Wikipedia Based Approach for Clustering Keyword of Reviews

Ming Jiang[1, 2], Wencao Yan[1, 2], Xingqi Wang[1, 2], Jingfan Tang[1, 2], Chunming Wu[3]
(1. Institute of Software and Intelligent Technology, Hangzhou Dianzi University, Hangzhou, 310018, China)
(2. Zhejiang Provincial Engineering Center on Media Data Cloud Processing and Analysis, Hangzhou Dianzi University, Hangzhou, 310018, China)
(3. College of Computer Science and Technology, Zhejiang University, Hangzhou, 310027, China)
Email: jmzju@163.com, yanwencao@gmail.com, xqiwang@163.com, tangjf@hdu.edu.cn, wuchunming@zju.edu.cn

***Abstract*—A novel method based on Wikipedia for clustering keyword of reviews is proposed. Users can quickly finding the themes they interest through it. First the method extracts keywords, then calculates word similarity based on Wikipedia to generate similarity matrix, finally uses k-means to cluster. The performance is better than the methods which based on How-net and Word-net. The accuracy is around 77%.**

***Index Terms*—Wikipedia, keyword of reviews, semantic similarity, similarity matrix, cluster,**

I. Introduction

With the development of Internet, more and more people like to surfer on the Internet and leave a large number of reviews online. Such as the famous Internet shopping mall Amazon, thousands of customer reviews are leaved on popular products. It is a big problem that how a user can find the comments he interests among a lot of reviews. If we cluster reviews via the associate of subject, by this way the users can filter reviews conveniently and avoid browsing unwanted information.

With the progress of information communication especially the Internet as an emergence of new media, we have already got rid of the shackles of poor information and came into an information-rich society. Modern society is that of an information explosion, the source of information is no longer a problem but how to get it fast and accurately is a major issue. Currently a variety of technologies such as information retrieval, information filtering and information extraction are deployed around this purpose. As the network is too large, much of the related information usually is scattered in many different places and different times. General information retrieval tools are based on keywords; the results have a high degree of information redundancy. A lot of irrelevant information is returned just because it contains the same keyword.

Clustering keyword of reviews[1] is a challenge work. For example, on the one hand car and bicycle can be clustered in the same class because they all belong to transport. On the other hand, they can be divided into different classes because one belongs to Motor vehicle and the other one belongs to non-motorized. In other words, all words should be considered when clustering keyword of reviews. Therefore Similarity matrix can contribute to the clustered performance.

A method which based on Wikipedia for clustering keyword of reviews is proposed in this paper. General steps are as follows: first, extracting keyword of reviews. Second, building word similarity matrix, a method based on Wikipedia to calculate semantic similarity. At last, according to the word similarity matrix, k-means algorithm is used to cluster keyword of reviews. It solves a non-supervised clustering keyword problem. Experimental results show that our method performs better than existing methods.

The remainder of this paper is organized as follows. Section II reviews related work in the area of word clustering. Section III presents the detail of the method which based on Wikipedia for clustering keyword of reviews. In Section IV, we describe the experiment and evaluate the results, followed by conclusion in Section V.

II. Related Work

In the word extraction Minqing and Bing Liu[2] identified product-features using association rule mining. Peter D.Turney[3] tagged words with part of speech and extracted words with rules which based on English grammar. Such as the rule: “the first word is an adjective, the second word is a noun”. Our method is different from theirs. First, the method which we used is based on Chinese grammar and word frequency. Second, the goal is not the same. Our method aims to get the keyword of reviews, but they want the product-feature or emotional word.

Word clustering as a basic work of natural language process has a widely range of applications. It plays an important role in semantic disambiguation[4], subject extraction[5], text classification and information retrieval[6]. In semantic-based word clustering, Qun Liu and Sujian Li[7] proposed a method using How-net to calculate semantic similarity. How-net is a tree of sememe hierarchy. Through this system we can calculate similarity between words conveniently. In pragmatic features based word clustering, Evgeniy Gabrilovich and Shaul Markovitch[8] identified word similarity based on Wikipedia. Andrea Moro and Robert Navigli[9] also

