Understanding Travel Destinations From Structured Tourism Blogs

Lei Guo
School of Management, Harbin Institute of Technology, China, cnleiguo@hit.edu.cn

Ziru Li
School of Management, Harbin Institute of Technology, China

Wenjun Sun
School of Management, Harbin Institute of Technology, China

Understanding Travel Destinations From Structured Tourism Blogs

Lei Guo[1*], *Ziru Li*[1], *Wenjun Sun*[1]
[1]School of Management, Harbin Institute of Technology, China

Abstract: The increasing popularity of tourist generated content has created abundant opportunities for people to understand the opinions and experiences of prior tourists. However, till now no framework has been presented to automatically discover useful patterns from structured tourism blogs. In this paper, we present a method to mine the tourism information such as frequented spots and popular travel service within a given travel destination. These information can help us understand the travel destination and enable the website to recommend interesting travel spots. First, we introduce the method for compact pattern mining and sequential pattern mining. Then we propose a framework to analyze the structured tourism blogs. Particularly, sequential pattern mining was conducted to discover the frequented spots and their correlations. Then compact pattern mining was conducted to detect the spot associated travel service like shopping, etc. Finally, the experimental results based on an online tourism blog dataset (in Chinese) illustrate advantages of the proposed method.

Keywords: tourism blogs, sequential pattern, travel routes, travel service

1. INTRODUCTION

The rapid growth of Web 2.0 and social media content has change the way to generate and disseminate information. Tourists increasingly share their tourism experiences about travel destinations and travel services. At the same time, User-Generated Content (UGC) has provided people with a better way to understand travel destination and made them more and more rely on previous tourists' experiences to plan their trip [1]. Particularly, the online travel market has formed a large size in China. According to CNNIC, China's online travel booking users reached 133 million by June 2013. The report of iResearch shows that China's online travel market transactions amounted to 218.12 billion Yuan with an increase of 27.7% in 2013.

The increasing availability of UGC has created abundant opportunities for people to understand the opinions and experiences of prior tourists. In this area, many scholars conducted researches and found many interesting phenomena and significant findings. However, there are still great challenges for computers to understand the natural language especially Chinese language. In recent years, many innovations have been created in travel social media. For example, the GPS technology was applied to generate geo-location when taking photos. Significant results have been found based on geo-tagged photos. Furthermore, to make a better user experience, some online travel community platform, like qunar.com and mafengwo.cn began to provide users with a structured way to post blogs. People can post locations, relevant photos and reviews in their blogs. Structured tourism blogs make it available for us to extract this kind of travel routes. However, to the best of our knowledge, there is no framework has been presented to automatically discover useful patterns from structured tourism blogs.

In this work, we propose a data mining method to discover tourism information from structured tourism blogs. We try to explore:

- For a specific travel destination, where do most of the tourists come from?
- What are the popular spots in this destination and what are the correlations of these spots?
- When go to a spot, what kind of travel service (e.g. shopping, hotel) can we enjoy in?

To answer these questions, we first introduce a method to mine compact patterns and sequential patterns.

[*] Corresponding author. Email: cnleiguo@hit.edu.cn

Then this method is employed to construct a framework to discover the frequented spots and popular travel service from the structured tourism blogs. The contributions of this paper lie in two aspects. First, we propose a sequential patterns based framework to discover tourism information which is useful for potential tourists to understand an unfamiliar travel destination quickly. Second, we propose a method based on compact patterns to mine the spots associated service. With these information, online travel service providers can recommend spot associated popular service for tourists based on their locations.

The rest of this paper is organized as follows. Section 2 discusses the related works. Then we introduce sequential pattern mining method and present the framework in Section 3. Experimental results are shown in Section 4. Finally we conclude the paper in Section 5.

2. LITERATURE REVIEW

In recent years, the increasing of tourist generated content has raised many scholars' attention. Research on travel data mining mainly includes tourism blogs analysis and GPS routes mining which we will discuss in the following subsections.

