

Textual opinion summarization based on Cluster_HITS model

Yancui Li¹, Hongyu Feng², Wenhe Feng^{2*}

¹ Department of Computer Science and Technology, Soochow University, Suzhou 215006, Jiangsu, China

² Henan Institute of Science and Technology, Xinxiang 453003, Henan, China

Received 1 June 2014, www.cmnt.lv

Abstract

Textual Opinion summarization aims to concentrate and refine the text data so as to generate a summary of the text regarding the expressed opinion. We collect and annotate an English multi-document corpus on product reviews, which provides a basic resource for the research on textual opinion summarization. Specifically, we incorporate the opinion and helpful information into the Cluster_HITS model to consider the impacts of them. Experimental results show that the proposed method apparently outperforms baseline in terms of ROUGE measurement.

Keywords: Opinion Summarization, Cluster_HITS Model, Opinion Information, Helpful Information

1 Introduction

E-commerce has changed our life style, many E-commerce sites, such as Amazon and Taobao, is the electronic commodity exhibition and trading platform, and allows users to comment on goods. These comments can not only provide shopping reference to potential users, but also help manufacturer analysis and understand the market response. The hot commodity often has hundreds of thousands of comments, which may include poor quality or even irrelevant comments. Reading these comments is time-consuming and laborious work. Text summarization can help users read quickly and efficiently. But text summarization focuses on scientific and technical literature and news et al. User comment texts are brief, subjective and have diversity styles, its structure is flexible and loose. Opinion Summarization is summarizing the views and emotion of users, in order to help users digest the emotional information of comment text. Opinion Summarization can help users better understand a lot of emotional information in Internet, and can provide support for the search engine, question answering system, topic detection and tracking etc.

The research of Opinion Summarization can be divided into two categories depending on the output: one category is outputting the features of the commodity, such as opinion target, opinion word, opinion holder and so on [1, 2]; the other is extracting a series of ordered sentences to represent comments [3,4]. At present there are few research and this paper mainly focus on it. The traditional study of opinion summarization is searches for the optimum sentence sequence by extracting and ordering sentences present in the input document set with high score calculate by opinions and the coherence score [5,6]. Actually, in comment text, sentiment is related with the topic of sentence. Recently sentiment analyses are focusing on the classification of the emotional tendency [7,8], and there is few research about summarization of the textual emotion.

Generally, there are two approaches to automatic summarization: extraction and abstraction. Extractive methods

work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In contrast, abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original. We use extraction approach in this paper.

Unsupervised method is the mainstream approach to Opinion summarization. In [9], the authors collect and annotate a Chinese multi document corpus on product reviews. Then, a novel PageRank framework to generate opinion summarization is proposed, with the advantage of considering both the topic relation and opinion relation among reviews. Experimental results on the corpus demonstrate that the proposed method substantially outperforms existing approaches in terms of ROUGE measurement. Reference the annotation method of [9], we annotate 30 topics of English comment corpus in this paper, then experiment using opinion and quality information based on Cluster_HITS model to Opinion Summarization.

2 Cluster_HITS Model

Hyperlink-Induced Topic Search (HITS; also known as hubs and authorities) is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg [10]. The idea behind Hubs and Authorities stemmed from a particular insight into the creation of web pages when the Internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that it held, but were used as compilations of a broad catalog of information that led users directly to other authoritative pages. In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs [11]. Authority and hub values are defined in terms of one another in a mutual

* Corresponding author's e-mail: wenhefeng@gmail.com

recursion. An authority value is computed as the sum of the scaled hub values that point to that page. A hub value is the sum of the scaled authority values of the pages it points to. Some implementations also consider the relevance of the linked pages.

The algorithm performs a series of iterations, each consisting of two basic steps:

Authority Update: Update each node's Authority score to be equal to the sum of the Hub Scores of each node that points to it. That is, a node is given a high authority score by being linked to pages that are recognized as Hubs for information.

