Extraction of hints and advice from hotel reviews for improving small hotel management

Hirokawa, Sachio
Research Institute for Information Technology, Kyushu University

Okada, Makoto
Osaka Prefecture University

Hashimoto, Kiyota
Osaka Prefecture University

出版情報：Proceedings of the 2012 IEEE 14th International Conference on Commerce and Enterprise Computing, CEC 2012, pp.166-170, 2012-12-01
バージョン：
権利関係：

Extraction of Hints and Advice from Hotel Reviews for Improving Small Hotel Management

Sachio Hirokawa
Research Institute for Information Technology,
Kyushu University, Fukuoka, Japan
hirokawa@cc.kyushu-u.ac.jp

Makoto Okada, Kiyota Hashimoto
Osaka Prefecture University
okada@mi.s.osakafu-u.ac.jp, hash@kis.osakafu-u.ac.jp

***Abstract*— There are various kinds of and huge amounts of hotel information both from providers and customers. Hotel information by hotels and travel agents are reliable. A much large number of reviews by general users are available, which might be less reliable compared to the official ones. However, those reviews are helpful, since we can hear their personal experience and opinion. This paper proposes a method to find hints and advice for improving small hotel management by extracting and analyzing the feature words of reviews. This paper focuses on the secondary major words and compares their occurrence probabilities in the business customer's review with the family customer's review. A visual interpretation is proposed by mapping the feature words on Word Net.**

Keywords-Hotel Reiew; Blog; Feature Extraction; WordNet; SME

I. INTRODUCTION

A. Extraction of Tourism Information on Web

By development of the Internet, a lot of tourism information is overflowing on Web. The information by the hotel or a tourist agent is fair and can be trusted. However, it tends to less interesting and less impressive. Compared with those official information, blogs are much more interesting, since they contain personal experience and opinions, although they might be unreliable.

There are two types of blogs -- personal blogs and community blogs. A personal blog is written by a blogger and contains his/her experience and opinions. A community blogs are written by crowds of users on their common concerns. TripAdvisor.com provides a forum of community blogs. Any user can write his/her opinion concerning to tourism. This paper analyzes the hotel reviews of the site which are written in Japanese.

There are researches and services which analyze reputation information using blog documents. For example, the trend analysis investigates popular words and their movement. They analyze which portion and which attributes of target goods are evaluated in blogs. In such research, the pairs of an object and an evaluation value are analytic unit. Positive/Negative analysis of blogs are popular which tries to evaluate blogs whether they are favorable or critical.

There are researches which focus tourism information as theme of the information extraction from Web or newspaper articles. Matrienez et. al. (Martinez2009) proposes an NLP interface for tourism search engine. (Esparcia2010) shows a possible services in extracting name, place, price, time and period from tourism information. This service would be a basis for updating and providing necessary information for individual requirement. The quantity of the information dissemination by ordinary users increases by leaps and bounds in recent years. Much attention have been paid for the concrete experiences and reputation information by ordinary users. (Ishino2011) proposes a classification method of tourism blogs using machine learning technology with several attributes, such as the names of tours, location and destination. (Hao2010) proposes a method of regional feature extraction. Yin et. al. (Yin2010) proposes a method to make a variety of rankings for the same search result by choosing appropriate feature words. Nakatoh et. al. (Nakatoh2011) uses WordNet to compare the difference of feature words extracted from tourism blogs of several areas.

The main concern of the conventional research of tourism information has been to help a general user. The purpose of the trend analysis using blog is the information dissemination to the general user. As far as the authors know, there is no research of harnessing these pieces of information for activity of tourism industry. This paper proposes a method for extracting hints and advice from hotel reviews for improving SME(Small and Medium Enterprises) hotel management.

B. Hotel Reviews on TripAdvisor

This paper analyzes the hotel reviews of TripAdvisor.com. However, the purpose is not the general analysis such as positive/negative evaluation of the reviews or trend analysis which a large-scale hotel takes notice of. It is the purpose that a hotel of a minor scale notices the hint for employing the special feature of its own. This paper compares the hotel reviews by their categories, by their locality and by the second major feature words that appear in the reviews.

