2011 International Conference on Financial Management and Economics
IPEDR vol.11 (2011) © (2011) IACSIT Press, Singapore

The Analysis of Hotel Customer Generated Contents in Weblogs

Rueijiau Sung, Chaochang Chiu [+], Peiyu Hsieh, Huiling Chou

Department of Information Management
Yuan Ze University, Chungli, Taiwan

[1]**Abstract.** This research applies text mining approach to reveal and analyze the customers' comments about price, food, facilities, environment, and service of the hotels in Taiwan. We adopt a proposed heuristic n-phrase rule to identify the polarity customer's opinions and provide position maps to visualize the pros and cons of respective hotels. The research findings about different views and opinions of customers' perception or experiences with the hotels may provide hotel management with useful insights of customer feedbacks and clues for future improvement.

Keywords: Text mining; classification; heuristic n-phrase rule; correspondence analysis

1. Introduction

The advent of the Internet has extended consumer options for gathering product information from other consumers and offering their own consumption-related advice by engaging in electronic word-of-mouth (eWOM). Hennig-Thurau et al. (2004) refer to eWOM communication as any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet. Importantly, eWOM provide hotel managers a cost-effective and unbiased window to assess customer-perceived service quality and glimpse competitors" customer satisfaction levels, while helping to improve their business management (Litvin et al., 2007).

Blogs related to tourism have become popular and been regarded as a valuable source for data, as it provides insight into users and their behaviors which known as the "blogosphere" (Carson, 2008). Blogs is the higher perceived credibility of consumer opinions as compared to traditional tourist information sources. Customer generated contents (CGC) provides tourism organizations and enterprises with valuable market intelligence and ongoing market research opportunities (Akehurst, 2009). These CGC can help hotel managers learn consumer preferences for different hotel attributes, both internal and external. As the content rapidly expands day by day, the hotel managers cannot read and check all the contents potentially mentioning their hotel. In this study, we develop a semi-automatic information retrieval system to extract online customer eWOM by feature based opinion mining technique. We propose heuristic n-phrase rule to find out customer opinions about price, food, facilities, environment, and service of the hotels in Taiwan. Then, we use correspondence analysis to examine the association of the customers" opinions on different hotels. It is hoped that this study can provide hotel managers useful insights for online tourism eWOM management and strategy planning.

2. Blogs Analysis in Tourism

Prior researches about CGC on travel blogs have shown the potential application as a marketing tool and increase competitiveness in tourism. Carson (2008) suggests that valuable data can be drawn from travel

[1] [+] Corresponding author. Tel.: + (886-955073838); fax: +(886-34352077).
E-mail address: 101chiu@gmail.com

blogs and finding more effective ways to reduce noise in locating blogs. Volo (2010) points out the travel blogs can be an effective, unobtrusive method of collecting data.

Opinion mining, also known as Sentiment Analysis, provides a useful mechanism to determine the subjectivity, sentiment, appraisals or feeling of an author expressed in texts on specific topics (Liu, 2010). It can effectively reduce the use of manual labor in identification, storage, and analysis of business intelligence. Previous works have been proposed opinion mining at the document, sentence, or feature level. However, the document-level and sentence-level sentiment analysis can determine the overall document or sentence but do not indicate which specific features of an object are evaluated positively or negatively.

Feature-Based Opinion Mining is mainly focus on two successive steps: First, identify object features that have been commented on. Secondly, determine the respective opinion on the feature (Liu, 2010). Our mapping of opinion words to the features is most similar to that of (Liu et al., 2005). We propose the heuristic n-phrase model to extract the opinion features which is suitable for Chinese, first locate feature words, then identify opinion polarity.