Hub Update: Update each node's Hub Score to be equal to the sum of the Authority Scores of each node that it points to. That is, a node is given a high hub score by linking to nodes that are considered to be authorities on the subject.

In this paper, HITS model only consider the relation between sentence and cluster, we call it Cluster_HITS model. In this model, hubs is the topic center of cluster algorithm obtained, authorities is the sentence of text, as shown in FIGURE 1.

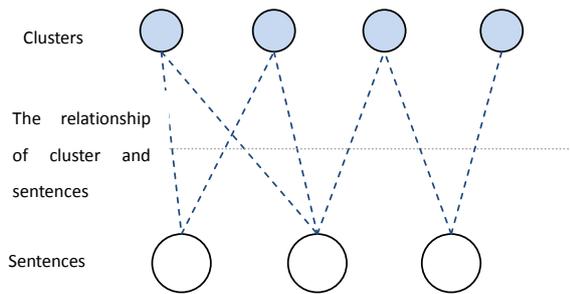


FIGURE 1. Cluste_HITS model.

In text summarization, start by building a directed graph $G = \langle S, C, E_{sc} \rangle$, S is the set of sentences of with some theme (authorities), C is the set of cluster hubs(hubs), $E_{sc} = \{s_{ij} \mid s_i \in S, c_j \in C\}$ presentment the relation of sentence and cluster. The value of e_{s_i, c_j} is depend on σ_{ij} , σ_{ij} is similarity of the sentence s_i and cluster c_j , we use cosine similarity algorithm to compute the similarity of s_i and c_j .

For iterator $(t+1)^{th}$, sentence s_i 's authority node value $Hub^{(t+1)}(s_i)$ and the hub of cluster node value. $Auth_{clusters}^{(t)}(c_j)$ are determined separately by the authority and hub node of iterator t^{th} , see the Equation (1) and (2)

$$Hub^{(t+1)}(s_i) = \sum_{c_j \in C} \sigma_{ij} Auth_{clusters}^{(t)}(c_j) \tag{1}$$

$$Auth_{clusters}^{(t+1)}(c_j) = \sum_{s_i \in S} \sigma_{ij} Hub^{(t)}(s_i) \tag{2}$$

In order to guarantee the convergence of iteration form, $Hub^{(t+1)}(s_i)$ and $Auth_{clusters}^{(t)}(c_j)$ are nominalized after each iterator as Equation (3) and (4).

$$Hub^{(t+1)}(s_i) = \frac{Hub^{(t+1)}(s_i)}{\sum_{s_i \in S} Hub^{(t+1)}(s_i)} \tag{3}$$

$$Auth_{cluster}^{(t+1)}(c_j) = \frac{Auth_{cluster}^{(t+1)}(c_j)}{\sum_{c_j \in C} Auth_{cluster}^{(t+1)}(c_j)} \tag{4}$$

For numerical computation of scores, the initial scores of all sentences and clusters are set to 1 and the above iterative steps are used to compute the new score until convergence. Finally, we get the sentence score $Score(s_i) = Hub(s_i)$, namely we use the authority scores as the saliency scores for the sentences. The sentences are then ranked and chosen into summary.

3 Textual Opinion Summarization based on Cluster_HITS

3.1 TEXTUAL OPINION SUMMARIZATION CORPUS

Because the corpus of opinion summarization is rare, so we collect 30 topics of product comments from amazon, each topic contain positive and negative comment. The comments include electronic product comments, book comments, movie comments and household item comments. Each Topic contains 500 texts comments, including the content of comment, the score of author and the vote information of other people for the comment, i.e., how many people think this comment is useful. The more people vote for the comment, the better quality of this comment.

The text must be segmented by sentence before automatic summarization. Then we choose 3 annotators dependently annotate text opinion summarization each topic. The criterion of extract summarization is choosing the comment sentence which opinion and content are apparent frequently. Form each topic we choose 120 words as summarization. Figure 2 gives the result of on annotator about "Kingston 8 GB Class 4 SDHC Flash Memory Card SD48GB".

