Using syntactic rules to combine opinion elements in Chinese opinion mining systems

[*1]Shih-Jung Wu, [2]Rui-Dong Chiang
*1*Department of Innovative Information and Technology, Tamkang University, Taiwan*
wushihjung@mail.tku.edu.tw
*2*Department of Computer Science and Information Engineering, Tamkang University, Taiwan*
chiang@cs.tku.edu.tw

Abstract

Most Chinese opinion mining systems use the specific pattern and nearby approach to combine relevant opinion elements (feature words and opinion words) to express the opinion tendencies of authors. In this paper, we propose a rule-based ad hoc method to study the combination problem of Chinese opinion elements. We extracted the opinion elements of articles based on lexicons and then combined them with the different sentence patterns and grammars to analyze the authors' opinions. Because the articles on the online communities such as blogs, wikis, online forums, etc. do not have a defined format, there are often opinion comments that do not refer to the topic, resulting in information loss and significantly reduced recall. Therefore, the "default topic" method is proposed to correct this type of problem. Additionally, there might be errors when using the nearby approach to combine opinion elements. Thus, we propose the concept of "clause priority" to increase precision. After 20 months of long-term tracking and analysis, the experimental result indicates that the method proposed in this paper had good precision, recall, and F1 of opinion tendency analysis for review articles.

Keywords: *Opinion mining, Sentiment analysis, Sentences pattern*

1. Introduction

As electronic malls have flourished, people's spending habits have changed. In the past, people usually acquired the evaluation of a product through word of mouth from friends and relatives before making a purchase. Now, consumers can obtain related opinions and other people's comments on a product through various online forums before considering whether to purchase the product. According to the 2009 Global Online Consumer Survey Report published by The Nielsen Company[1], a market research company in the United States, approximately 70% of consumers trust product-related online comments and opinions. Consumers also comment on the service of companies or sellers. Therefore, companies or sellers can find out the reputation of both their products and their services through comments published by consumers on the online forum and then adjust the strategies targeting the related products or deficiency in service.

Opinion mining reviews are typically analyzed at various resolutions [5]. One is document-level opinion mining identifies the overall subjectivity or sentiment expressed on an entity in a review. The other is sentence-level which can associate opinions in detail with specific aspects of entity. Some related studies define the opinion elements to express people opinions toward entities such as products, services and their attributes for sentiment analysis [6-9]. In this paper, we focus on subjective sentences for analysis. We use four types of opinion elements (topic, feature, item, opinion word) to form an opinion sentence from sentences with subjective opinions. Topic is the subject of people's comments. That is usually a company or vendor name. Feature represents products or related services for the topic. If the comment has more detail attributes about feature, we use items to represent them. Opinion words are usually emotional words or adjective word which to express the evaluation refers to the reviewers or people's comments.

---

[1] http://blog.nielsen.com/nielsenwire/consumer/global-advertising-consumers-trust-real-friends-and-virtual-strangers-the-most/

This paper studies the combination problem of Chinese opinion elements. Most of the studies on the combination of opinion elements use the specific pattern and nearby approach to combine the relevant opinion elements[1-3]. But, Qiu et al. [4] found that part of the feature words or opinion words extracted based on patterns were words in other domains or words that are meaningless. Unlike professional reviews written by experts, articles posted on the forums are usually unformatted and very colloquial, these articles are often contain new words and opinion elements that belong to other domains. Because designing a universal Chinese opinion mining system for all the domains is difficult, our study focused on a single domain by assuming that all the feature words and opinion words of the domain are complete. In this paper, the opinion elements of articles were extracted based on lexicons and combined with the sentence patterns (general sentences, equative sentences, and comparative sentences) and context dependency to analyze the authors' opinion tendencies. Moreover, we applied the method of "default topic" to solve the problem of missing topics for the unformatted articles posted on the online forum and applied "clause priority" to correct the combination error when using the nearby approach.

This paper is organized into six sections. The first section introduces the research motivation and objective. The second section reviews related studies on opinion elements combination. Default topic and clause priority are introduced in third section. The forth section describes in detail the method of opinion elements combination and the sentence patterns and grammars proposed in this study. The fifth section presents the related experimental result, and the discussion and future work are given in the last section.

