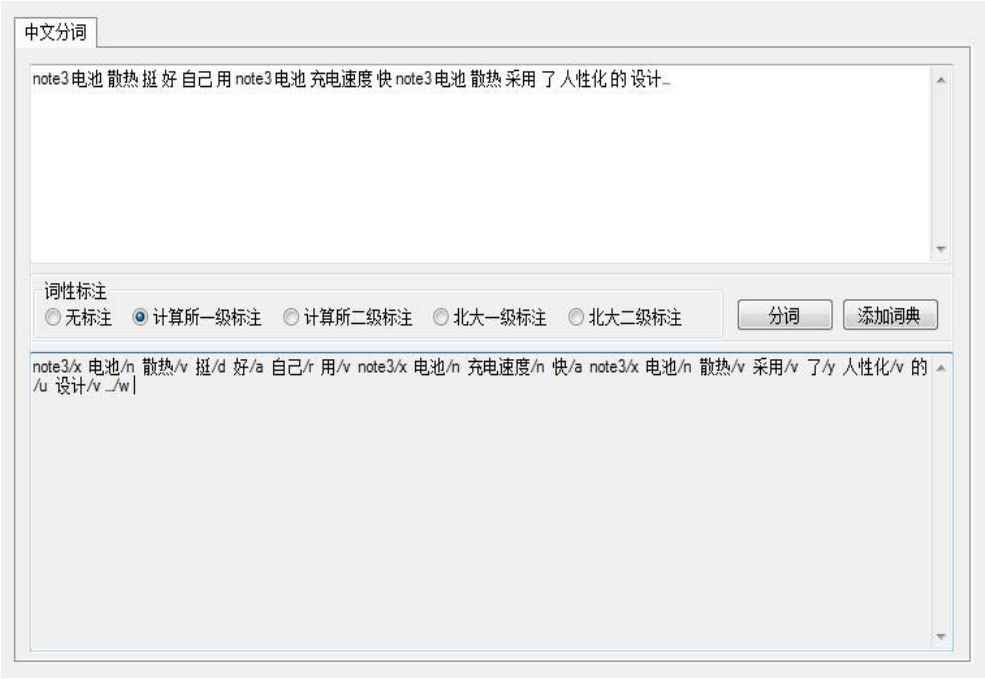


Figure 4. Word Examples

Due to the limited candidates are not well exhibit a review of all the evaluation target, and therefore need to use Bootstrapping method in these limited on the basis of the evaluation object to extend candidate attributes, in order to more fully show the review of all objects.

4. Based on the Evaluation Object Extraction Method Bootstrapping

Bootstrapping [11], that is self-expanding, is a semi-supervised learning methods. The method is by manual intervention to get the seeds, and then automatically increment iterative training until convergence. In each round of iteration, it will have a new dimension of data, with these new labeling data retrain the model, which in turn can generate new data, and so on ad infinitum, until the end of the final convergence. Bootstrapping methods adopted to extract object, easily lead to the theme of this paper, therefore offset each iteration process, not just a statistical frequency of occurrence of each object, but to use the word attribute and relationship data sets intimacy calculate a score for each candidate several pre-made object to leave the high score into the set of objects.

Bootstrapping method evaluation object extraction process:

1. Selecting the part of speech in line with the rules of Section 2 of the candidate as the initial study evaluated data sets.
2. The data on the learning data set score using the following formula Rating [12].
3. Selecting the first five data getting the highest score to add to the data set (data selected scores must be greater than the predetermined threshold value β).
4. From the experimental data set in the new part of speech rules randomly selected template again, repeat steps 2 and 3 operations, until that there haven't be the evaluation object template for new qualifying.
5. Pick up the qualifying template phrase, put it as a final evaluation of the object to be extracted.