The speed rating and its small capacity and its low price make this an ideal choice for picture frames and some of the less expensive cameras and camcorders. This gives us a massive amount of storage capacity and has worked without a hitch. Speed is good enough and works great. Nice and quick data transfer, love the plastic case it comes with, and the locking feature on the side of the card itself. A quality product at a good price. One thing you need to consider is that it's a SDHC card, that won't work on all devices and won't be read by a standard SD card reader, you'll need a SDHC reader. Great card, very reliable with more than enough space for occasional RAW shooting.

FIGURE 2. An example of our corpus.

After labelling, we have carried on the statistics to the corpus. Table 1 gives the compression ratio of statistical

results, and Table 2 gives some statistics of manual annotation.

In table 1, Original Sentence is the total number of 30 topics of the corpus. Marked sentence is the average number of manual annotation results in 30 topics'. Sentence compression ratio is made by marked sentence and Original sentence.

In Table 2, Author (123) indicates the person who annotated the English corpus. Total sentence is sentence numbers chosen by every author. Average sentence is the average number of sentence of these topics. Total words are total number of every author of the chosen sentences in these 30 topics. Average words are the average number of words in these topics.

Artificial abstract is very subjective and the annotation results certainly exists some subjective differences, because the annotation is made by people with different semantic understanding and knowledge background. If the annotation result is very different, it is controversial, subjective and little believable. Otherwise, it is little controversial, subjective and believable.

We used the ROUGE toolkit for evaluation, which has been widely adopted for automatic summarization evalua-

tion. It measures summary quality by counting overlapping units such as the n-gram, word sequences and word pairs between the candidate summary and the reference summary. ROUGE-N is an n-gram recall measure computed as Equation (5):

$$ROUGE - N = \frac{\sum_{S \in \{Re fSum\}} \sum_{n-gram \in S} Count_{match}(n-gram)}{\sum_{S \in \{Re fSum\}} \sum_{n-gram \in S} Count(n-gram)} \quad (5)$$

Where n stands for the length of the n-gram, and count match (n-gram) is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. Count (n-gram) is the number of n-grams in the reference summaries. ROUGE toolkit reports separate scores for 1, 2, 3 and 4-gram, and also for longest common subsequence co-occurrences. Among these different scores, unigram-based ROUGE score (ROUGE-1) has been shown to agree with human judgment most. We show three of the ROUGE metrics in the experimental results: ROUGE-1 (unigram-based), ROUGE-2 (bigram-based), and ROUGE-W (based on weighted longest common subsequence).

TABLE 1 The statistic results of corpus

original sentence	Marked sentence	Sentence compression ratio	number of words	Marked number of words	word compression ratio
250516	688	0.27%	1920973	11059	0.58%

TABLE 2 Manual annotation data statistics

author	Total sentence	Average sentences	Total words	Average words
Author1	238	7.93	3637	121.23
Author2	199	6.63	3276	109.2
Author3	251	8.37	4146	138.2

TABLE 3 ROUGE value of manual annotation

language	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU4
English	0.409	0.136	0.383	0.173	0.154

We use ROUGE toolkit* to measure the summarization performance, which is widely applied for summarization evaluation. It measures the quality of a summary by counting the unit overlaps between the candidate summary and a set of reference summaries. The full explanation of the evaluation toolkit can be found in [12]. In general, the higher the ROUGE scores, the better summarization performance. Table 3 give the performance of human annotators by ROUGE value.

3.2 FRAMEWORK OF OPINION SUMMARIZATION

In this section, the experiment uses Cluster_HITS to carry out the research. This paper will combine the sentence information, opinion information and helpful information through the model as Figure 3. The model considers the relationship between sentence, clustering, the emotional relationship of sentences and quality information of the sentence. In this model, the upper layer is the extension, including opinion information and review of the helpful

information, the middle layer is a sentence level, and the bottom is the clustering layer. In this model, clustering center, opinion information and review of helpful information is the central node (hubs) and the sentence is the authoritative node (authorities).