2.1 Tourism blogs analysis

The tourism blog can be retrieved by search engine. Sharda et al. [2] presented a conceptual model to build a Blog Visualizer system based on the tourism blogs retrieved by blog search engines like Google Blog Search engine. The Blog Visualizer can provide a virtual experience of the trip to support better tour planning. Davidov and Rappoport [3] presented a framework to discover road and transport network based on a small set of seed terms describing a geographical region.

With the development of online travel community, more studies began to analyze the information posted on professional tourism websites. Pang et al. [4] proposed a framework to discover location representative tags from tourism blogs and then selected relevant photos to visualize these tags. This framework can summarize a given tourist destinations both textually and visually. Yuan et al. [5] presented a method based on frequent pattern to extract interesting information from massive tourism blogs. Furthermore, Ye et al. [6] presented a method to extract feature of the tourist reviews about travel destinations. It provided an opportunity to understand the views and attitudes of tourists.

2.2 GPS routes mining

Zheng et al. [7] presented a model to mine interesting locations and classical travel sequences from users' GPS trajectories data. Fujisaka et al. [8] proposed a fundamental model to find geographic social patterns from user movement histories made by mass mobile micro-bloggers. Particularly, the emergence of geo-tagged photos has led to an amount of research [9]. Lu et al. [10] investigated the trip planning problem systematically. They presented a novel automatic trip planning framework by leveraging massive geo-tagged photos and textual travelogues. Arase et al. [11] proposed a method to detect people's frequent trip pattern. They categorized the photo collections based on trip themes and then mined frequent trip pattern for each trip theme. Zeng et al. [12] built a model to detect interesting points by incorporating the trajectories and geo-photos. Kurashima et al. [13] proposed a travel route recommendation method that makes use of the geo-tagged photos shared by social network sites.

3. METHODLOGY

3.1 Compact pattern and sequential pattern

Frequent pattern are item set that commonly appear in a dataset with frequency more than a specified threshold. Frequent pattern mining is one of the most popular data mining methods that was originally developed for market basket analysis [14]. Particularly, we introduce a compact pattern mining method as follows:

given a database $T = (t_1, . . ., t_m)$ of transactions where transaction $t_i = \{t_{i1}, . . ., t_{in}\}$ and a specified minimum relative support (rs_{min}), the task of compact patterns mining is to find all the compact 2 items sets (e.g. $\{t_{ij}, t_{i(j\pm1)}\}$) with relative support no less than rs_{min}. The difference between our method and the traditional frequent pattern mining is that only the items occurred adjacently is valid pattern. It means that there is no gap between the items.

As an illustration, Figure 1 shows a simple database with 5 transactions. With a minimum relative support of rs_{min} = 60%, a total of 2 compact pattern with 2 items can be found in this database.

Transactions	Item sets
1	{ *a b c d e* }
2	{ *b a d c e* }
3	{ *c a b d* }
4	{ *b c a e* }
5	{ *e c b d a* }

(a) A simple database with 5 transactions

2 items sets	Relative support
{ *a b* }	60%
{ *b c* }	60%

(b) the compact patterns with 2items (rs_{min} = 60%)

Figure 1. A simple example for compact pattern mining

Just like frequent pattern, compact pattern doesn't take into account the order of transactions. However, in many applications the orderings are significant [15]. In this work, it is interesting to know whether people travel to somewhere in sequence, e.g., travel to *Kuanzhai Alley* first and then go to *Wensu Park* later. Sequential pattern mining need to be employed in these applications. In this work, we define the task of sequential pattern mining as follow: given a database $V = (v_1, . . ., v_m)$ of vectors where vector $v_i = \{v_{i1}, . . ., v_{in}\}$ and a specified minimum relative support (rs_{min}), the task of sequential pattern mining is to find all the compact sequences with 2 items (e.g. $\{v_{ij}, v_{i(j+1)}\}$) with relative support no less than rs_{min}. Here, sequential patterns are more advanced than compact patterns and the only difference is that sequential patterns are items with direct correlations.