Bootstrapping method used in the process of scoring formula is:

$$Score(n) = (Score_{pjc}(n) + Score_{e-s}(n) + Score_{p-s}(n) + Score_{m-s}(n))/a \quad (1)$$

Among them, $Score_{pjc}(n)$ indicate that the number of candidates to be evaluated before and after evaluation of the ten positions contain the word adjacent evaluate information. $Score_{e-s}(n)$ represents words (phrase) support, namely the number of words or phrases appear in the corpus. $Score_{p-s}(n)$ represents pure support [13], referring to the candidate to be evaluated as a noun or noun phrase appears in the sentence, and the sentence is no longer contains the number of sentences other candidates to be evaluated. $Score_{m-s}(n)$ presents template support that candidate evaluation template object is extracted from the corpus of times out. Refers to a factor of 4 paper selected threshold β is set to 200.

Brief code Bootstrapping method is as follows:

Input: A candidate set B = integrity evaluation object {B1, B2, ... Bn};

Comments preprocessed experimental data set A = {A1, A2, ..., Am}; $B \subset A$.

Output: Comment text evaluation object set C .

Process:

1. The data set in the data B by scoring {b1, b2, ... bi ... bn}
2. For （i=1, i<=n, i++）
3. For (j=1, j<=5, j++)
4. If (bi>bi+1)
5. dj=bi
6. Else dj=bi+1
7. End if
8. End for
9. End for
10. {d1, d2, ..., d5}→The initial learning of the data set D.
11. Then the g data from A collection, and then calculate the score.
12. While (ai >= β) and （Speech rules extracted data representative $\not\subset D$）
13. For （i=1, i<=g, i++）
14. For (j=1, j<=5, j++)
15. If (bi>bi+1)
16. dj=bi
17. Else dj=bi+1
18. End if
19. End for
20. End for
21. {d1, d2, ..., d5}→Learning data set D（At this point then the above data back

padding.）

22. End while

23. Will set A compliance with set out D combination.→The final object set E.

5. The Experimental Analysis

Experiments for mobile e-commerce sales reviews research paper from Zhongguancun online in crawling the three groups of users comment on a Android phone, the number of each group respectively for 2180, 1920, 1563 as the experimental data set data1, data2, data3. Experimental environment is Microsoft Visual Studio 2010 platform, using the C++ language, first of all the three datasets were developed by Institute of Computing Technology of Chinese lexical analysis system ICTCLAS its segmentation, POS tagging, and then use special stop list its filtration, concentration of the processed data in line with the rules of section 2 of speech phrases into a test program, the final results will calculate the exact rate, recall rate, and average to determine the performance of this method.

$$P(accuracy) = A / B \quad (2)$$

$$R(recall) = A / C \quad (3)$$

$$F - measure = 2PR/(P + R) \quad (4)$$

Wherein P means the accuracy, and R means the recall. A formula 2 refers to the number of extracting feature words right, B refers to the number of withdrawn all words. A number of formulas 3 refers to the number of extracting feature words right, C standard answer feature words.F-measure refers to the harmonic mean.

5.1 Experiment 1: The Results Compare to Data1 were not Used this Method and Evaluated after the Integrity of the Candidate Phrase

The data1 were not used this method and evaluate the integrity of the candidate phrase to compare the final results, Table 1 shows some experimental results.

Table 1. Part of the Results of Experiment Two

Property	This paper method			Do not use the phrase candidate integrity		
	Accuracy	Recall	The harmonic mean	Accuracy	Recall	The harmonic mean
Battery(电池)	0.801	0.702	0.748	0.702	0.700	0.702
Android system(安卓系统)	0.723	0.687	0.705	0.693	0.701	0.693
Front camera(前置摄像头)	0.832	0.690	0.754	0.798	0.660	0.798
Phone signal(手机信号)	0.800	0.720	0.758	0.711	0.697	0.711
Cooling effect(散热效果)	0.767	0.801	0.784	0.723	0.782	0.723
Screen Resolution(屏幕分辨率)	0.757	0.760	0.758	0.765	0.722	0.765

You can see if the phrase is not used to determine the integrity of the candidate by the experimental results do not showing a good evaluation of the object, its precision and recall rate of this method is better high.Do not tend to make more use of intactness judgment clutter phrases into the extraction to be evaluated, resulting in decreased accuracy of the experimental results, and before carrying out extraction of the object if the first evaluation of these phrases integrity determination, it not only can the filter out useless phrase, but also can choose the more important information phrase.Therefore, it can be seen through this set of experiments to determine the importance of the integrity of the candidate phrase, we in order to prove the feasibility of this approach, the paper turn into a program data2 experiment, harmonic mean of the results shown in Figure 4 below compares, where harmonic mean 1 is the result of this article refers to the method, harmonic mean 2 is not used method results integrity.

