Automatic Restaurant Information and Keyword Extraction by Mining Blog Data for Chinese Restaurant Search

Chien-Li Chou[1(✉)], Min-Ho Tsai[1], Chien-Ho Chao[1], Hsiao-Jung Lin[1], Hua-Tsung Chen[2], Suh-Yin Lee[1], and Chien-Peng Ho[3]

[1] Department of Computer Science, National Chiao Tung University, 1001 Dahsueh Road, Hsinchu 30010, Taiwan
{fallwind, sylee}@cs.nctu.edu.tw, {shjk42l0, pass5l8224, smilecatxiii}@gmail.com
[2] Information and Communications Technology Lab, National Chiao Tung University, 1001 Dahsueh Road, Hsinchu 30010, Taiwan
huatsung@cs.nctu.edu.tw
[3] ICL/Industrial Technology Research Institute, 195 Chung Hsing Road, Section 4, Chutung, Hsinchu, Taiwan
cpho@itri.org.tw

Abstract. Restaurant search and recommendation system is a very popular service in many countries. In those systems, most of the restaurant information such as restaurant name, address, phone number, and introduction are collected manually. In this paper, we propose a restaurant information extraction method which can automatically extract restaurant information from online reviews of restaurants in blogs. In addition, by calculating TFIDFs of words in blog posts, the hot keywords can be discovered and ranked. For restaurant search, users are allowed to search by keywords, areas, and/or extracted hot keywords. The experimental results show that the proposed method can achieve over 90 % average accuracy of hot keyword extraction and about 95 % mean average precision for restaurant search. In user study, the fact that the proposed system is more useful than Google search in restaurant search is presented.

Keywords: Information retrieval · Opinion mining · TFIDF · Food and restaurants · Restaurant search

1 Introduction

With the rapid growth and affordable cost of Internet bandwidth, more and more web contents are generated by not only business content providers but also customers and users. Nowadays, blogs, the web pages for users to post their words, are widely spread and used. People post their moods, thoughts, and comments for something such as what they buy, where they go, and what they eat. According to the statistics from MBAonline.com in 2012, two million blog posts were written in one day. Such a huge

W.-C. Peng et al. (Eds.): PAKDD 2014 Workshops, LNAI 8643, pp. 700–711, 2014.
DOI: 10.1007/978-3-319-13186-3_62

number of data make the search results noisy and redundant. Therefore, mining useful information in such big data becomes a vital issue.

Many blog posts are written for recording the dining. Users write down their comments for the food and environment in restaurants they went to. This kind of blog posts is not only a record of life but also useful information for other people. For example, when people want to have dinner in an unfamiliar city, they usually search the reviews about restaurants there on the Internet by keywords. However, users have to spend much time to find and read the unorganized reviews. People usually want to collect more reviews about the desired restaurant to confirm if it is really good. Thus, an automatic restaurant information and keyword extraction system can reduce the time spent on collecting the similar reviews of restaurants and can extract correct restaurant information such as address and phone number. Some websites such as Yelp [1], Tabelog [2], and HOT PEPPER [3] provides restaurant search services to users for America, English, and Japanese restaurants. For Chinese restaurants, there are few restaurant search sites as good as the above sites. Therefore, we focus on restaurant information and keyword extraction for Chinese restaurants by mining contents of blog posts.

In this paper, we propose a method for automatic restaurant information and keyword extraction based on the techniques of information retrieval and pattern mining. Using the extracted information and keywords, we can develop a crowd sourcing restaurant search system to view the real comments from other people instead of business campaigns.

The remainder of this paper is organized as follows. The related literatures are reviewed in Sect. 2. The proposed method for restaurant information and keyword extraction is described in detail in Sect. 3. In Sect. 4, comprehensive experiments including quantitative and qualitative evaluations are conducted and the experimental results are presented. Finally, we conclude this work in Sect. 5.