2. Related work

In most studies of Chinese opinion mining systems, they use parsers or POS tools to search for and combine opinion elements based on specific patterns and nearby approach. Because most methods in these studies are similar, we only introduce related studies on Chinese below.

Li et al. [6] noted that feature words in English are mostly nouns or noun phrases, whereas in Chinese, they could be nouns, verbs, or combinations of a verb and noun; This study analyzed the Chinese reviews in the mobile domain by first using Chinese Lexical Analysis System of Institute of Computing Technology (ICTCLAS)[2] to punctuate the sentence and identifying the preset feature word and then looking for adjectives to the right of this feature word as the opinion words for combination. Guo et al. [10] propose Multi-aspect Sentiment Analysis for Chinese Online Social Reviews. Their method applies the Latent Dirichlet Allocation (LDA) model to discover aspects automatically. Zhang et al. [11] used the parser developed by Harbin Institute of Technology (HIT-IR LTP) to analyze and find the preset word dependency in a sentence for combining feature words and opinion words and, finally, gave different scores according to the modification of different adverbs. Chen and Yao [1] proposed an algorithm to extract related feature words and opinion words from articles and defined the relationship between the topic word and the opinion word as a 2-tuple <topic, sentiment>. The topic can be a certain brand or the feature of this brand, and this topic is established according to the ontology; sentiment is the evaluation of topic. Therefore, after parser analysis, the specific sentence structure was combined with related opinion elements. The main methods in the analysis were NSUBJ (nominal subject), DOBJ (direct object), and AMOD (adjectival modifier). Peng and Shih [2] applied the CKIP[3] word segmentation tool to articles on the PTT[4] forum, used mainly specific patterns to combine opinion words in the article and output, and then calculated its polarity.

Opinion polarity is determined mainly to judge whether the article or sentence conveys a positive or a negative opinion tendency. Opinion polarity helps users to rapidly find out the quality of a product or the opinion tendency of other people for a certain product or a topic. Xu et al.[12] identify the opinion polarity based on S-HAL (Sentiment Hyperspace Analogue to Language). Hatzivassiloglou and McKeown [13] proposed the prediction of the opinion words of unknown tendency with conjunctions: if in the sentence, there are adjectives in front of and behind "and", these two adjectives give the same opinion tendency, whereas "but" means that the two adjectives on the two sides have the opposite

---

[2] http://ictclas.org/

[3] http://ckipsvr.iis.sinica.edu.tw/

[4] http://www.ptt.cc/index.html/

opinion tendency [4, 14, 15]. Some scholars also consider the determination of negation words, while others consider that a negation word in front of the opinion word or in the same sentence will reverse the original tendency of opinion word[2, 4, 8, 14, 16-20]. Similar to Peng and Shih [2], in this paper, we considered only a negation word in front of the opinion words as changing the polarity of opinion words. Negation words behind were not considered. In addition, Morinaga et al. [21] suggested that for opinion elements extracted from articles in a specific domain, there was correspondence between feature words and opinion words. Therefore, Ding et al. [17] considered context dependency, which means that, in certain situations, the tendency of an opinion cannot be determined only by opinion words and feature words and opinion words must be combined to determine the tendency of an opinion. Although the opinion tendency is divided into positive and negative only and degree of tendency is not determined in this paper, we also consider context dependency.

3. Default topic and Clause priority

This section introduces the concepts of default topic and clause priority.

3.1. Default topic

There is no restriction on the length of articles on Mobile01[5] (an online forum in Taiwan). If the author mentions a topic in the article, multiple sentences are usually used to thoroughly describe this topic. The author would most likely use the subsequent clauses for supplementary description. Considering only clauses with topics would disregard some clauses describing this topic because some clauses do not mention a topic. As a result, there would be an incomplete opinion expression, and subsequent analysis would result in information loss. Therefore, we propose the concept of default topic. If a topic is mentioned, but is not mentioned in subsequent clauses, we would use the topic mentioned in the previous sentence for subsequent clauses, avoiding incomplete opinions when the subsequent clauses do not mention a topic.

Example 1: "A 公司的收訊真的還算蠻 OK 的, …,客服方面的素質也算還不賴!!"("The reception of company A is really good, …, the quality of customer service is not bad at all!")