Vectors	Sequence
1	{ *a b c d e* }
2	{ *b a d c e* }
3	{ *c a b d* }
4	{ *b c a e* }
5	{ *e c b d a* }

(a) A simple database with 5 vectors

Sequences with 2 items	Relative support
{ *a b* }	40%
{ *b c* }	40%
{ *c a* }	40%

(b) the sequential patterns with 2items (rs_{min} = 40%)

Figure 2. A simple example for sequential pattern mining

As an illustration, Figure 2 shows a simple database with 5 vectors. With a minimum relative support of rs_{min} = 40%, a total of 3 sequential patterns with 2 items can be found in this database.

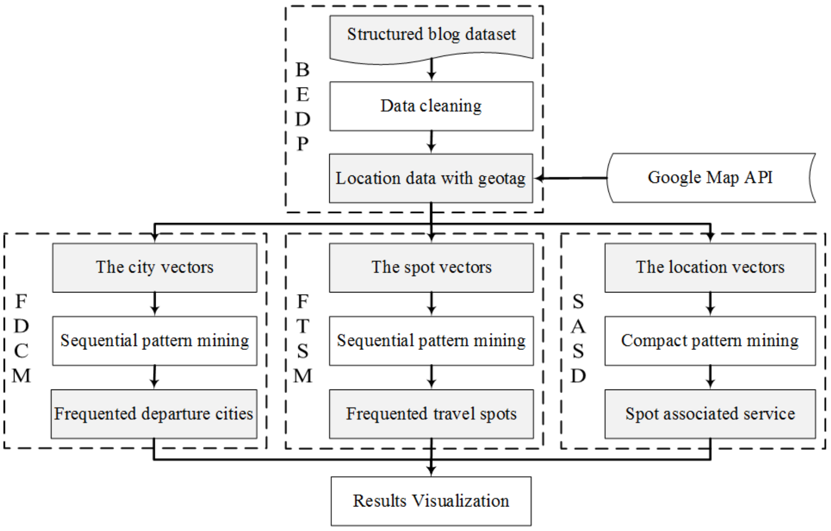


Figure 3. The method framework

3.2 The framework

We present a framework to mine valuable information from the structured tourism blogs. As shown in figure 3, the framework includes four parts as follows: blogs extraction and data preprocessing (BEDP), frequented departure cities mining (FDCM), frequented travel spots mining (FTSM) and spot associated service detection (SASD).

3.2.1 Blogs extraction and data preprocessing

Many methods and tools can be applied to crawl webpages. In BEDP subsystem, we developed a crawler to automatically download the structured tourism blogs. To ensure the quality of data, the blogs posted by unregistered users will be removed. Table 1 shows a sample of the data.

In previous study, Kurashima et al. [13] assumed that the collection of each photographer's geo-tagged photos is a sequence of visited locations. Similarly, we assume that the set of textual locations post on each blog is a sequence of travel route. To keep the sequence information, the field *Cities* and *TravelRoutes* were stored in class of vector. Particularly, each location in field *TravelRoutes* has a type label (e.g. spot, food and shopping).

In order to assign geotags to the textual location names, we developed a program based on Google Map API to obtain latitude and longitude of each city and spot.

Table 1. A sample of the structured tourism blog

Field name	Contents
Username	qunar_zhuang0
PostTime	2013-12-26
Title	The trip to *Chengdu*
Cities	{*Shanghai, Chengdu*}
TravelRoutes	{*Kuanzhai Alley* /spot, *Wensu Park* /food, *Jiong Box* /entertainment, *Dufu Thatched Cottage* /spot, *Jinsha Ruins* /spot, *Shufeng Yayun* /shopping}

3.2.2 Frequented departure cities mining

The FDSM subsystem is used to find the frequented departure cities to a specified travel destination from the database of *Cities*.

First, all the data items with only one city are removed and the remained data is stored as a vectors set. Then we conduct sequential pattern mining and the frequent sequence with 2 cities are generated from the *Cities* vectors. Finally, the departure cities and the frequent weight can be visualized on the map base on geotags.