2 Related Work

To extract the restaurant information, such as restaurant name, address, and phone number, from unstructured text content of blog posts, called "named entity recognition [4]", many studies for English websites were well conducted [5, 6]. However, it is hard to recognize the named entities in Chinese since the quality of Chinese word segmentation technique is insufficient to segment the correct phrase of named entities.

Many researchers focused on opinion mining from the online reviews. Hu et al. [7] created rules based on the number of frequencies from user reviews to extract the product related characteristics, and then classified the reviews into positive ones or negative ones by the extracted product characteristics. Jindal et al. [8] analyzed the components of the comparative sentences discovered from online reviews to extract the comparative targets and characteristics. The comparative sentences are categorized into four types. For each type appropriate rules were generated and applied to extract the comparative advantage target. Gu et al. [9] mined popular menu items of restaurants from web reviews by analyzing the post frequencies. Kato et al. [10] extracted the onomatopoeia from the online reviews by calculating the TFIDF of words and used the onomatopoeia to search the desired restaurants of users.

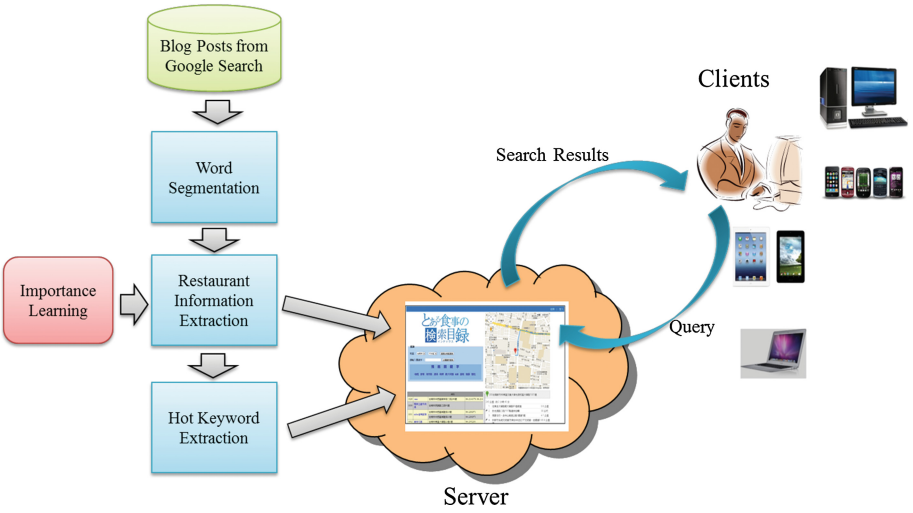

Fig. 1. The system framework

As an application, restaurant recommendation is an interesting topic for researchers. Yu et al. [11] developed a context-aware travel planning system which can recommend where to live, where to go, and where to dine. Gupta et al. [12] proposed a personalized location based restaurant recommendation system. The user preferences and location were taken into consideration to recommend the restaurants. Chu et al. [13] also developed a context-aware Chinese restaurant recommendation system. Kitayama et al. [14] constructed a restaurant information retrieval system by learning the relations among search properties based on the operational context. Association rule mining is applied to extract the relations among search properties.

3 Proposed Method

The proposed system can be divided in to two stages: (1) offline restaurant information and keyword extraction stage and (2) online restaurant search stage. The system framework is shown in Fig. 1. In the offline stage, the restaurant information such as store name, telephone number, and address is extracted, and the keywords of restaurants are then computed by analyzing contents of blog posts. In the online search and recommendation stage, users can type the keyword to search the desired restaurants, and the proposed system will return the related restaurants to users. Furthermore, the system can recommend the restaurant indirectly related to the keyword. The details of the two stages are described in the following.

3.1 Offline Information and Keyword Extraction Stage

As shown in Fig. 2, the offline stage consists of four steps is described below.