3. The Research Design

3.1 Data acquisition and Preprocessing

A web crawler program is developed to spider customer generated articles from Wretch, Yahoo Blogs, and Yam blogs. The article websites, topics, abstracts, HTML source codes, full text are retrieved and put them in the initial CGC databank. Then, we apply the CKIP (Chinese Knowledge Information Processing), a Chinese word segmentation system developed by Academic Sinica, to word segmentation and part-of-speech tagging. The feature words are usually nouns or noun phrases, while the opinion words are usually adjectives. The noun, noun phrase are filtered as the candidate features, and adjective, adverb, verb as the candidate opinion words. We use term frequency with a threshold to select the frequent features. Features granularity is used to construct a lexicon with expressions describing different hotel features after examination of frequency lists for candidate features. There are five sub-categories of features are identified including ***Price, Food, Services, Facilities, and Environment***. Further, we identify the polarity of these candidate opinion words base on the positive and negative opinion words from HowNet and modified by domain experts. These positive and negative opinion words are represented as “pros” and “cons”

3.2 Replace features and opinion words with [feature] and [opinion polarity]

For the diversity of Chinese words combination, the different words combination could result in similar or completely different meanings. We replace the related terms with the predefined feature and opinion polarity signal (pros, cons). For example, spacious room→[pros] [facilities]; unpalatable food →[cons] [food]; beautiful scenery→[pros] [environment]; service not bad→[service] [pros]. This replacement can ensure to find general Chinese language patterns.

3.3 Heuristic n-phrase rule

We propose heuristic n-phrase rule to identify the “opinion polarity” on a “feature” in processed sentences. With a given number, n, to extract the terms either side of “feature”. The steps are as follow: First, we identify the “feature” in a processed sentence. Then, we check the first phrase prior to the “feature” to determine whether it is “opinion polarity”. If match, we put these two words together, “opinion polarity”+ “feature”. If not, we check the first phrase after the “feature”, And so on. In addition, if there are more than two conjoint “features”, we will check if these “features” are the same or not. If the “features” are different, then calculated them separately. For example, a reduced sentence “opinion polarity” + “$feature_1$” + “$feature_2$” will be transformed to “opinion polarity” + “$feature_1$” and “opinion polarity” + “$feature_2$”. The architecture of heuristic n-phrase rule is shown in Figure 1.

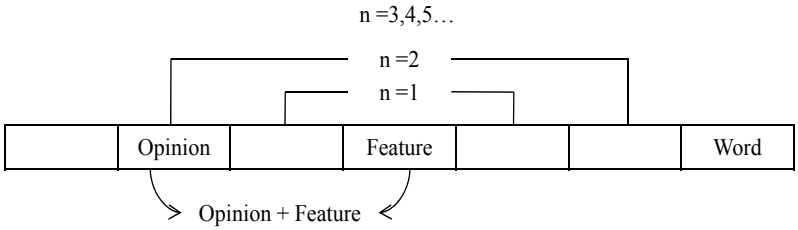

Figure 1. Architecture of Heuristic n-phrase Rule

3.4 Classification and Evaluation

The commonly used classifier, such as C5.0, Neural-Network (NN), Support vector machine (SVM), are use to demonstrate the performance of the heuristic n-phrase rule. We adopt five-fold cross-validation and use average accuracy as evaluation measure.

4. Evaluation of Heuristic n-phrase Rule and Experiment Results

We conduct experiment and evaluate the proposed method. The samples used in this study are 826 reviews of hotels using the keyword "hotel" automatically spider from "Yahoo Blogs", "Wretch", and "Yam Blogs" web sites from April 1, 2009 to March 31, 2010. After text preprocessing and filtering, the remaining 2790 sentences are labeled as 2523 positive opinion sentences and 267 negative opinion sentences.

For the length of these sentences have been significantly reduced, we demonstrate the heuristic n-phrase rule with n=2 and n=3, that is, bi-phrase and tri-phrase. The commonly used classifier NN, C5.0, and SVM are used to measure the classification performance. In general, using heuristic n-phrase rule, both n=2 or n=3, the accuracy results of "opinion polarity" + "feature" are over 85%, although the results are different due to diversified classifiers. In addition, heuristic tri-phrase rule achieves the highest value (bold words) in each feature and most of the accuracy results are over 90%. Heuristic tri-phrase rule is obviously performs better than bi-phrase rule. The experimental results show that the proposed heuristic n-phrase rule can demonstrate its feasibility and effectiveness and achieve a satisfactory performance in Chinese.

5. Correspondence Analysis

Correspondence analysis is conducted to visually show the relationships between features and opinions polarity, with distance on the map represents the correspondence (closeness), closer proximity means greater perceived similarity. Correspondence map can provide the relative performance of key factors between competitors and reflect what the customers perceived. It is necessary for an hotelier to understand the market opportunities and competitive threats, in particular, the competitors in the same field.