3.2.3 Frequented travel spots mining

The FTSM subsystem is the main concern in this work. This subsystem provides a way to find the most popular travel spots and their correlations from the database of *TravelRoutes*.

There are various types of locations in the database of *TravelRoutes*. First, we extract all the spots locations from the database of *TravelRoutes* and then generate spots vectors in original order. For example, we can generate a spots vector {*Kuanzhai Alley*, *Dufu Thatched Cottage*, *Jinsha Ruins*} from field of *TravelRoutes* in Table 1.

To find the frequented travel spots, we conduct sequential pattern mining to generate the frequented spot sequences. Here we only generate sequences with two spots for two reasons. First, sequences with more spots are more like to be insignificant. Besides, Sequences with two spots are enough for us to construct a frequented spot network.

Using the sequential pattern mining, we can find frequented spots and their correlation weight (support). The larger the weight is, the more frequent the spots correlation is. Base on the sequential pattern, take spots as nodes and correlations as edges, we can generate a frequented spot network where the edge weights denote the sequence support.

3.2.4 Spot associated service detection

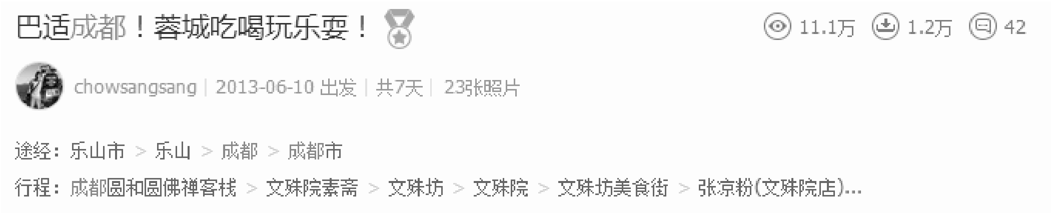
As mentioned above, we have presented a method to find the frequented spots. However, tourists are also interested in the tourism service like shopping and hotel. In the SASD subsystem, we present a method to detect the popular service around the spots. We name it *spot associated service*.

To this end, we conduct compact pattern mining to discover the frequented pattern with two locations from the database of *TravelRoutes*. After removing the patterns with only locations in type of spot or with none location in type of spot, we can find where the tourists went before or after they go to a specific spot. Hence we can recommend corresponding popular services to tourists based on their locations.

4. EXPERIMENTS AND RESULTS

4.1 Data

The data in this study were collected from qunar.com (NASDAQ: QUNR), one of the most popular Chinese travel agency. Qunar.com provides an online forum for customers to review their travel experiences. Figure 4 provides a sample blog list page where we can get the blog title and the travel cities. Figure 5 shows a sample blog content page where we can extract the travel routes.



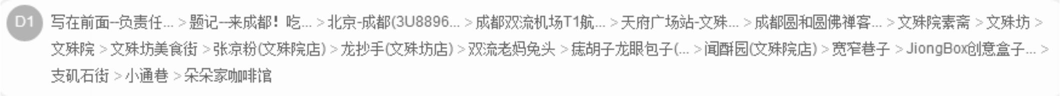
巴适成都！蓉城吃喝玩乐耍！ 11.1万 1.2万 42

chowsangsang | 2013-06-10 出发 | 共7天 | 23张照片

途经：乐山市 > 乐山 > 成都 > 成都市

行程：成都圆和圆佛禅客栈 > 文殊院素斋 > 文殊坊 > 文殊院 > 文殊坊美食街 > 张京粉(文殊院店)...