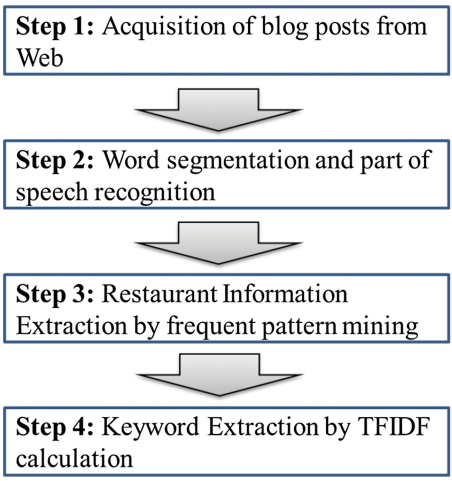

Fig. 2. The flow chart of offline information and keyword extraction

Step 1. Acquisition of blog posts

In this step, we collect the blog posts related to restaurants from Google search engine. We use one keyword "dining record" (in Chinese) and another keyword for area such as Taipei, etc. to search the blog posts containing reviews of restaurants. The keyword for area can reduce the noise in search results since it narrow the search space. Every returned result page $\boldsymbol{P}_i$ is then processed to extract the restaurant information.

Step 2. Word segmentation and part of speech (POS) recognition

For each page returned by Google, we parse the title T_i and main body B_i of the result page P_i. The parsed texts are then analyzed by CKIP Chinese word segmentation system [15]. CKIP can segment Chinese sentences into phrases or words and recognize their POS. We only keep nouns for the following steps since names of restaurants and keywords are usually nouns. Therefore, after the segmentation and filtering, a set of nouns $N^{T_i} = \{N_1^{T_i}, N_2^{T_i}, \ldots, N_X^{T_i}\}$ and another set $N^{B_i} = \{N_1^{B_i}, N_2^{B_i}, \ldots, N_Y^{B_i}\}$ are obtained from T_i and B_i, respectively.

Step 3. Restaurant Information Extraction

To extract restaurant information, first, we extract names of restaurants. According to observations, the title of a blog post for dining record usually consists of the name of the restaurant introduced in the post. That is, if a noun appears in N^{T_i} and N^{B_i} at the same time, it is most likely part of the name of the restaurant. If there are more than one nouns consecutively appearing in the same order in both the sets, the nouns are concatenated to form an extended noun as a candidate. For example, assume that the title of a blog post is "Noodle Store: the good place for lunch", and the main body is "Looking for lunch? Come to Noodle Store." After segmentation and filtering, the title is segmented into nouns "noodle", "store", "place", and "lunch", and the main body becomes "lunch", "noodle", and "store". Since the words "noodle" and "store" appear consecutively and in the same order in both sets, we can concatenate the two words to

"noodle store." The word "lunch" appears in both sets, so it also becomes a candidate. "Noodle store" and "lunch" cannot be combined since the orders of these two words are not the same. To select one of the candidates to be the name of the restaurant, we define the name score S_{name} as

$$S_{name}(C) = \alpha \times Importance(C) + (1 - \alpha) \times Freq(C) \times Length(C), \quad (1)$$

where C is a candidate, α is the weight, $Freq(C)$ is the candidate appearing frequency in the post, $Length(C)$ is the number of words in C, and $Importance(C)$ is the importance of C. To define the Importance of a candidate, we randomly collect 300 names of restaurants as training data to train the importance of words. A name of restaurant is regarded as a transaction, and a word is regarded as an item. Frequent pattern mining [16, 17] is then applied to find the common words for names of restaurants. For a candidate C, words in C is regarded as items. Then we generate a set I_C containing all possible itemsets $\{I_1, I_2, \ldots\}$ for C. The importance of candidate C can be defined as

$$Importance(C) = \sum\nolimits_{I_i \in I_C} (Length(I_i) \times Support(I_i)), \text{ if } I_i \text{ is in } FPS, \quad (2)$$

where FPS is the frequent pattern set, $Length(I_i)$ is the number of items of I_i, and $Support(I_i)$ is the support of the frequent pattern corresponding to I_i. The candidate with the highest name score is selected to be the name of the restaurant in the post. After selecting the name of the restaurant, the other restaurant information is then extracted by searching the text nearby the name in main body. Usually, telephone numbers are in some specific formats, such as 0x-xxx-xxxx, 09xx-xxx-xxx, etc. Using this constraint, we can easily extract the phone number. Simultaneously, the address is extracted by searching the area names downloaded from the website of post office.