In Example 1, the author wants to express two opinions: "A 公司收訊 OK" ("the reception of company A is good") and "A 公司客服不賴" ("the quality of customer service of company A is not bad"). The topic contained in the first sentence is "A 公司/company A", and it is stated that the quality of its reception is good. The second sentence, however, does not mention "A 公司/company A" as the topic. Here, if the default topic is not set to be "company A", we would not be able to know that "客服素質還不賴" ("the quality of customer service is not bad") is describing "A 公司/company A", resulting in missing content in subsequent clauses. Therefore, using default topics can avoid the situation of missing topics.

3.2. Clause priority

The pairing method for opinion elements mainly combines the nearby approach and clause priority. The nearby approach considers the opinion word closest to the feature word as the opinion word that describes this feature. When pairing opinion elements, the author often does not express a complete opinion in one clause but first mentions the feature in the prior clause and then uses the opinion word in the following clause to describe this feature. However, when a feature is mentioned in prior clauses and another feature and opinion word are also mentioned in the next clause, the feature and opinion word in this clause is considered with priority. The prior feature is replaced by the subsequent feature.

Example 2: "B 公司推出的方案，爛的網路…"("the plan of company B, terrible network….")

In Example 2, "爛/terrible" is the main opinion word, which, according to clause priority, is considered to describe the feature of the same clause, "網路/network" rather than the feature of the prior clause "方案/the plan". In this example, if only the nearby approach is considered, the wrong

---

opinion elements would be extracted for pairing, resulting in output error. However, if the concept of clause priority is applied, the system will first determine that the feature "網路/network" appears in the clause and then generate the correct output result.

4. Sentence patterns and grammars

Regarding the analysis of opinion tendency of articles on the online forum, three sentence patterns are summarized in combination with default topic, clause priority, and statistical correction. We explain in detail the sentence patterns and grammars in this section. Statistical correction is used to correct the opinion tendency of authors with contradicting opinion output. In the following, we introduce three sentence patterns: general sentence, equative sentence, and comparative sentence. General sentence may contain equative sentence, comparative sentence and negative sentence and so on. Generally, sentences do not belong to equative sentence and comparative sentence which are regarded as general sentence to analysis.

4.1. General sentence pattern

When pairing opinion elements of general sentences, the default topic, i.e., the previous topic, will be used when there is no topic in the subsequent sentence. When analyzing articles, the feature, item, and opinion word will be paired according to the pairing method of opinion elements in above subsection clause priority. In the following, the pairing method and the output results are shown in Example 3.

Example 3： "我只辦過一支 A 公司的，話說我家的收訊真的爛的可以，2G 的訊號才一格，講一講還會斷線，還好我辦的是 0 月租，再送 200 元通話費，用了一年多，還沒收到帳單過。" ("I only used a plan of company A. I can only say that the reception in my house is really terrible, the 2G signal only has one bar, and it will disconnect after talking for a short while. However, fortunately my plan has no monthly fee, along with a 200-Yuan bonus. After more than one year, I have not received any bill yet.").

Here, the topic is "A 公司/company A", features are "收訊/reception" and "訊號/signal", the item is "2G", and the opinion words are "爛/terrible", "一格/one bar", and "斷線/disconnect". At the beginning of analyzing this article, the concept of default topic is used to set the topic as "A 公司/company A", which will be put into the tuple. The search for the next opinion element continues and the feature "收訊/reception" is found. There is no feature in the current tuple and no need to determine the correspondence. The feature is put directly into the tuple, and the compositions of the tuple become (A 公司/company A, 收訊/reception). Then, the next opinion word of the opinion element "爛/terrible" is found. Therefore, the opinion word is put into the tuple, and now the tuple becomes (A 公司 /company A, 收訊/ reception, Ø, 爛/terrible). Then, the aforementioned process of pairing opinion elements is repeated until the end of the article. The final output results are (A 公司/company A，收訊 / reception，Ø，爛/terrible), (A 公司/ company A，訊號/signal, 2G，一格/one bar), and (A 公司 /company A，Ø，Ø, 斷線/disconnection).