Figure 4. A page of blog list



D1 写在前面--负责任... > 题记--来成都！吃... > 北京-成都(3U8896... > 成都双流机场T1航... > 天府广场站-文殊... > 成都圆和圆佛禅客... > 文殊院素斋 > 文殊坊 > 文殊院 > 文殊坊美食街 > 张京粉(文殊院店) > 龙抄手(文殊坊店) > 双流老妈兔头 > 痣胡子龙眼包子(... > 闻酣园(文殊院店) > 宽窄巷子 > JiongBox创意盒子... > 支矶石街 > 小通巷 > 朵朵家咖啡馆

Figure 5. A page of travel routes

Chengdu, a famous tourism city in China, was chosen to be the travel destination for experiments. Keyword *Chengdu* (in Chinese) is used to retrieve data on the search engine of travel.qunar.com. All the tourism blog list, travel routes and location type (spot, hotel, food, shopping, entertainment) involve *Chengdu* from May 2011 to May 2014 were extracted to establish a structured blog dataset. To ensure the quality of data, the repeated blogs and blogs posted by unregistered users were removed and 18089 useful blogs were remained for experiments.

4.2 Frequented departure cities mining

As we can see in figure 4, the travel cities sequence (e.g. {*Leshan*, *Chengdu*}) was post on the page of blog list. Hence, sequential pattern mining method can be applied to discover the frequented departure cities from where tourist went to Chengdu mostly.

First, we need clean the blog list. Some blogger didn't post their departure city and only one city (*Chengdu*) is shown on the page. After removing this kind of single data, 6668 valid data were used to generate cities vectors. Then in the sequential pattern mining section, all the compact sequences with two cities were generated. In this work, the main concern is to find the frequented departure cities to Chengdu, only the sequences directed to *Chengdu* (e.g. {*Leshan*, *Chengdu*}) were remained in the final dataset. Finally, 1159 cities (nodes with geotag) and 1158 correlations (edges with weight) are visualized in Figure 6. Some cities far from Chengdu are not shown on the map. It's obvious that the visitors in Chengdu mostly came from Chongqing and Beijing.

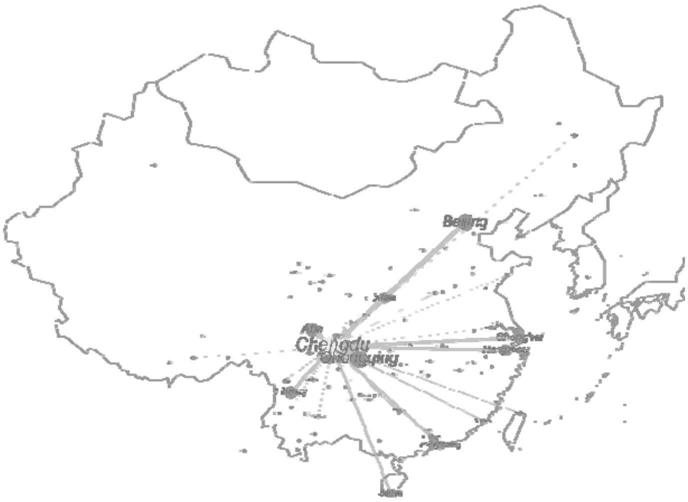


Figure 6. The frequented departure cities to Chengdu

4.3 Frequented travel spot mining

As we have obtained the tourist generated travel routes from the blogs, hence in this section, we want to discover the frequented spots in Chengdu.

To this end, we first extract all the Chengdu spots from the travel routes dataset. We keep the spots in original order and then construct 6847 spot route vectors. The names for the same spot could be slightly different in the dataset. To match spot names, we wrote a python program using an approximate fuzzy string matching technique called FuzzyWuzzy. Four algorithms are employed to calculate the strings similarity. The program takes a pair of spot names as input and returns the matching score of the names. We kept all name pairs whose matching scores are greater than 0.8. Then we manually reviewed matched spot names and merged them into one.