To validate the correctness of the extracted name, address and phone number of a post, we use the extracted name and the area as the keyword to search by Google. We select the top K results and repeat step 1 to 3. For each result, we can obtain a set of name, address, and phone number. The final restaurant information is decided by major voting scheme. If no information is voted by more than one page, we may select the wrong name of the restaurant. Hence, we select the next name candidate as the name, and repeat the above procedure until a correct name is found or no candidate can be processed.

Step 4. Keyword Extraction

Extracting keywords from the text of a blog post is an important task for restaurant search and indexing. For keyword extraction, we calculate the TFIDF value of every word. TF and IDF for a noun N_j in the post P_i are defined as

$$TF_{N_j}^{P_i} = \frac{n_{N_j}^{P_i}}{|N^{T_i}| + |N^{B_i}|} \quad (3)$$

$$IDF_{N_j} = \log \frac{|P|}{\left|\left\{i : N_j \in P_i\right\}\right|}, \quad (4)$$

where $n_{N_j}^{P_i}$ is the number of N_j appearing in P_i. And the TFIDF of a keyword N_j for a restaurant R is defined as

$$TFIDF_{N_j}^{R} = \sum_{P_i \in R} TF_{N_j}^{P_i} \times IDF_{N_j}. \quad (5)$$

The nouns with top 20 high TFIDF are selected as the keywords of the restaurant. For each area, we aggregate all the keywords of restaurants in the area, and accumulate the TFIDF values of the same keywords of different restaurants. The top 10 keywords are selected to be the area keywords, which can be recommended to users as hot keywords.

However, the long phrases may be segmented in the step of word segmentation, which causes that a long phrase is hard to be a keyword. Therefore, we propose a keyword expansion method to recover the long keywords. For each keyword of a restaurant, we search the main body to find the position of the keyword. If the previous and next words of the keyword in main body are nouns, we concatenate those words together to form an expanded keyword. For example, "noodle" is a keyword of the restaurant. "Seafood noodle" appears in the main body but the word "seafood" is not a keyword. We can concatenate "seafood" and "noodle" to form an expanded keyword since "noodle" is a keyword and the previous word of "noodle", "seafood," is a noun. That is, with keyword expansion, when a user search by long keyword, the system can still return the correct results.

3.2 Online Search Stage

Search by Keyword
Users can type their desired keywords or choose one of the hot keywords extracted from blog posts, as the interface shown in Fig. 3. The restaurants contain the keyword(s) are retrieved and ranked by the TFIDF of the keyword(s) in the restaurant. Users can view the related paragraphs of the corresponding blog posts in the system for judging if they go to the restaurant. As elaborated in Fig. 4, this mechanism emphasizes the texts related to the keyword and provides a quick review of the blog posts for the restaurant to users.

Search by Location
Users can search the nearby restaurants if the system obtains the location information from the device or inputted by users. Google Map is used to calculate the distances between the location of users and the restaurant addresses. Google navigation can plan the routes among multiple destinations. Search by location and search by keyword can work together.

4 Experimental Evaluation

4.1 Prototype System

We develop a prototype system to evaluate the proposed method for restaurant search. Figure 4 shows the interface of the proposed interface consisting of area search

Fig. 3. The search interface of the proposed system

Fig. 4. The prototype System

component, keyword search component, hot keyword search component, list of search results component, online map component, route planning component. The online map and route planning component applies Google Map API to acquire the user location and to plan the route to restaurants. By clicking a restaurant name listed in the search result, the page will show all the blog posts related to the restaurant, as shown in Fig. 5.

4.2 Experimental Setting

We collect blog posts for restaurants in 19 areas in Taiwan as listed in Table 1. The keyword for Google search is set to "dining record AREA" (in Chinese), where AREA is the name of an area. At least 50 restaurants for each area and at least 4 blog posts for each restaurant are collected. Totally 1099 restaurants and 5483 blog posts are used to conduct the experiments. For better evaluation of the proposed method, we conduct both quantitative and qualitative experiments.