doi:10.4304/jsw.9.9.2246-2250

presented a method. In the Evgeniy Gabrilovich and Shaul Markovitch's method which regarded each article title in Wikipedia as a concept and creates tfidf[10] model. The value of each dimension in word vector represents the tfidf value between word and corresponding article. Then the word similarity can be calculated by word vector. Yutake Matsuo[11] et al used the results returned by search engine to cluster words. They counted the number of target words co-occurrence in websites, then calculated the word similarity on basis of it. Wang[12] et al tried to use bilingual parallel corpus for word clustering, their method dialectic opinion of the combinations of a statistical translation model and a mutual information clustering algorithm. Jerome R.BeUegarda[13] et al used the resources of document classification to word clustering. The documents were collated and classified before, the co-occurrence of words expanded to the document class. The document class as the feature-dimensional vector represent of the target word. On the one hand it can alleviate the data sparseness problem, on the other hand it reflects the distribution of the target word in particular document set. Guihong Cao[14] et al brought fuzzy clustering algorithm into word clustering. The fuzzy clustering model attached to each word in multiple degrees, in order to reflect the word ambiguity. There are also many other clustering methods[15][16][17]. The approach proposed to calculate word similarity in this paper is also based on Wikipedia. Our method is different from others on two main aspects. First, our method represents the word by similarity matrix, but Evgeniy Gabrilovich and Shaul Markovitch builded word vector space model. The experimental results show that similarity matrix is much better. Second, this paper processes Chinese while others' methods handle English. Obviously, Chinese and English are quite different.

III. The Proposed Techniques

Figure 1 shows processes and framework of the system. The system of keyword clustering contains five main steps. 1, part-of-speech tagging; 2, extracting keywords; 3, semantic similarity model based on Wikipedia generated; 4, similarity matrix of words generated; 5, using k-means algorithm to cluster keywords;

A. Part of Speech Tagging

Keyword is generally represented by noun or noun phrase, so part-of-speech tagging is necessary. Chinese lexical analysis system ICTCLAS provided by Chinese Academy of Sciences is used to tag reviews. The words are labeled as noun, adverb, adjective and so on. The following example shows a labeled review. "appearance /n beautiful /a quality /n feel / n nice / a". Each word is marked out its part of speech. Each review is labeled by Chinese lexical analysis system and stored in one line. Then preprocess the tagged reviews and delete unnecessary content including removing stop words, particle and quantifier. After this treatment, reviews can be used for next step.

B. Extracting Keywords

According to Chinese expression, keywords are almost noun and there will be emotional adjectives nearby. Therefore, the nouns in table I are probably keywords. Since Chinese lexical system is not always right, so there will be some words are marked unreasonable. Such as "technology /n problem /n" actually it should be as a noun "technology problem /n". The patterns in table II can distinguish noun more accurately therefore it can improve the system's performance.

Then word frequency is counted. According to reviews' characteristics, a keyword will be mentioned many times by different people so the low-frequency terms are removed. After the processes above, a set of nouns contains outlier is got. Finally removing some unrealistic terms by human, a set of keywords is presented.

TABLE I.
NOUN EXTRACTION MODE

Situation	Combination
1	noun
2	noun + noun
3	noun + noun morpheme

TABLE II.
KEYWORD EXTRACTION MODE

Situation	Combination
1	adjective + noun
2	noun + adjective
3	noun + adverb + adjective

TABLE III
WORD SIMILARITY BASED ON WIKIPEDIA MODEL GENERATION ALGORITHM

```
Input:
    D: article in Wikipedia
Output:
    vector space model of word
Steps:
    (1) Filtering the articles which are less than 300 words
    (2) Labeling the filtered articles
    (3) Building the dictionary including removing
        stop words and low-frequency words
    (4) Establishing vector space model of words.
        The value of each dimension in word vector represents
        the tfidf between word and corresponding article
```

C. Semantic Similarity Model Based on Wikipedia

We reference Evgeniy Gabrilovich and Shaul Markovitch's method to calculate word similarity. The difference has been described in Section II. The table III is described the algorithm.

There are many reasons lead us to choose Wikipedia as our corpus. First, Wikipedia is open to us and we can get it for free. Second, Wikipedia is well organized. Each article represents a theme so it is ideal for computing word similarity. Finally, Wikipedia is written by the people around the world which is consistent with people's daily expressions.