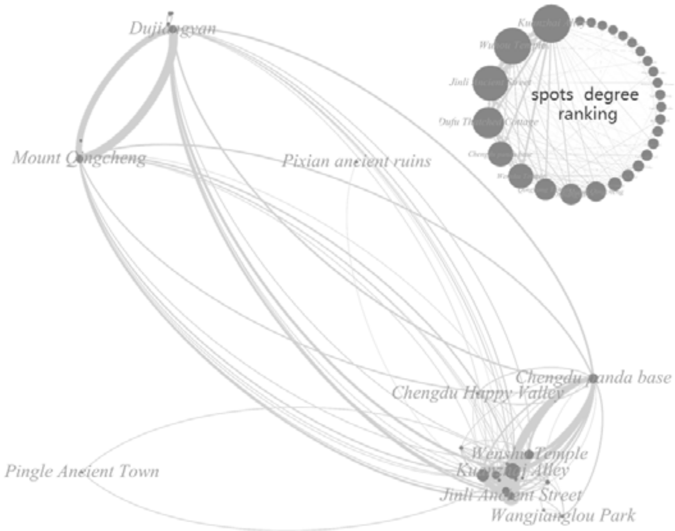


Figure 7. The frequented travel spots in Chengdu

We conducted sequential pattern mining experiment with rs_{min} = 0.2% to extract all the compact sequences with two spots. Finally, 28 spots (nodes with geotag) and 205 correlations (edges with weight) were generated and the results were shown in Figure 7. According to the spots degree ranking, *Kuanzhai Alley* was the most popular spot in Chengdu. Also we can easily learn that Chengdu tourism mainly concentrated in three areas. The first area is downtown of Chengdu including many spots (e.g. *Kuanzhai Alley*, *Wenshu Temple*, *Jinli Ancient*

Street), the second area is *Chengdu panda base* and the third area includes two spots *Dujiangyan* and *Mount Qingcheng*.

4.4 Spot associated service mining

In this section, *Kuanzhai Alley* was chosen as the focus spot and 5030 travel routes include the spot *Kuanzhai Alley* were extracted as the data for the fellowing experiment.

We conducted a compact patterns mining experiment with rs_{min} = 0.5% to extract the compact 2 items sets. Only the patterns including *Kuanzhai Alley* (e.g. {*Kuanzhai Alley*, *Liaolaoma Trotter*}) were remained in the final dataset. 24 locations (nodes with geotag) and 23 correlations (edges with weight) are generated and the results were shown in Figure 8. Here we found 23 popular services (food, shopping and entertainment) associated to *Kuanzhai Alley* and most of the services are in type of food.

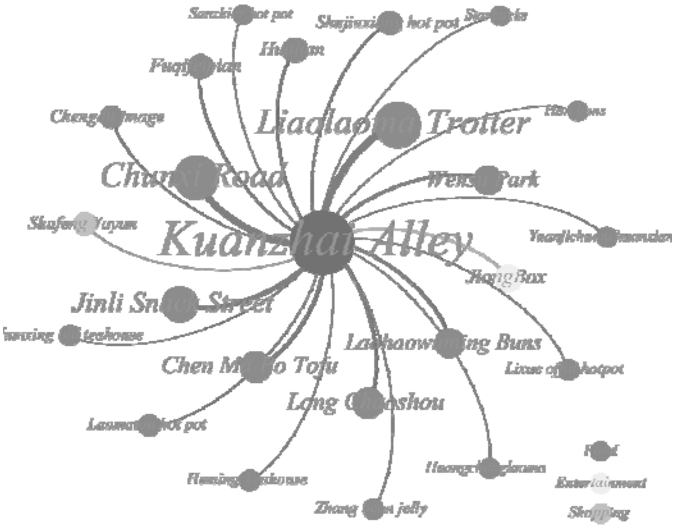


Figure 8. ***Kuanzhai Alley*** **associated service**

5. CONCLUSIONS

The rapid growth of Web 2.0 and social media content in online travel have drawn great amount of attention both in academia and in industry. Research on automatically discover tourism information from UGC have both academic and practical significance.