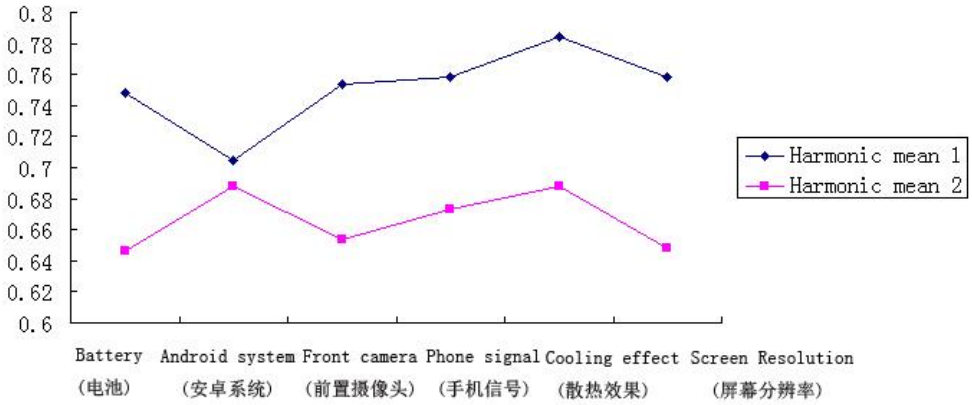


Figure 5. The Results of this Method and Do Not Use the Phrase Candidate Integrity Compare the Harmonic Mean

As can be seen from the above two sets of experimental results, the proposed method is feasible and determine the integrity of the candidate phrase accuracy of the experimental results to improve a great help, but because the program needs to evaluate each candidate phrase all its integrity, and therefore the time complexity of the program high in efficiency to be improved.

5.2 Experiment 2: Compare this Method with Keyword Matching Methods Evaluated Score Results

Select the experimental data set data3, we made to conduct the evaluation of this method to extract the object, then keyword matching method to extract the evaluation object with its harmonic mean comparison.Keyword matching method used to extract the evaluation object is still experimental environment Microsoft Visual Studio 2010 platform, using C++ language.The results shown in Figure 5, which refers to the harmonic mean a result of this method, the value 3 is a method to reconcile the results of the evaluation keyword matching.

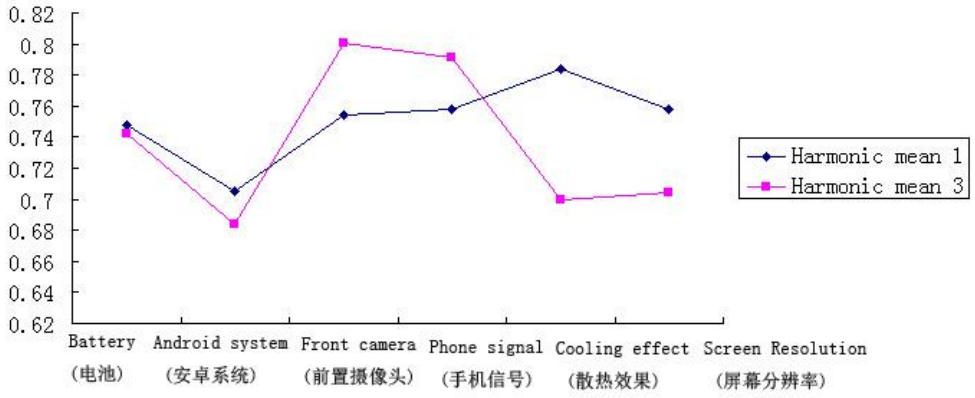


Figure 6. Harmonic Mean Compare this Method with Keyword Matching Method

As can be seen from the experimental results, keyword matching method to reconcile the average of fluctuations on the part of the evaluation object extraction is very accurate, but for some object extraction error is greater, is not stable, and the method of this paper, although the precise the degree to be improved, but not so great before and after the error and downs, if there is a lot of ups and downs will lead to incomplete extraction of the object, the object can not be fully described in the text extraction, so at this point this method can still be selected.And because this method involves the integrity of the evaluation, a little high time complexity of the algorithm than keyword matching method, this point remains to be to continue to improve.