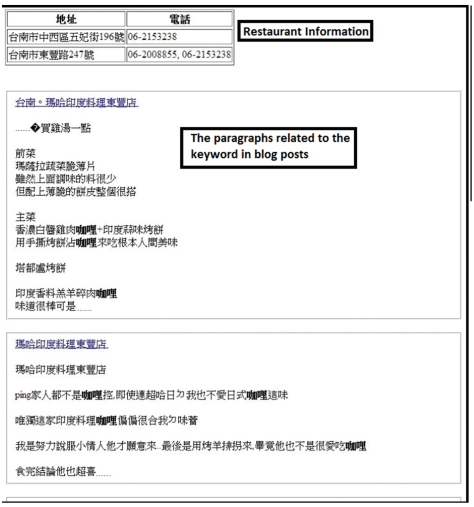

Fig. 5. The quick review of the blog posts for the selected restaurant

Table 1. List of Taiwan areas used to search the blog posts

List of the areas used in the proposed method		
台北市 (Taipei city)	彰化縣 (Changhua county)	屏東縣 (Pingtung county)
新北市 (New Taipei city)	雲林縣 (Yunlin county)	基隆市 (Keelung city)
桃園縣 (Taoyuan county)	南投縣 (Nantou county)	宜蘭市 (Yilan county)
新竹縣 (Hsinchu county)	嘉義縣 (Chiayi county)	花蓮縣 (Hualien county)
新竹市 (Hsinchu city)	嘉義市 (Chiayi city)	台栗縣 (Taitung county)
苗栗縣 (Miaoli county)	台南市 (Tainan city)	
台中市 (Taichung city)	高雄市 (Kaohsiung city)	

Quantitative Experiments

For quantitative experiments, we apply three measurements for evaluating the accuracy of hot keyword extraction, the average precision (AP) of search results, and the mean average precision (MAP) of search results. Within an area, the accuracy of hot keyword extraction is defined as

$$\mathrm{Acc} = \frac{\#\ representative\ hot\ keywords}{\#\ hot\ keywords\ extracted}. \tag{6}$$

Whether a hot keyword is representative or not is decided by users.

Given a set of hot keywords $H = \{H_1, H_2, \ldots, H_Z\}$, the AP within an area is defined as

$$\mathrm{AP} = \frac{\sum_i Precision(H_i)}{|H|}, \tag{7}$$

where *Precision* (H_i) is the precision of the search results of H_i. And the MAP for the proposed system is defined as

$$\text{MAP} = \sum\nolimits_i AP_i / \# \, areas. \tag{8}$$

Qualitative Experiments
We invite 12 university students to rate the user experiences of our proposed system. Each participant performs 19 search tasks (One search task for one area) and rates score $1 \sim 5$ on the following options.

- Convenience (1 is not convenient for users, and 5 is convenient.)
- Practicability (1 is less practicability, and 5 is more practicability.)
- Smoothness on use (1 is hard on use, and 5 is smooth on use.)
- Is it better than Google search for restaurant search? (1 means Google is much better than the proposed system, and 5 means the proposed system is much better than Google.)

4.3 Experimental Results

Figure 6 shows the accuracy of hot keyword extraction. Most of the accuracies of hot keyword extraction are greater than 90 %. That is, the extracted hot keywords are representative enough for users. However, in Taoyuan County, the accuracy of hot keyword extraction is only 70 %. It means that in all extracted 10 hot keywords, three of them are uninformative to be search keywords. The reason is that some of the blog posts collected in the area of Taoyuan County are written by the same author. The author often uses nicknames in his posts, and the nicknames are usually extracted as keywords since the TF and IDF of the words are high. The average accuracy of hot keyword extraction is 91.58 %, which provides sufficient information for users.