We manually group hotels into four different styles, that is, urban, scenery, recreation area, and hot spring hotels, based on the hotel's core resource in order to more clearly demonstrate the state of competition. The top five hotels by number of customer reviews in each style of hotel are selected as the samples. Table 1 denotes the 1363 sentences that are selected with heuristic tri-phrase rule from the reviews about these 25 hotels. 1173(86.1%) of the sentences are pros opinions and the remaining 190(13.9%) are cons opinions. Figure 2 provides the correspondence map on the relative proximities of the five sub-categories of features (price, food, facility, environment, service) and the opinions polarity (pros, cons) modify on these features of scenery style hotels. The horizontal axis (Dimension 1) accounted for 52.86% and vertical axis (Dimension 2) for 28.58%, accordingly the associations of each feature and customer opinions polarity are explained on Dimensions 1 and 2 (81.44%).

Table 1. A Summary of Pros and Cons Reviews of Four Style Hotels

Hotel	Price		Food		Facility		Environment		Service		P	C	Total
	P	C	P	C	P	C	P	C	P	C			
Scenery	**14**	**5**	**71**	**12**	**88**	**20**	**192**	**13**	**53**	**9**	**418**	**59**	**477**
H_1	5	1	23	3	17	7	55	5	26	2	126	18	144
H_2	3	1	6	1	5	3	39	1	3	2	56	8	64

…	2	1	20	2	26	2	24	3	10	1	82	9	91
Urban	**30**	**6**	**173**	**18**	**91**	**7**	**31**	**6**	**32**	**9**	**357**	**46**	**403**
H_i	3	1	21	2	14	2	6	2	15	1	59	8	67
….	2	2	19	1	14	1	9	1	4	2	48	7	55
Hot Spring	**16**	**9**	**106**	**11**	**91**	**18**	**41**	**9**	**40**	**9**	**294**	**56**	**350**
H_n	2	4	53	3	12	4	14	2	4	1	85	14	99
…	3	1	15	3	29	4	6	1	17	3	70	12	82
Recreation Area	**10**	**5**	**21**	**5**	**40**	**8**	**24**	**6**	**9**	**5**	**104**	**29**	**133**
H_k	2	1	2	1	19	3	12	2	1	1	36	8	44
…	3	1	10	1	16	2	2	1	4	1	35	6	41
Total	**70**	**25**	**371**	**46**	**310**	**53**	**288**	**34**	**134**	**32**	**1173**	**190**	**1363**

* P: Pros, C: Cons

In the graph, we can see the customer opinions of these five scenery hotels are located in four different quadrants, except two, Fushoushan and Ming Chr, in the same quadrant. This indicates that customers reflect the different pros and cons on the features of these hotels differently. The Promised Land Resort is mostly related to pros opinions of food. The UNI-Hotel is mostly related to pros opinions of facility. Tai Ping Mountain Resort, Fushoushan Farm, and Ming Chr Resort, government-owned business and operate by themselves or outsourcing, are mostly related to "Environment-Pros" and "Service- Cons". Obviously, these hotels get the different position in the minds of consumers, and also demonstrate the business operating characteristics and competitive advantages/disadvantages. Figure 3 demonstrates the correspondence map of urban style hotels. The contribution of both dimensions (Dimensions 1 and 2) demonstrates to explaining the association of the feature and customer opinions polarity is fairly high (81.89%).