6. Conclusion

Comments can effectively extract most of the evaluated experimentally proved to be evaluated based on Bootstrapping extraction in Chinese reviews of the proposed method, the accuracy, the recall and the harmonic mean all have greatly improved, you can use this method to extract evaluated in order to help consumers understand the information on more goods.

The results can be seen by this method is not able to extract all of the evaluation object one by one out, so the next step needs to be improved in order to increase the accuracy of the method results. At the time of this article did not consider the experimental optimization problem, so the experiment is the next step should be to improve the efficiency of the place.

References

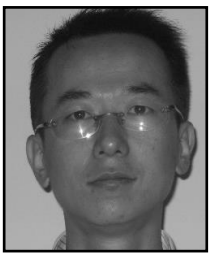
[1] X. Song, S. Wang, H. X. Li, “Research on Comment Target Recognition for Specific Domain Products, Journal of Chinese Information Processing, vol. 20, no. 1, **(2006)**, pp. 17220.
[2] A. Arun and P. Srinivasan, “Automated query generation of Rdbms for information and knowledge extraction”, Proceedings of 2013 International Conference on Information Communication and Embedded Systems, **(2013)**; Chennai.
[3] M. Mannai, B. Abdessalem and W. Karaa, “Bayesian information extraction network for medline abstract”, Proceedings of 2013 International Conference on Computer and Information Technology (WCCIT), **(2013)**; Sousse.
[4] M. Gamon, A. Aue and S. Oliver, “Mining customer opinions from m tex”, Proceedings of the 6th International Symposium on Intelligent Data Analysis, **(2005)**; Madrid, Spain.
[5] G. Qiu, B. Liu and J. Bu, “Opinion word expansion and target extraction through double propagation”, Journal of Computational Linguistics, vol. 37, no. 1, **(2011)**, pp. 9-27.
[6] J. Macqueen, “Some methods for classification and analysis of multivariate observations”, Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, **(1967)**.

[7] Y. Duan, W. Zhu and Q. Chen, "Semantic Annotation of Species Description Text in Chinese by Combining Naive Bayes Algorithm with Bootstrapping Method", New Technology of Library and Information Service, vol. 5, no. 5, **(2014)**.
[8] T. He, B. Wen and L. Song, "Emotion tendentiousness recognition and extraction research.", Proceedings of the COAE, **(2008)**; Harbin.
[9] ICTCLAS Project Team, "ICTCLAS Chinese word segmentation system", EB/OL, http://ictclas.org/, **(2009)**.
[10] H. Xu, L. Sun and T. Yang, "Third Chinese propensity analysis evaluation summary report", EB/OL, http://ir.sdu.edu.cn/ccir2011/coae 2011_register.html, **(2010)**.
[11] H. Wang and H. Chen, "Feature extraction method based on Bootstrapping in English product comment", Journal of Shandong University (Natural Science), **(2014)**.
[12] X. Song, S. Wang and H. Li, "Research on Comment Target Recognition for Specific Domain Products", Journal of Chinese Information Processing, vol. 1, **(2010)**.
[13] Y. Zhao, H. Liu and B. Qin, "HIT_ IR_OMS : Sentiment Analysis System", Proceedings of the COAE, **(2008)**; Harbin.

Authors

Geng Yushui, He is a professor of School of Information, Shandong Polytechnic University. Currently he is serving as the director of School of Information. His research interests are Cloud Computing, Enterprise Business Processes System and the Internet of Things.He has led more than 10 research projects.

Zhang Lishuo, She is a student of Shandong Polytechnic University in School of Information. His research interests include big data and data mining.



Sun Tao, He is a PhD, Associate Professor, his current research interests focus on big data, data integration and cloud computing.