The hotel reviews were collected from the TripAdvisor, a tourism portal site, as the target of analysis. The site provides more than 50 million "word-of-mouth" data concerning to hotels, sightseeing spots and restaurants around the world. They have photos as well as reviews. This paper analyzes 82,720 reviews written in Japanese.

Figiure1. *An example of Review in "TripAdvisor"*

The hotel reviews were collected from the TripAdvisor, a tourism portal site, as the target of analysis. The site provides more than 50 million "word-of-mouth" data concerning to hotels, sightseeing spots and restaurants around the world. They have photos as well as reviews. This paper analyzes 82,720 reviews written in Japanese.

II. BASIC ANALYSIS OF DATA

This section describes a basis analysis of the data using the location and the category of a review. Table I(a) shows the number of reviews with respect to area and category. The most largest category is "business" which has around 30% of reviews, followed by "couple", "family" and "solo" all of which have 20%. The smallest category is "friend" which has 10%. The largest area is "Kanto" which has 30%, followed by "Kyushu", "Kinki" and "Chubu" which have around 15-20%. The other area, "Tohoku", "Chugoku" and "Shikoku" have less than 6%. The most dominant users are "business" customers in "Kanto" where the capital Tokyo is. This analysis would not give any useful hints for small hotels in local area outside of "Kanto".

Table I(b) shows the share of each areas in a category. "Kanto" dominates around 30% in every category. It will be worth while to note that "Kyushu" has 20% in "family", "couple" and "friend" categories. The reason why "Tohoku" has the lowest share in every category is not clear from this table alone. But, we would guess that it is because of the earthquake and the nuclear accident. Table I(c) shows the share of each category in an area. We can see that "business" customers occupy 30% in all area except for "Kyushu" area, where the share of "family" and "couple" customers are larger than that of "business" customers.

Table I Category and Area

(a) The number of reviews

	To	Ka	Cb	Ki	Cg	Si	Ky	Total
fml	945	5901	3215	2342	662	491	5240	18796
bus	1864	7348	3427	4544	1417	937	4151	23688
cp	924	6127	3017	3443	667	526	4937	19641
sol	828	3541	1641	2615	653	448	2521	12247
frd	407	2456	1200	1606	327	246	2106	8348
Ttl	4968	25373	12500	14550	3726	2648	18955	82720

(b) The share of an area in a category

	To	Ka	Cb	Ki	Cg	Si	Ky
Family	0.05	0.31	0.17	0.12	0.04	0.03	0.28
Business	0.08	0.31	0.14	0.19	0.06	0.04	0.18
Couple	0.05	0.31	0.15	0.18	0.03	0.03	0.25
Solo	0.07	0.29	0.13	0.21	0.05	0.04	0.21
friend	0.05	0.29	0.14	0.19	0.04	0.03	0.25

(c) The share of a category in an area

	To	Ka	Cb	Ki	Cg	Si	Ky
Family	0.19	0.23	0.26	0.16	0.18	0.19	0.28
Business	0.38	0.29	0.27	0.31	0.38	0.35	0.22
Couple	0.19	0.24	0.24	0.24	0.18	0.20	0.26
Solo	0.17	0.14	0.13	0.18	0.18	0.17	0.13
friend	0.08	0.10	0.10	0.11	0.09	0.09	0.11

III. COMPARISON OF REVIEWS IN CATEGORY

A. *Feature Words of Category*

This section compares the feature words of categories. Table II shows the top 20 frequent words of all reviews. The words, "man", "visit", "submission" and "experience" are common words in all reviews. The word, "sanitariness", "service", "feeling" and "assessment" would concern with the guests' opinion. The words found in "business" reviews such as "price", "location", "railway station" would be the key issue of business customers and the criteria how they choose the hotel.

The above guess would be reasonable. However, justification of the arguments is not strong enough to convince ourselves and is not useful for small hotel management. The rest of this section focuses on detailed comparison of "business" and "family".