D. Similarity Matrix of Words Generated

During keyword clustering, we note that whether to cluster two words together not only related with similarity of the two words but also related with similarity between the two words and other words. For instance, a set contains four elements is going to be clustered into two classes. If the set is {apple, bananas, car, bike}, it should be divided into {apple, bananas} and {car, bike}. If the set is {motorcycles, tricycles, car, bike}, it should be clustered into {car, motorcycles} and {tricycles, bike}. Car and bike can be clustered into different classes in the different situations. Considering this case, similarity matrix is used to cluster words. Establishing similarity matrix for the set of keywords, the element a_{ij} in matrix represents the similarity between $word_i$ and $word_j$. The similarity between $word_i$ and $word_j$ is calculated through their word vectors' cosine similarity. The equation is as follow.

$$a_{ij} = \frac{word_i * word_j}{|word_i| * |word_j|} \tag{1}$$

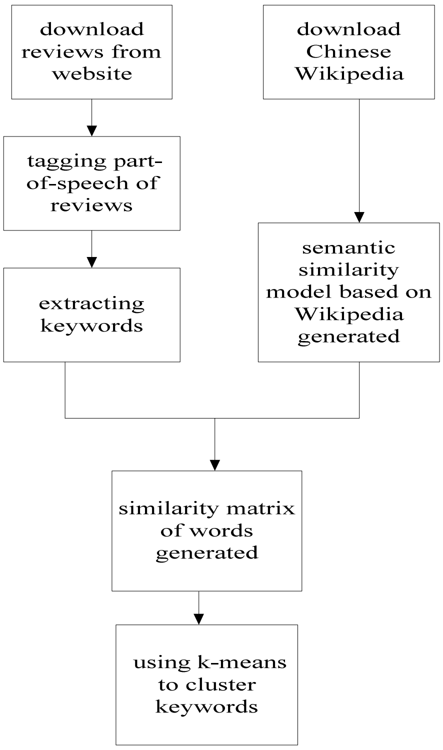

download reviews from website
download Chinese Wikipedia
tagging part-of-speech of reviews
extracting keywords
semantic similarity model based on Wikipedia generated
similarity matrix of words generated
using k-means to cluster keywords

Figure 1. Overview of project architecture and flow

E. Using k-means Algorithm to Cluster Keywords

After word similarity matrix is got, it's time to cluster the keyword of reviews. There are many clustering methods but most of them are not suitable for this case. The method based on hierarchical can't modify the data once it is split or merged. The density-based and model-based methods relate to the density of data, but the densities of different sets of clustering words are quite different. This paper adopts the method based on division, the classic k-means algorithm. Cosine similarity is used to measure the distance between words and criterion function of k-means algorithm.

IV. Experiments

A. Data

The experimental data used in this paper is showed in table IV. We random select three different areas of product review from Amazon and choose two different products in each area.

TABLE IV
THE INFORMATION OF REVIEWS

Area of reviews		Number of reviews	Number of keywords
Kitchenware	Rice Coolers	173	27
	Grinder	147	27
Electronic products	Computer	138	17
	Printer	127	29
Apparel	shirt	94	18
	shoes	122	21
Total		801	139

B. The Software Implementation

The software was carried out in the eclipse, using java programming language. We can use the buttons to load train data and test data. It will show total result in the white area. The detail of results will export as a table. The table can be open in office software. Following is the mean of each button. 'Load train data' (loading corpus such as Wikipedia corpus), 'train' (training the data loaded), 'Load test data' (loading data for testing), 'cluster' (clustering the words of test data), 'Export result' (exporting the detail of results as a table).