The average precisions are illustrated in Fig. 7. Same as the accuracy of hot keyword extraction, most of APs are high enough to provide correct search results to users. In some specific areas, about only 80 % APs are achieved, and we observe that some conditions of the areas. First, the areas with low APs are less famous in restaurants so that few people write the blog posts for restaurants in those areas. Second, a famous night market is in the area. Blog posts written for the night market make the restaurant information noisy because there are many vendors in the night market. The author introduces multiple vendors, and the information of different vendors is mixed and hard to separate. As a result, a blog post focuses on multiple restaurants makes the search results noisy. The MAP for restaurant search is 94.85 %, which is high enough for users to obtain desired restaurant information.

4.4 User Study

In the qualitative experiments, the average convenience score rated by 12 participants is 4.25, the average practicability score is 3.75, and the average smoothness score is 4.0. The result shows that the participants put a premium on the proposed system in

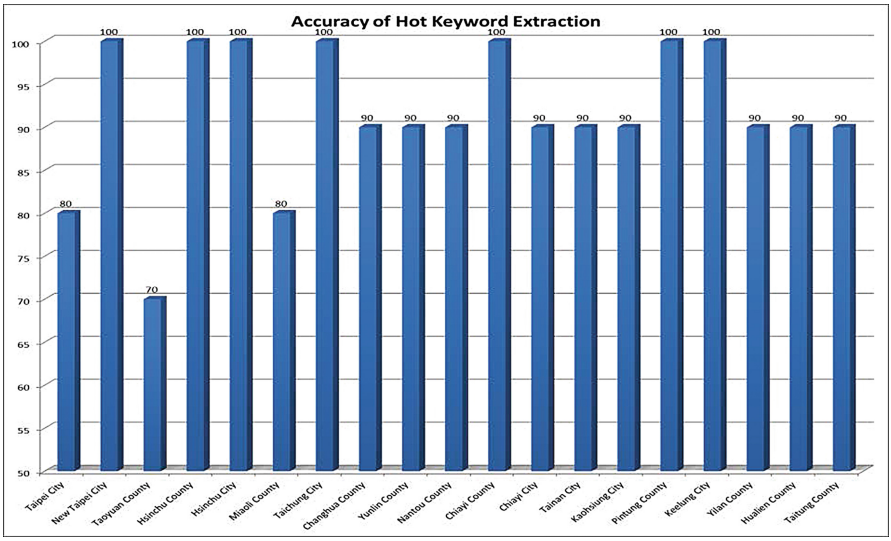

Fig. 6. The accuracy of hot keyword extraction

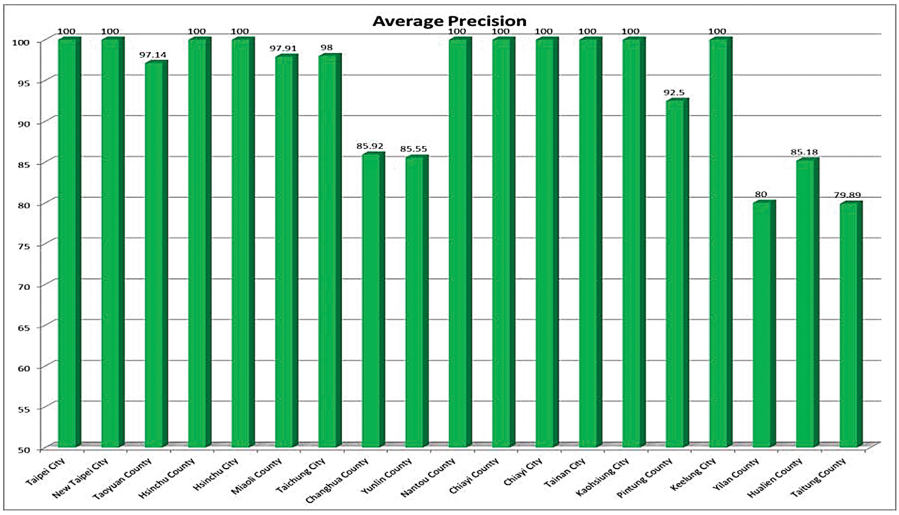

Fig. 7. The average precision of areas

convenience and smoothness. That is, the proposed system is well designed for good user experience. Because the number of restaurants is not sufficient to make every participant satisfied, the practicability score is lower than the other two scores. In the last question "Is it better than Google search for restaurant search?" of this user study, the average score is 4.75. Most of all participants give 5 points for this question. It depicts that the proposed system is really useful for users who want to search a restaurant.