4.2. Equative sentence pattern

The equative sentence pattern is the case where opinion elements of the same type are connected with conjunctions. The standard pattern of an equative sentence is "A conjunction B conjunction C conjunction…D" ("A, B, C, and D are the same type of topic, feature, item, or opinion word"). Because only one type of opinion elements can be put in the tuple, when the opinion elements on the same level are connected with conjunctions, there must be a process of separation. A new tuple is added for each level of opinion elements. Then, opinion elements are paired according to the description above. Example 3.4 is an explanation of equative sentences.

Example 4: "A 公司, B 公司和 C 公司的收訊都很不錯" ("the reception is good for companies A, B, and C.")

In Example 4, companies A, B, and C (A 公司, B 公司 and C 公司) are connected with conjunctions "," and "和/and", complying with the equative sentence pattern above. Therefore, the sentence is separated into three tuples, which are then treated as general sentences. The final output results are (A 公司/ company A，收訊/ reception，Ø，不錯/ not bad), (B 公司/ company B，收訊/ reception，Ø, 不錯/ not bad), and (C 公司/company C，收訊/reception，Ø, 不錯/not bad).

4.3. Comparative sentence pattern

In articles on the online forum, there are often comparisons between products or companies, and there are two types of comparative sentences.

"A…比較(more)…opinion word" ("A could be any combination of the topic, feature, and item").

This is a relatively simple sentence that only describes the unilateral good (or bad), and the pairing processing is the same as in general sentences. However, if the topic does not appear in the clause of comparative sentences, the default topic will be used as shown in Example 5.

Example 5: "A 公司比較省" ("company A is more economic.")

The output is (A 公司/ company，Ø，Ø，省/ economic).

"A…比 (than)…B…opinion word" ("A and B can be any combination of the topic, feature, and item").

For this comparative sentence, if there is no topic in front of "比/than", the topic appearing in the previous sentences will be used as the default topic; the feature part is also centered around "比/than". If the topics of the prior and subsequent sentences are different, and a feature is mentioned in the prior sentence but not the subsequent one, then the feature in the prior sentence is used for the subsequent sentence. We will base on compare sentence properties to give opposite polarity of opinion word in subsequent sentence (Add a negative word "not/不" before the opinion word which belongs to subsequent sentence). Detail description demonstrates in Example 6.

Example 6: "A 公司的速度真的比 B 公司的快" ("the speed of company A is faster than that of company B.")

Using "比/than" as the center, a feature, "速度/speed", is found to be in front but not behind. Thus, this feature will also be used as the feature for the part behind "比/than". The result is two sentences: (A 公司/company A，速度/speed，Ø，快/faster) and (B 公司/company B，速度/speed，Ø, not 快/not faster).

5. Discussion of experiments

This section discusses the experimental results. First, we present the experimental results of data in the telecommunication domain on the Mobile01 online forum, and then we discuss, with data, the factors that affect the precision and recall. We display the result of long-term tracking next. The approach of data assessment is to measure precision, recall, and F1, which are defined as following equations:

$$\text{precision} = A/(A+C1+C2) \tag{1}$$

$$\text{recall} = (A+C1)/B \tag{2}$$

$$\text{F1-Measure} = 2* \text{precision}* \text{recall}/(\text{precision}+ \text{recall}) \tag{3}$$

A is the number of complete sentences output by the system that are correct outputs, B is the number of manually labeled complete sentences, *C1* is the number of sentences output by the system that are incorrect in meaning, and *C2* is the number of extra complete sentences found by the system.

Table 1 is the schematic of the system output.

Table 1. System output schematic.

	Complete sentence output by the system	*Manually labeled complete sentence*
Correct	A	B
Incorrect	C1+C2	0

5.1. Experimental results

We use articles from the four months from July to October in 2011 for the experiment, and the amount of data is shown in Table 2. Sample articles which discuss the same topic in Mobile01 always include one original article and many follow up posts. After analysis, a post article usually output 4 to 5 or more tuples. For the experiments, Intel® Core™2 Duo Processor E8400 (6M Cache, 3.00 GHz, 1333 MHz FSB) computers were contain 4 GB RAM. The algorithm is developed under python. These articles are analyzed with default topic, clause priority, and the sentence patterns and grammars proposed in this paper. The experimental results are shown in Table 3. The averages of precision, recall, and F1 are 80.8%, 92.7%, and 0.861, respectively. According to the experimental result, our methods show a stable precision, recall and F1 in every month.