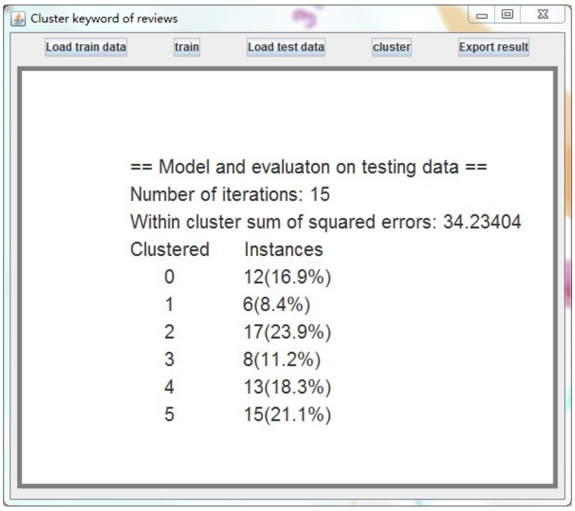
The interface can see from Figure 2.

C. Experimental Results

Besides based on Wikipedia and cosine similarity keyword clustering method (WSC) we use in this paper. We also use based on How-net and cosine similarity keyword clustering method (HSC) and based on Wikipedia and Chebyshev distance keyword clustering method (WCC) to compare with our approach. The results are showed in table V and Figure 3.

D. Analyzing Results

The results show that the method based on Wikipedia is better than the method based on How-net. The main reasons are as follows. First, Wikipedia's vocabulary is richer than How-net's. Second, How-net is organized by a group of experts while Wikipedia is written by the people around the world. Therefore, the Wikipedia is much better to express natural language. When calculating the distance between words, the results show that cosine similarity is better than chebyshev distance. Cosine similarity between word vectors is actuary a combination of similarity in each dimension of word vectors. It is close to the people judging similarity of words so that's the reason why cosine similarity performs well. According to the analysis above, we know the advantages of the method this paper proposed which led to perform better than other methods. Besides, the results are acceptable.



```
== Model and evaluaton on testing data ==
Number of iterations: 15
Within cluster sum of squared errors: 34.23404
Clustered    Instances
    0        12(16.9%)
    1        6(8.4%)
    2        17(23.9%)
    3        8(11.2%)
    4        13(18.3%)
    5        15(21.1%)
```

Figure 2 The software interface

TABLE V
THE RESULTS OF THREE METHODS

Area of reviews		WSC	HSC	WCC
Kitchenware	Rice Coolers	74.07%	70.37%	55.55%
	Grinder	77.77%	66.66%	77.77%
Electronic products	Computer	76.74%	47.05%	76.74%
	Printer	72.41%	68.69%	68.96%
Apparel	shirt	77.78%	55.56%	72.22%
	shoes	80.95%	52.38%	66.66%

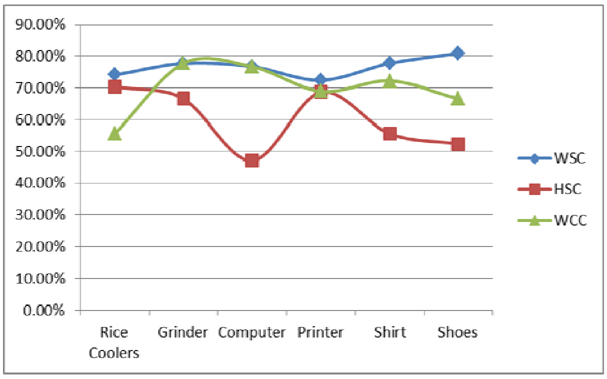


Figure 3. The Results of three methods

V. CONCLUSION

In this paper a method which based on Wikipedia for clustering keyword of reviews is proposed. It can divide reviews into different classes according the clustered keywords. Users can quickly finding themes they interest through this method. Its performance is better than the method based on How-net and Word-net.

Of course there is a lot of work we have to do to improve our method. For example the keyword of reviews not only can be a noun, it also may be a verb phrase or other forms. Another work is filtering Wikipedia article. If we can remove duplicate articles, it can reduce the word vector' dimensions at the same time improve the efficiency of the method.

ACKNOWLEDGMENT

This work is supported by the National High Technology Development 863 Program of China (No.2011AA01A107) and the Zhejiang Provincial Technical Plan Project (No. 2011C13008).

REFERENCES

[1] Marie Candito, Enrique Henestroze Angioano, Djame Seddah. A word clustering approach to domain adaptation. Proceedings of the 12th International Conference on Parsing Technologies. Pages 37-42.

[2] Minqing Hu, Bing Liu. Mining and Summarizing Customer Reviews. KDD'04, August 22-25, 2004, seattle, Washington, USA.