Table II. Top 20 Frequent Words

df	word	df	word
84255	man	62595	assessment
82720	visit	61569	theme
82720	submission	61294	word-of-mouth
82720	experience	53699	hotel
75840	sanitariness	46504	room
74898	land site	27741	railway station
74559	service	25752	exercise
74427	feeling	24277	work
74003	value	23292	breakfast
67031	guest room	22263	stay

The followings are the top 10 feature words of reviews in "business" category. From these words, we guess "a hotel in a shopping quarter in front of a railway station and close to a convenience store" would be an ideal hotel for business trip.

work, business trip, business, single, shopping quarter, youth, eating and drinking, ,in front of a station, convenience store, railway station

On the other hand, the followings are the top 10 words of reviews in "family" category. These words remind us the situation where parents and their children stay at a hotel. The words of "daughter" and "mam" would imply the requirement for women.

family, child, age, with children, daughter, accompany, children, mam, joy, parents

B. *Business vs Family in Word Frequency and Word Occurrence Probability*

This section gives a much more sharp comparison of feature words of reviews of "business" category and of "family" category. We used the difference diff(wi) = pr(business|wi)-pr(family|wi) of probabilities. Here pr(X|wi) is the probability of a review that contains the word wi to be in category X. To be precise, we used "normalized" probability as follows: pr(X|wi) = (#{d|(d in X) and (wi in d)}/(#{d|wi in d}))/(#X/#D), where #X is the number of reviews in the category X and #D is the total number of all reviews.

Table III. Feature Words Compared

(a) Feature Words of "Business"

word	df(w)	diff	pr(b|wi)	pr(f|wi)
work	24277	-3.3814	0.0259	3.4073
lan	1018	-1.7428	0.2832	2.026
single	3442	-1.5145	0.2711	1.7856
inn	1364	-1.2761	0.4162	1.6923
shopping quarter	3701	-1.2682	0.4340	1.7022
business	15271	-1.2270	0.5196	1.7466
subway train	2313	-1.0786	0.4795	1.5581
railway station	27741	-0.9798	0.5281	1.5079
convenience store	6234	-0.9450	0.5775	1.5225
internet	2106	-0.9371	0.5287	1.4658

(b) Feature Words of "Family"

word	df(w)	diff	pr(b|wi)	pr(f|wi)
family	19900	4.1040	4.1568	0.0528
child	4113	3.6617	3.7407	0.0790
accompany	2090	2.5027	2.7485	0.2458
adult	1344	2.4789	2.6556	0.1767
pool	3592	2.0673	2.2899	0.2226
beach	2029	1.9283	2.0367	0.1084
together	1198	1.5894	2.0645	0.4751
summer	1109	1.5702	1.9617	0.3915
resort	4118	1.5267	1.7591	0.2324
sea	5144	1.4817	1.7252	0.2435

It is interesting to note that "bullet train" is in the 9th position of feature words of "business" category, which can be considered as locations of business hotels. The word "internet" would imply the demand of internet connection in the hotel. Newly found feature words for "family" category are "beach", "pool", "summer", "resort" and "sea", which remind us an image in which a family goes pools and sea in their summer holidays.

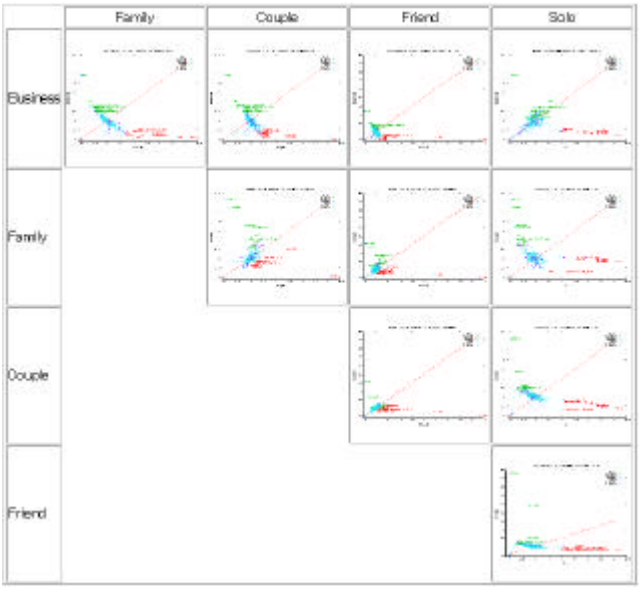

Figure 2. Ratio of Word Occurrences in Category

Figure 2 display the two dimensional comparison of the probabilities of pr(X|wi) and pr(Y|wi). Pr(X|wi) is the probability of a review that contains the word wi to be in the category X. We can see a correlation between "business" and "solo" and a correlation between "family" and "couple". On the other hand, there is a negative correlation between "business"and "family". A lesson we should learn fro these figures is that a small hotel should not run after the two categories of "business" and "family". They have different requirements which cannot be fulfilled unless the hotel has a large capacity. A small hotel should focus on "family" or on other category.