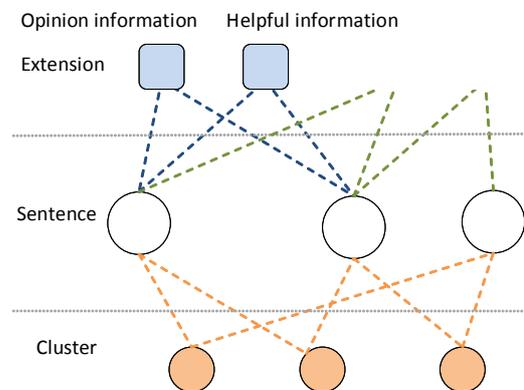


FIGURE 3 Cluster_HITS model with emotional and quality information.

* <http://www.berouge.com/Pages/default.aspx>

Using this model we revise the value of authorities and hubs in the follow way.

$$Hub^{(t+1)}(s_i) = \sum_{c_j \in C} \sigma_{ij} Auth_{cluster}^{(t)}(c_j) + \sum_{o_j \in O} \alpha_{ij} Auth_{opinion}^{(t)}(o_j) + \sum_{h_j \in H} \omega_{ij} Auth_{helpful}^{(t)}(h_j) \quad (6)$$

$$Auth_{cluster}^{(t+1)}(c_j) = \sum_{s_i \in S} \sigma_{ij} Hub^{(t)}(s_i) \quad (7)$$

$$Auth_{opinion}^{(t+1)}(o_j) = \sum_{s_i \in S} \alpha_{ij} Hub^{(t)}(s_i) \quad (8)$$

$$Auth_{helpful}^{(t+1)}(h_j) = \sum_{s_i \in S} \omega_{ij} Hub^{(t)}(s_i) \quad (9)$$

In Equation (6), $Hub^{(t+1)}(s_i)$ is the t+1 times value of authority node of the sentence. $Auth_{cluster}^{(t+1)}(c_j)$ is the t+1 times value of the clustering center node. $Auth_{opinion}^{(t+1)}(o_j)$ is the t+1 times value of the affective information node. $Auth_{helpful}^{(t+1)}(h_j)$ is the t+1 times value of review of quality information node.

The sentences in this model construct the feature vector based on Unigram. In Equation 6, for iterator (t+1), $Hub^{(t+1)}(s_i)$ indicate the authority node value, $Auth_{cluster}^{(t+1)}(c_j)$ indicate the hub of cluster node value. $Auth_{opinion}^{(t+1)}(o_j)$ is the opinion information of center node. $Auth_{helpful}^{(t+1)}(h_j)$ is the helpful information of center node. $Auth_{opinion}^{(t+1)}(o_j)$ and $Auth_{helpful}^{(t+1)}(h_j)$ presentation emotion feature and comment quality feature separately. α_{ij} is the weight of emotion feature, and ω_{ij} is the weight of comment quality. In order to guarantee the convergence of iteration form, $Hub^{(t+1)}(s_i)$, $Auth_{cluster}^{(t+1)}(c_j)$, $Auth_{opinion}^{(t+1)}(o_j)$ and $Auth_{helpful}^{(t+1)}(h_j)$ are nominalized after each iteration as Equation (10)-(13).

$$Hub^{(t+1)}(s_i) = \frac{Hub^{(t+1)}(s_i)}{\sum_{s_i \in S} Hub^{(t+1)}(s_i)} \quad (10)$$

$$Auth_{cluster}^{(t+1)}(c_j) = \frac{Auth_{cluster}^{(t+1)}(c_j)}{\sum_{c_j \in C} Auth_{cluster}^{(t+1)}(c_j)} \quad (11)$$

$$Auth_{opinion}^{(t+1)}(o_j) = \frac{Auth_{opinion}^{(t+1)}(o_j)}{\sum_{o_j \in O} Auth_{opinion}^{(t+1)}(o_j)} \quad (12)$$

$$Auth_{helpful}^{(t+1)}(h_j) = \frac{Auth_{helpful}^{(t+1)}(h_j)}{\sum_{h_j \in H} Auth_{helpful}^{(t+1)}(h_j)} \quad (13)$$

In the Equations above, C is cluster node set, O is the opinion information set, H is comment helpful information. After modify the Equation, we can combine the opinion

and helpful information to Cluster_HITS model.