[3] Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, July 8-10, 2002. pp 417-424. NRC 44946.

[4] DAN T, RADU I, NANCY I. Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. Proceedings of the 20th international conference on Computational Linguistics Article No. 1312.

[5] Chen Jiong, Zhang Yongkui. Novel chinese text subject extraction method based on word clustering. Journal of computer applications 2005, 25(4).

[6] Saeedeh Momtazi, Dietrich Klakow. A word clustering approach for language model-based sentence retrieval in question answering system. Proceedings of the 18th ACM

conference on Information and knowledge management. Pages 1911-1914.
[7] Li Feng, Li Fang. A New approach Measuring Semantic Similarity in How-net 2000. Journal of Chinese information processing. 2007, 21(3).
[8] Gabrilovich, E. and S. Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. Proceedings of IJCAI, 1606-1611.
[9] Andrea Moro, Roberto Navigli. WiSeNet: building a Wikipedia-based semantic network with ontologized relations. Proceedings of the 21st ACM international conference on Information and knowledge management. Pages 1672-1676.
[10] Akiko Aizawa .The feature quantity: an information theoretic perspective of Tfidf-like measures. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. Pages 104-111.
[11] Yutaka Matsuo, Takeshi Sakaki, Koki Uchiyama, Mitsuru Ishizuka. Graph-based word clustering using a web search engine. Proceedings of the 2006 Conference on Empirical Method in Natural Language Processing. Pages 542-550.
[12] WANG Y Y, LAFFERTY J, WAIBEL A. Word Clustering with Parallel Spoken Language Corpora[C]// Proceedings of the Fourth International Conference on Spoken Language. 1996:2364-2367.
[13] JEROME R B, JOHN W B, Y L CHOW, NOAH B C, DEVANG N. A Novel Word Clustering Algorithm based on Latent Semantic Analysis[C]// Proceedings of the Acoustics, Speech, and Signal Processing. IEEE Computer Society, 1996:172-175.
[14] CAO G, SONG D, BRUZA P. Fuzzy K-Means Clustering on a High Dimensional Semantic Space [J]. Lecture notes in computer science, 2004.
[15] Xiaodan Xu. A New Sub-topic Clustering Method Based on Semi-supervised Learning [J]. journal of computers, vol.7 no.10 October 2012
[16] Yukai Yao. K-SVM: An Effective SVM Algorithm Based on K-means Clustering. Journal of computers, vol.8,no.10 October 2013
[17] Yu Zong. A Clustering Algorithm based on Local Accumulative Knowledge. Journal of computers, vol.8 no.2 February 2013

Ming Jiang He received the B.S. degree and M.S. degree in science in 1996 and 2001 respectively, and Ph.D. degree in Computer Science in 2004, all from Zhejiang University, China. He is currently a Professor in college of Computer Science, Hangzhou Dianzi University, China. His research interests include network virtualization, Internet QoS provisioning, and network multimedia processing.

Wencao Yan He received his BSc in Software Engineering from Hangzhou Dianzi University in 2012. Currently he is a master student in this university. His primary research area focuses on natural language processing and data mining.

Xingqi Wang He received his Bachelor and Master Degree from Harbin Institute of Technology in 1997 and 1999, respectively, and Ph.D degree from Zhejiang University in 2002. He is an associate professor in college of Computer Science, Hangzhou Dianzi University, China. His entire major are Computer Science. As a researcher, he visited CERCIA, University of Birmingham, UK from 2005 to 2006. His research interests include machine learning, data mining and multimedia content analysis.

Jingfan Tang He received the Ph.D. degree in Computer Science in 2005 from Zhejiang University, China. He is currently an Associate Professor in college of Computer Science, Hangzhou Dianzi University, China. His research interests include network virtualization, quality assurance, process improvement and legacy system re-engineering.

Chunming Wu He received the B.S. degree, M.S. degree and Ph.D. degree in Computer Science from Zhejiang University, China, in 1989, 1992 and 1995 respectively. He is currently a Professor in college of Computer Science, Zhejiang University, and the director of NGNT laboratory. His research fields include Network Multimedia processing, reconfigurable network technology, network virtualization and artificial intelligence.