The next section analyzes "business" category and "family" category with secondary major words for discovering some hints for small hotel management.

IV. Comparison of Major Words vs Secondary Major Words for SME

Figure 3 displays the graph of the left upper corner, i.e. Business vs Family, of Figure 2. Each point represents a word wi and their probability pr(business|wi) as the y-axis and pr(family|wi) as the x-axis. We can see that there is a negative correlation between pr(business|wi) and pr(family|wi). The shape of the points represent the frequencies of the words as follows. A green point represents a major word with df(wi)>50000. In other words, a green point means that there

are more than 50000 reviews that contain the word wi. A green star mark means that 50000>=df(w)>10000. The pink square mark means that 10000>=df(w)>5000. Blue lozenge mark means that 5000>=df(wi)>1000. The green words and the red words that locate far away from the diagonal line are the feature words of each category. Thus, "family", "child","accompany","adult", "pool", etc in the lower right corner are the feature words of the category "family". On the other hand, the words "work", "lan", "single","inn" etc. in the left upper corner are the feature words of the category "business". We can observe that the frequencies of those words are less than 5000.

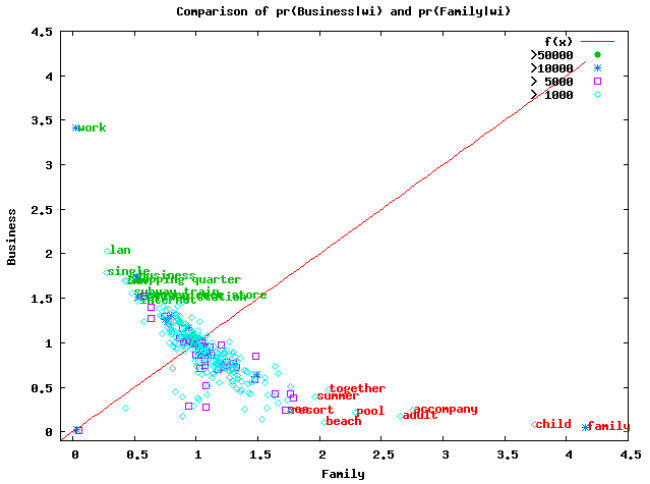

Figure 3. Comparison of Word Occurrence Probability

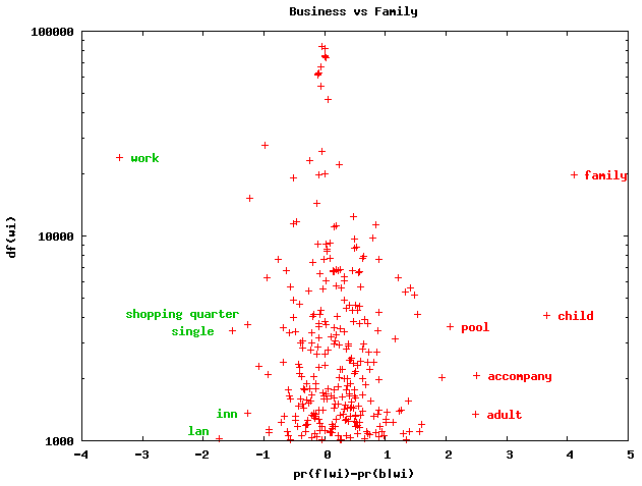

Figure 4. Frequency and Probability Gap

The points which are far away from the diagonal line figure 3 are the feature word of "business" and "family". Figure 4 displays the same data with different axes. The x-axis displays the difference pr(business|wi) and pr(family|wi). The y-axis displays the frequency df(wi) of the word. We can observe that there are many words with large differences of probabilities with frequencies of less than 5000. These second major words can be used as management strategy of small hotels.