4 Experimental Results

The corpus used in this experiment of this chapter is from English comments of 30 topics. Manual annotation will be evaluated by ROUGE-1.5.5. The scale of summarization for English is 120 words. We get opinion information by the MALLET machine learning toolkit, use maximum entropy supervision method, and set all parameters of classification algorithm as default.

We give three reference systems in order to compare them with the summarization based on Cluster_HITS model. The systems explain as follow:

- ✓ Random: The sentences will be selected randomly in every topic, which will be the text sentiment. The results of report are an average of randomly 20 times' selections because of the random of results.
- ✓ MaxSim: We choose the most similar sentences to others in every topic to construct the text sentiment abstract of this topic.
- ✓ Human: The result of text sentiment abstract selected from every topic artificially.

4.1 ADD OPINION INFORMATION

In the Cluster_HITS algorithm, the numbers of cluster are shown in the Table 4. In the experimental process, the reported experimental data by the K-means are an average value of 20 experiments, because the K-means algorithm is random.

TABLE 4 Clusers of Cluster_HITS

Types	Chinese	English
Numbers Of Topic	200	500
K-means	15	30
AGNES	15	30

During the experiment, different number of clusters have a certain influence on the experiment, so this article choose a better number of clusters to make K-means and AGNES also get a good effect. In the Chinese corpus, the numbers of K-means and AGNES clustering are both 15 and the numbers are both 30 in the English corpus.

After making sure the number of clusters, this article compared the result of Cluster_HITS method based on emotions and other summarization method. The results of English abstract are respectively given in Table 5. And Figure 4 to 7 give the influence of opinion feature with different feature weight value to textual opinion summarization results.

Kmeans_HITS and AGNES_HITS on the table refer to the basic Cluster_HITS algorithm using K-means clustering algorithm and AGNES clustering algorithm. ME and Term means emotion detection by using maximum entropy classification methods and emotion detection Term-counting methods. In Table 5, on ROUGE-1, consider the method of topic information between sentences and does not consider the topic information of the Random rate, the average increase is 2 percentage points. The further integration of the subject information and opinion information, summarization effect is nearly 2 percentages than only

considering the topic effect. The other metric of summarization also has the corresponding improvement. This shows that in the text opinion summarization, opinion relevance are equally important as topic relevance, and the relationship between them has important influence on the summarization result.

In the course of the experiment we found that the change of "opinion" feature weight has effect at opinion summarization. Figure 4 to 7 shows the summarization results with different weights of the opinion feature. Because the Kmeans_HITS changed obviously, so only the results of the experiment are given in the below.

TABLE 5 Results of English text sentiment abstract based on Cluster_HITS

Method	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU4
Random	0.3116	0.0409	0.2850	0.1221	0.0909
MaxSim	0.3258	0.0536	0.2871	0.1246	0.1068
Kmeans_HITS	0.3314	0.0604	0.2963	0.1286	0.1103
AGNES_HITS	0.3330	0.0544	0.2966	0.1273	0.1084
Kmeans_HITS_ME	0.3607	0.0778	0.3249	0.1427	0.1293
Kmeans_HITS_Term	0.3640	0.0869	0.3323	0.1467	0.1322
AGNES_HITS_ME	0.3600	0.0860	0.3263	0.1442	0.1342
AGNES_HITS_Term	0.3616	0.0847	0.3292	0.1458	0.1358
Human	0.4092	0.1357	0.3828	0.1731	0.1538

In Figure 4 to 7, ME and Term show the text sentiment results of emotion detection of maximum entropy classification method and Term-counting method respectively. Similar to the PageRank method, when the "opinion" feature weight increases, the rate of accuracy increases. When reaching the peak point, feature weight increases again, the effect on the decline. Moreover, the data in the figure also shows that adding opinion information, textual opinion summarization results are better than without opinion information. Further illustrate that the opinion information is helpful for summarization and we can't only considered the opinion information when extract opinion summary.