In this paper, using the structured tourism blogs data generated by prior tourists, we mine the frequent spots and popular travel service within a given travel destination. These information can help us understand the travel destination and enable the website to recommend interesting travel spots. In this work, we regard the set of textual locations post on each blog as a sequence of travel route. First, we introduced the method for compact pattern mining and sequential pattern mining and then constructed a framework to analyze the structured tourism blogs. Specifically, a crawler was developed to automatically download the structured tourism blogs. Besides, a program based on Google Map API was written to obtain the geotag of each location. Then sequential pattern mining was conducted to discover the frequented spots and their correlations. Later, based on the compact pattern mining method, we detected the spot associated travel service like shopping, etc. Finally, we illustrate the benefits of this framework by applying it to an online tourism blog dataset (in Chinese). As a result, our proposed framework showed clear advantages by providing a well presented information for potential tourists to understand the travel destination quickly.

In the future, we would like to improve the framework by incorporating the photos and detailed reviews about travel destinations. Besides, the spatial relationships of locations could be taken into account for generating a better travel routes.

ACKNOWLEDGEMENT

This research is partially supported by NKBRPC (2014CB340506) and the NSFC(71071042).

REFERENCES

[1] Shen, J., Cheng, Z., Shen, J., Mei, T., & Gao, X. (2014). The Evolution of Research on Multimedia Travel Guide Search and Recommender Systems. In MultiMedia Modeling (pp. 227-238). Springer International Publishing.

[2] Sharda, N., & Ponnada, M. (2008). Tourism blog visualizer for better tour planning. Journal of Vacation Marketing, 14(2), 157-167.

[3] Davidov, D., & Rappoport, A. (2009). Geo-mining: discovery of road and transport networks using directional patterns. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1 (pp. 267-275). Association for Computational Linguistics.

[4] Pang, Y., Hao, Q., Yuan, Y., Hu, T., Cai, R., & Zhang, L. (2011). Summarizing tourist destinations by mining user-generated travelogues and photos. Computer Vision and Image Understanding, 115(3), 352-363.

[5] Yuan, H., Guo, L., Xu, H., & Xiang, Y. (2013). Frequent Patterns Based Word Network: What Can We Obtain from the Tourism Blogs?. In Knowledge Science, Engineering and Management (pp. 15-26). Springer Berlin Heidelberg.

[6] Ye, Q., Law, R., Li, S., & Li, Y. (2011). Feature extraction of travel destinations from online Chinese-language customer reviews. International Journal of Services Technology and Management, 15(1), 106-118.

[7] Zheng, Y., Zhang, L., Xie, X., & Ma, W. Y. (2009). Mining interesting locations and travel sequences from GPS trajectories. In Proceedings of the 18th international conference on World wide web (pp. 791-800). ACM.

[8] Fujisaka, T., Lee, R., & Sumiya, K. (2010,). Discovery of user behavior patterns from geo-tagged micro-blogs. In Proceedings of the 4th International Conference on Uniquitous Information Management and Communication (p. 36). ACM.

[9] Xu, Z., Chen, L., (2014). Mining geo-tagged photos. Communications of the CCF, 10(5), 31-36. (in Chinese)

[10] Lu, X., Wang, C., Yang, J. M., Pang, Y., & Zhang, L. (2010, October). Photo2trip: generating travel routes from geo-tagged photos for trip planning. In Proceedings of the international conference on Multimedia (pp. 143-152). ACM.

[11] Arase, Y., Xie, X., Hara, T., & Nishio, S. (2010). Mining people's trips from large scale geo-tagged photos. In Proceedings of the international conference on Multimedia (pp. 133-142). ACM.

[12] Zeng, Z., Zhang, R., Liu, X., Guo, X., & Sun, H. (2012). Generating tourism path from trajectories and geo-photos. In Web Information Systems Engineering-WISE 2012 (pp. 199-212). Springer Berlin Heidelberg.

[13] Kurashima, T., Iwata, T., Irie, G., & Fujimura, K. (2013). Travel route recommendation using geotagged photos. Knowledge and information systems, 37(1), 37-60.

[14] Borgelt, C. (2012). Frequent item set mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6), 437-456.

[15] Liu B. (2011). Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Data-Centric Systems and Applications.