Table 2. Telecommunication data volume of Mobile01.

Month	*Number of articles*	*Number of sentences*
July	2275	23071
August	2055	21889
September	2020	20812
October	2025	20945

Table 3. Overall data from July to October.

	July	*August*	*September*	*October*
Precision	600/732=0.819	499/619=0.806	474/588=0.806	441/552=0.799
Recall	631/668=0.945	538/590=0.912	504/549=0.918	478/512=0.934
F1	0. 876	0.856	0.858	0.861

5.2. Long-term tracking

In this section, we analyze the data from July of 2011 to February of 2013, and the results of long-term tracking are shown in Figure 1. It also shows the line graph for precision, recall, and F1 value for each month. The averages of precision, recall, and F1 are 84.0%, 89.4%, and 0.865, respectively. According to the results of long-term tracking, the precision and recall of the sentence pattern and grammar proposed in this study worked well and was stable.
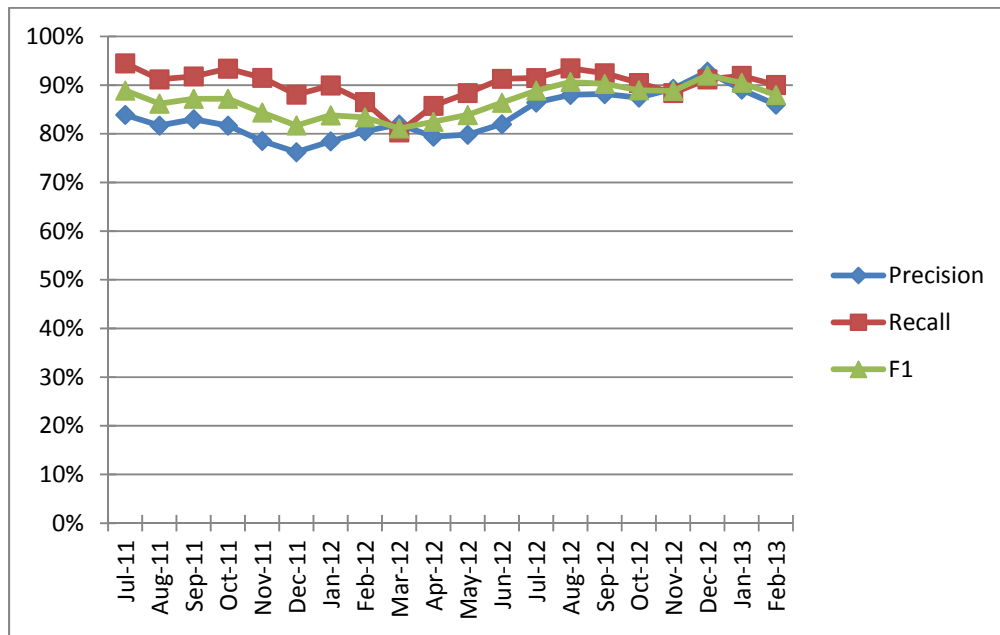
Figure 1. Line graph of overall long-term tracking data of each month

6. Discussion and future work

This paper studied the combination problem of Chinese opinion elements. The opinion elements of articles are extracted based on lexicons, and the concepts of "default topic" and "clause priority" are used in opinion element combinations. Then, the three sentence patterns and grammars we summarized are combined to express the authors' opinions and determine the opinion tendency of the authors. For the experiment, we used articles in the domains of telecommunication from the Mobile01 forum. The experimental results indicate that, under short-term data analysis and long-term tracking, the methods we proposed had good and stable efficiency.

In the future, the classification of subjective sentences and objective sentences shall be studied, and the classification result can be integrated into the grammar to reduce the error of this part, subsequently increasing precision. Additionally, in this study, question sentences and hypothetical sentences were considered to not carry any subjective opinion, and this type of sentences was first excluded to improve the precision, which also resulted in the loss of hypothetical sentences with the subjective opinions of authors. Because there are still many narrative sentences and hypothetical sentences in review articles, it is necessary to improve the analysis of narrative sentences and hypothetical sentences to improve the precision and recall, respectively.