4.2 ADD HELPFUL INFORMATION

The comment quality is irregularity because lack of edit and manager. The comment with high quality is helpful for user to recognize the product and have reference value. While the low quality comments even irrelevant to the topic, this comment is no reference value. We use the K-means cluster algorithm to clustered text into 30 topics. Table 6 gives the result of the opinion summarization Cluster_HITS. The opinion detect is used word count method, and set the weight of opinion feature as 2.0. Feature weight of helpful is calculated by the total comments as $h + \frac{\log(\text{total comments})}{2}$.

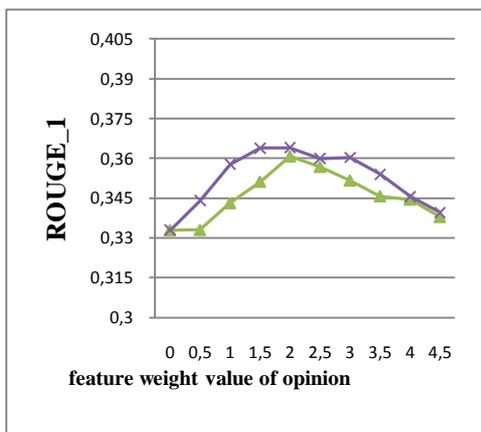


FIGURE 4 Results of ROUGE-1.

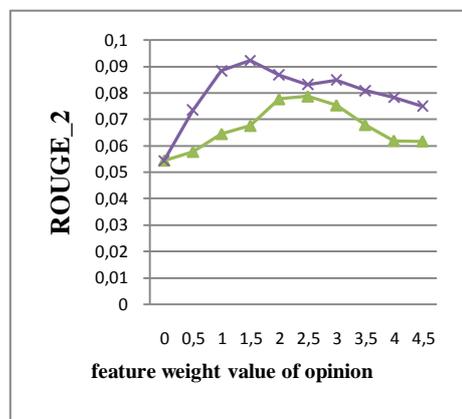


FIGURE 5 Results of ROUGE-2.

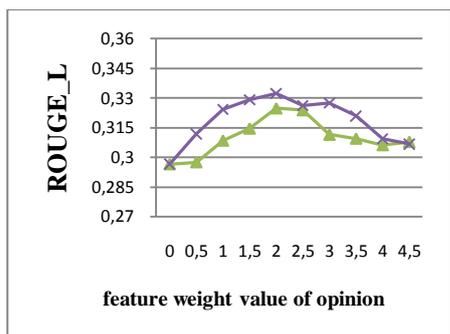


FIGURE 6. Results of ROUGE-3.

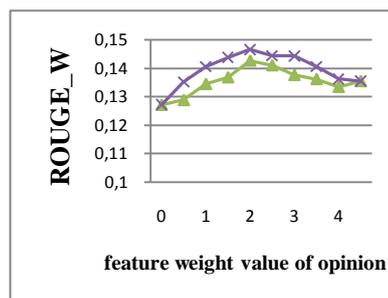


FIGURE 7 Results of ROUGE-4.