V. INTERPRETATION ON WORDNET

In the previous section, we proposed a method of feature extraction of hotel reviews in two different categories. The feature words are chosen according to the difference of the probabilities of the word in the two categories. The method can be applied assuming that a small hotel is competing with a large hotel with a variety of services from family customers to business customers.

The feature selection, however, is not enough for hints and advice. Because a simple list of words requires some further interpretation. We apply the method of mapping words on WordNet as introduced in (Nakatoh2011). The two kinds of feature words are colored with red and green and are mapped on WordNet. In Figure 5, the red nodes represent the feature words of "business", while the green nodes represent the ones of "family". The hypernym of green nodes, i.e. the feature words of family customers, is "individual". The hypernym of red nodes, i.e. the feature words of business customers, is "artefact". The two notions capture the demands of main clients of the two categories.

Figure 5. Feature Words on WordNet

VI. CONCLUSION AND FURTHER WORK

It is observed that frequency distribution of many data on Web follows Zipf's law. This applies to the frequencies of words in hotel reviews. Only a few words are very popular, while most words appear in a few reviews. Those diversified words of customers are valuable information for improving hotels. However, a small hotel would not able to respond all of these requests. A SME hotels have to focus on their guests and guests' voice for providing satisfactory stay. It is necessary to know the strength of own hotel and the requirement of guests.

The paper proposed a method to extract hints and advice from hotel reviews for improving small hotel management to compete with large hotels in the same area. The distinction are made in considering categories and in focusing second major words. The words of large difference in the probabilities are chosen as feature words. The extracted feature words are mapped with 2 colors on WordNet. Hypernym of the feature words of a category helps the interpretation of the feature words. The two categories of "business" and "family" are compared as an example.

By implementing a search interface, we plan to construct an interactive system to compare and analyze any situation and requirements.

REFERENCES

(Wu2011) X. Wu, S. Hirokawa, C. Yin, T. Nakatoh, Y. Tabata, Extraction and Comparison of Tourism Information on the Web, Proc. AROB2011, 2011

(Yin2010 C. Yin, T. Nakatoh, S. Hirokawa, X. Wu, J. Zeng, A proposal of search engine " XYZ "for tourism events, Proc. JCAI2010, 2010

(Esparcia2010) S. Esparcia, V. Sanchez-Anguix, E. Argente, A. Garcia-Fornes, V. Julian, Integratiing Information Extraction Agents into a Tourism Recommender System, Proc. HAIS2010, Springer LNAI 6077, pp.193-200, 2010

(Hao2010) Q. Hao, R. Cai, Ch.Wang, R. Xiao, J.-M. Yang, Y. Pang, L. Zhang, Equip Tourist with Knowledge Mined from Travelogues, Proc. WWW2010, pp.401-410, 2010.

(Ishino2011) Aya Ishino, Hidetsugu Nanba, Toshiyuki Takezawa, Automatic Compilation of an Online Travel Portal from Automatically Extracted Travel Blog Entries. Proceedings of the 18th international Conference on Information Technology and Travel &Tourism (ENTER2011), 2011

(Nakatoh2011) Tetsuya Nakatoh , Chengjiu Yin , Hiroki Matsuura , Sachio Hirokawa : Visualization of Tourism Information using WordNet, Proc. Of the 3rd International Conference on Awareness Science and Technology, 413-418, 2011-09-27

(Martinez2009) J. M. Ruiz-Martinez, D. Castellanos-Nieves, R. Valencia-Garcia, J. T. Fernandez-Brieis, F. Garcia-Sanchez, P. J. Vivancos- Vincente, J. S. Castejon-Garrido, J. B. Camon, R. Martinez -Bejar : "Accessing Touristic Knowledge Bases through a Natural Language Interface", Proc. PKAW2008, Springer LNAI 5465, pp.147-160, 2009

(Yin2010) C. Yin, T. Nakatoh, S. Hirokawa, X. Wu, J. Zeng,A proposal of search engine XYZ for tourism events,Proc. JCAI (International Joint Conference on Artificial Intelligence) Vol.1,pp.178-181,2010