7. References

[1] Chen Mosha, Yao Tianfang, Combining dependency parsing with shallow semantic analysis for Chinese opinion-element relation identification, Universal Communication Symposium (IUCS), pp. 299-305, 2010.
[2] Ting-Chun Peng, Chia-Chun Shih, Using Chinese part-of-speech patterns for sentiment phrase identification and opinion extraction in user generated reviews, Digital Information Management (ICDIM), pp. 120-127, 2010.
[3] Shu Zhang, Wenjie Jia, Yingju Xia, Yao Meng, Hao Yu, Extracting Product Features and Sentiments from Chinese Customer Reviews, Seventh International Conference on Language Resources and Evaluation, pp.1142-1145, 2010.

[4] Guang Qiu, Bing Liu, Jiajun Bu, Chun Chen, Expanding domain sentiment lexicon through double propagation, Proceedings of the 21st international jont conference on Artifical intelligence, pp.1199-1204, 2009.

[5] Zhen Hai, Kuiyu Chang, Jung-Jae Kim, C.-C. Yang, Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance, Knowledge and Data Engineering, Vol. 26, No. 3, pp. 623-634, 2014.

[6] Zhichao Li, Min Zhang, Shaoping Ma, Bo Zhou, Yu Sun, Automatic Extraction for Product Feature Words from Comments on the Web Information Retrieval Technology, Vol. 5839, pp. 112-123, 2009.

[7] Bing Liu, Lei Zhang, A Survey of Opinion Mining and Sentiment Analysis Mining Text Data, pp. 415-463, 2012.

[8] Soo-Min Kim, Eduard Hovy, Determining the sentiment of opinions, Proceedings of the 20th international conference on Computational Linguistics, Article No. 1367, 2004.

[9] Bing Liu, Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies, Vol. 5, pp. 1-167, 2012.

[10] Fu Xianghua, Liu Guo, Guo Yanyan, Wang Zhiqiang, Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon, Knowledge-Based Systems, Vol. 37, pp. 186-195, 2013.

[11] Changli Zhang, Daniel Zeng, Jiexun Li, Fei-Yue Wang, Wanli Zuo, Sentiment analysis of Chinese documents: From sentence to document level, Journal of the American Society for Information Science and Technology, Vol. 60, No. 12,   pp. 2474-2487, 2009.

[12] Tao Xua, Qinke Peng, Yinzhao Cheng, Identifying the semantic orientation of terms using S-HAL for sentiment analysis, Knowledge-Based Systems, Vol. 35, pp. 279-289, 2012.

[13] Vasileios Hatzivassiloglou, Kathleen R. McKeown, Predicting the semantic orientation of adjectives, Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, pp.174-181, 1997.

[14] Minqing Hu, Bing Liu, Mining and summarizing customer reviews, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177, 2004.

[15] Guang Qiu, Bing Liu, Jiajun Bu, Chun Chen, Opinion Word Expansion and Target Extraction through Double Propagation, Computational Linguistics, Vol. 37, No. 1, pp. 9-27, 2011.

[16] Bin Shi, Kuiyu Chang, Mining Chinese Reviews, Data Mining Workshops, pp. 585-589, 2006.

[17] Xiaowen Ding, Bing Liu, Philip S. Yu, A holistic lexicon-based approach to opinion mining, Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 231-240 , 2008.

[18] Ku Lun-Wei, Liu I-Chien, Lee Chia-Ying, Chen Kuan-hua, Chen Hsin-Hsi, Sentence-Level Opinion Analysis by CopeOpi in NTCIR-7, Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access, pp. 260-267, 2008.

[19] Qiang Ye, Wen Shi, Yi Jun Li, Sentiment Classification for Movie Reviews in Chinese by Improved Semantic Oriented Approach, Proceedings of the 39th Annual Hawaii International Conference on, pp. 53b-53b, 2006.

[20] Li Zhuang, Feng Jing, Xiao-Yan Zhu, Movie review mining and summarization, Proceedings of the 15th ACM international conference on Information and knowledge management, pp. 43-50, 2006.

[21] Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, Toshikazu Fukushima, Mining product reputations on the Web, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 341-349, 2002.