▲ EnBi-Rank-ME ✕ EnBi-Rank-Term

TABLE 6 Results of text Opinion summarization add helpful information

Method	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU4
Kmeans_HITS	0.331	0.060	0.296	0.129	0.110
Kmeans_HITS _{helpful}	0.351	0.074	0.319	0.140	0.126
Kmeans_HITS _{Term}	0.364	0.087	0.332	0.147	0.132
Kmeans_HITS _{Term,helpful}	0.376	0.093	0.343	0.152	0.140
Human	0.409	0.136	0.383	0.173	0.154

Kmeans_HITS_{Term,helpful} is Kmeans_HITS method adding opinion information and helpful information. The data in Table 6 shows that summarization result after adding helpful information increased by 1 percentage points in ROUG-1, and with other criteria also increased. This proves that the readers are more inclined to believe that the quality of a good comment, and the comment quality is helpful to textual opinion summarization.

5 Conclusions

This paper mainly researches the opinion summarization based on Cluster_HITS model. We first collect and annotate an English produce comments corpus, which include 30 topics with 500 comments each topic from

References

[1] Hu M., Liu B. (2004) Mining and summarizing customer reviews Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp.168-177.
 [2] Titov I., McDonald R. (2008) A joint model of text and aspect ratings for sentiment summarization. *Proceedings of ACL-08: HLT*, pages 308–316, Columbus, Ohio, USA, June 2008.
 [3] Carenini G., Cheung J. C. K, Pauls A. (2013) Multi-document summarization of evaluative text. *Computational Intelligence*, 29(4): 545-576.
 [4] Lerman, K., Blair-Goldensohn, S., and McDonald, R. (2009, March). Sentiment summarization: evaluating and learning user preferences. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 514-522.
 [5] Nishikawa H., Hasegawa T., Matsuo Y., et al (2010) Opinion summarization with integer linear programming formulation for sentence extraction and ordering. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, pp.910-918.

Amazon. Then we give a Cluster_HITS model for unsupervised textual opinion summarization. As for textual opinion summarization, we incorporate opinion and helpful information in the Cluster_HITS model respectively to consider the impacts of them. Experiment result shows that the opinion and quality information is useful for opinion summarization.

Acknowledgments

This research is supported by the National Natural Science Foundation of China, No.61331011, No.61273320. The science and technology research projects of Henan province education office, No.14A520080.

[6] Wang D., Liu Y. (2011) A pilot study of opinion summarization in conversations. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pp. 331-339.
 [7] Pang B., Lee L., Vaithyanathan S. (2002) Thumbs up?: sentiment classification using machine learning techniques 2002. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, pp.79-86.
 [8] Li S., Huang C. R., Zhou G., et al. (2010) Employing personal/impersonal views in supervised and semi-supervised sentiment classification. *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, pp.414-423.
 [9] Lin L. Y., Wang Z. Q., Li S. S., et al. (2014) Chinese Multi-Document Opinion Summarization via PageRank. *Journal of Chinese Information Processing*. 28(2):85-90(in Chinese with English abstract).

- [10] Kleinberg J. M. (1999) Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604-632.
- [11] Manning, C. D., Raghavan, P., and Schütze, H. (2008). Introduction to information retrieval (Vol. 1). Cambridge: Cambridge university press.

- [12] Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pp. 74-81.

Authors	
	<p>Yancui Li, 12. 02. 1982, China</p> <p>Yancui Li received the Master degree in in computer science and technology Soochow University, China in 2008. Now she is a Ph.D. candidate of Soochow University. She works as lecturer in School of Information Engineering, Henan Institute of Science and Technology since 2008. Her research interests include natural language processing and data mining. She has published more than 15 papers. Mrs. Li is member of China Computer Federation.</p>
	<p>Hongyu Feng, 22. 04. 1977, China</p> <p>Hongyu Feng received the Master degree in computer science and technology from South West Jiaotong University, China in 2006. The author's major field of study is Intelligent computing, natural language processing and computer application. She has published more than 15 papers. She now works in Henan institute of Science and Technology.</p>
	<p>Wenhe Feng, 20.11.1976, China</p> <p>Wenhe Feng received the Ph.D. degree in linguistics from Wuhan University, China in 2010. Currently, he is a researcher at Henan Institute of Science and Technology, China. His research interests include natural language processing and machine learning.</p>