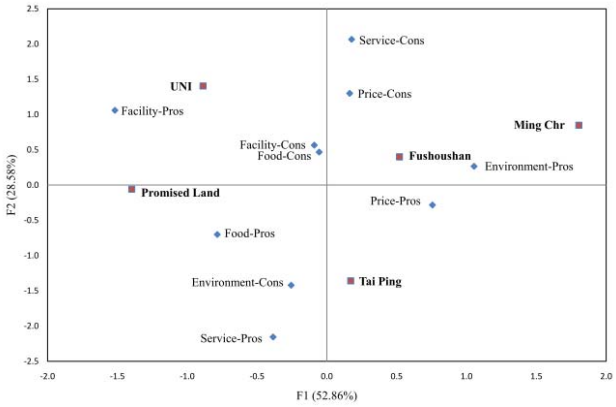

Figure 2 Correspondence Map of Scenery Hotels

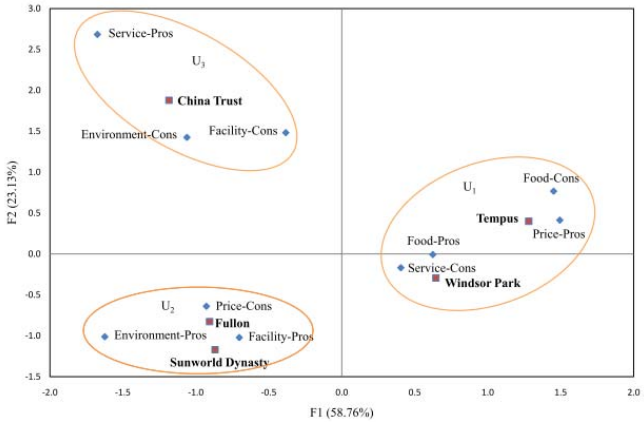

Figure 3 The Correspondence Map of Urban Style hotels

These five urban hotels seem to cluster into three groups (U1, U2, U3) with each showing a strong association with the different opinions of customers. The cluster U1 contains Tempus Hotel and Windsor Park Hotel related to “Price-Pros”, “Service-Cons”, and “Food- Pros/Cons”, although Windsor Park Hotel has gained more pros opinions on food than Tempus Hotel. We can see that these two hotels provide an acceptable price and roughly equal positive and negative comments in diet, but dissatisfied with the service in the minds of consumers. The cluster U2 contains Sunworld Dynasty Hotel and Fullon Hotel related to “Facility-Pros” and “Environment-Pros”, but “Price-Cons”. Finally, the cluster U3 contains China Trust Hotel related to “Service-Pros”, but cons opinions on environment and facility. Similar maps are also applied to hot spring hotels and recreation area hotels.

6. Discussion and Conclusion

We have proposed an analysis prototype of opinion mining of hotel customer generated contents. Although many studies have suggested that customer generated contents in blogs are useful and valid for application. There are some inherent limitations such as the weakness in blogs itself such as free format of text and the difficulty to generalize findings due to the small size of bloggers, majority are young people. Thus, the research results can reflect partial customers’ reviews, but not for all. In this paper, we have illustrated the whole opinion mining of hotel customer generated contents in Chinese-language weblogs and proposed technique for the automatic feature-opinion extraction as well as for the visualization of the detected opinions polarity on the different features in hotel industry. A heuristic-phrase rule is proposed to extract the feature-opinion parity in order to calculate the frequency of the customer opinion polarity on the features. The experimental results show that heuristic n-phrase rule demonstrates its feasibility and effectiveness and achieve a satisfactory performance in Chinese. Furthermore, we use correspondence analysis to visualize the competitive position based on the customer generated contents in blogs that those visualizations can provide valuable insight in consumer-directed market positioning strategy for hoteliers.

7. References

[1] Tourism Bureau, Republic of China (Taiwan) (2011). *Monthly Statistic Report on Tourism*, http://admin.taiwan.net.tw/statistics/File/201001/2010_1monthly.pdf.

[2] Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet? *Journal of Interactive Marketing*, 18(1), 38–52.

[3] Litvin, S.W., Goldsmith, R.E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism Management*, 29(3), 458-468.

[4] Carson, D. (2008). The “Blogosphere” as a Market Research Tool for Tourism Destinations: A Case Study of Australia”s Northern Territory, *Journal of Vacation Marketing,* 14(2), pp.111-119.

[5] Akehurst, G. (2009). User generated content: The use of blogs for tourism organizations and tourism consumers, *Service Business*, 3 (1), 51-61.

[6] Serena Volo (2010), Bloggers” reported tourist experiences: Their utility as a tourism data source and their effect on prospective tourists, Journal of Vacation Marketing, 16,297-311

[7] Liu, B. (2010). Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing*, Second Edition, (editors: N. Indurkhya and F. J. Damerau).

[8] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002), pp. 79–86.

[9] Liu, B., Hu, M., and Cheng, J., “Opinion Observer:Analyzing and comparing opinions on the web,” in *Proc. of the 14th international Conference on World Wide Web*, pp.342-351, 2005.

[10] CKIP (Chinese Knowledge and Information Processing) : http://ckip.iis.sinica.edu.tw/ CKIP//index.htm