5 Conclusion

We proposed an automatic restaurant information and keyword extraction system by mining the blog posts data for restaurant search. From the large number of blog posts, the restaurant information such as restaurant name, address, and phone number can be automatically extracted and validated by pattern mining techniques. By using the TFIDF approach, the representative words are extracted to be hot keywords for users' references. The experiment results show that the average accuracy of hot keyword extraction is over 90 % and the MAP of restaurant search is about 95 %. In user study, we observe that the extracted hot keywords and the restaurant information are more compact and practicable than the information searched from Google search.

References

1. Yelp. http://www.yelp.com
2. Tabelog. http://tabelog.com/
3. HOT PEPPER. http://www.hotpepper.jp/
4. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Invest. **30**(1), 3–26 (2007)
5. Alfonseca, E., Manandhar, S.: An unsupervised method for general named entity recognition and automated concept discovery. In: 1st International Conference on General WordNet, pp. 1–9 (2002)
6. Satoshi, S., Nobata, C.: Definition, dictionaries and tagger for extended named entity hierarchy. In: 4th International Conference on Language Resources and Evaluation, pp. 1977–1980 (2004)
7. Hu, M. Liu, B.: Mining and summarizing customer reviews. In: 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177 (2004)
8. Jindal, N., Liu, B.: Mining comparative sentences and relations. In: 21th National Conference on Artificial Intelligence, vol. 2, pp. 1331-1336 (2006)
9. Gu, Y.H., Yoo, S.J.: Mining popular menu items of a restaurant from web reviews. In: Gong, Z., Luo, X., Chen, J., Lei, J., Wang, F.L. (eds.) WISM 2011, Part II. LNCS, vol. 6988, pp. 242–250. Springer, Heidelberg (2011)
10. Kato, A., Fukazawa, Y., Sato, T., Mori, T.: Extraction of onomatopoeia used for foods from food reviews and its application to restaurant search. In: 21st International Conference Companion on World Wide Web, pp. 719–728 (2012)
11. Yu, C.C., Chang, H.P.: Towards Context-Aware Recommendation for Personalized Mobile Travel Planning. In: Vinh, P.C., Hung, N.M., Tung, N.T., Suzuki, J. (eds.) ICCASA 2012. LNCS, vol. 109, pp. 121–130. Springer, Heidelberg (2012)
12. Gupta, A., Singh, K.: Location based personalized restaurant recommendation system for mobile environments. In: International Conference on Advances in Computing, Communications and Informatics, pp. 507–511 (2013)
13. Chu, C.H., Wu, S.H.: A Chinese restaurant recommendation system based on mobile context-aware services. In: 14th International Conference on Mobile Data Management, vol. 2, pp. 116–118 (2013)

14. Kitayama, D., Matsuo, J., Sumiya, K.: Extracting relations among search properties based on the operational context of geographical information retrieval systems. In: Hong, B., Meng, X., Chen, L., Winiwarter, W., Song, W. (eds.) DASFAA Workshops 2013. LNCS, vol. 7827, pp. 179–192. Springer, Heidelberg (2013)
15. CKIP Chinese Word Segmentation System. http://ckipsvr.iis.sinica.edu.tw/
16. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: ACM SIGMOD International Conference on Management of Data, pp. 1–12 (2000)
17. Li, H., Wang, Y., Zhang, D., Zhang, M., Chang, E.Y.: PFP: parallel FP-growth for query recommendation. In: ACM Conference on Recommender Systems, pp. 107–114 (2008)