An Analysis System Development of User Point of Interests based on Social Big Data

*Bo-Hyun Yun
Computer Education Department, Mokwon University
ybh@mokwon.ac.kr

Abstract

A number of studies have been recently conducted on user location information due to the wide spread of smart devices. The present study analyzed the major point of interests (POI) according to gender, age, region and time by extracting user profiles using mass documents and services produced from Twitter where location services are offered, and confirmed the POI correlation according to those profiles. The results of the POI flow analysis confirmed monthly hot places, daily hot places, hourly hot places, and POI preference according to gender and age.

Keywords: *Point of Interests(POI), Tweet analysis, Location information, Big data, smart devices*

1. Introduction

While the studies using the terminal device and GPS has been conducted, the studies on the POI analysis in connection with SNS service attached with GPS information have only a short history. Location-based services are used to confirm and track the user or terminal location by receiving the GPS signals which GPS (Global Positioning System)-attached mobile devices such as smart phones and tablet PC have transmitted by satellite. Social Network Services (SNS) recently began to suggest some nice restaurants where users recommend as well as locations where users send messages through the GPS information.

However, the existing studies recommend the POI and offer services simply considering locations. To offer more exact services, it is necessary to consider the characteristics of users. In other words, the existing studies did not consider users' gender and age. The present study analyzed the real data and verified the importance of user profile information when providing location-based services.

This study is composed of as follows: The POI analysis and related researches were explained. Chapter 3 explains the real-time POI analysis system and Chapter 4 confirms the POI analysis results through practical examples. The conclusion is presented in Chapter 5.

2. Related Researches

A method [1] to recommend friends in SNS considering the similarity between the attribute data of POI types and user trajectory is suggested. The research [2] offers more detailed friend suggestion than did the simple clustering of users who have similar trajectory. However it also excludes the important aspects for friend suggestion, which are age and gender, causing the high probability of not selecting the recommended friends. The research [3] offers users services by using the location tag information attached to pictures. It gives recommendation to users by measuring the similarity between user trajectory and POI trajectory. However, it does not consider age and gender when giving the recommendation to users. The research [4] measures three factors to recommend the POI to users. First, the user-based collaborative filtering is applied for the measurement of the similarity between users. Second, the social influence from friends is measured. Third, regional influence is measured by using the power law distribution. This research recommends users regions by combining these three factors. Unlike the existing POI recommendation researches, the research [5] devised a method to

* Received: Mar. 21, 2015, Revised Apr.01, Accepted: Jul.08, 2015

continuously recommend the POI for tomorrow and following days by using user transfer characteristics and regional information. It gives the continuous POI recommendation by modeling the transfer patterns and regional characteristics as the Markov chain for a specific individual. Unlike the research above, the researches on POI in indoor environments [6, 7] have been actively conducted, which gives users POI recommendation based on indoor locations such as department store and large-scaled commercial districts. It uses a method to offer users services by analyzing the their visit history of the relevant indoor environments and the spots where they stayed long.

Even though the existing researches using regional information have been actively conducted, researches using user profile information (gender, age, etc.) have yet to be performed. The suggesting research shares a common aspect that the POI is recommended or analyzed using location information. However, user profile information is needed to provide the exact POI. Since the existing research assumes that the user trajectory (movement) itself has a meaning, whether it is a place that users actually prefer cannot be known. However, compared to the existing researches, it has a distinction in that the suggesting research considers whether actual users directly mentions a place, by using the text information of tweets, and reflects the real time aspects of SNS services.

3. Real-time POI Analysis

This chapter explains the POI analysis system. It roughly has four steps and the entire system is presented in Figure 1. Data preprocessing, establishing a POI dictionary, a data processing method for each step are explained in the following chapter.

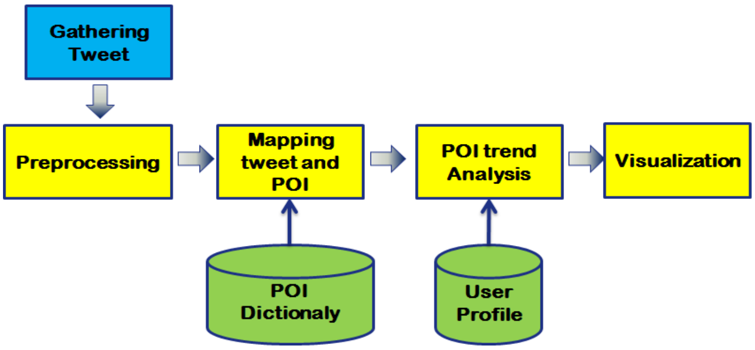


Figure 1. System diagram

3.1. Collection of tweet documents

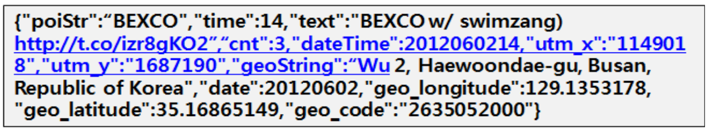
Tweet documents use the original texts of tweets collected from ETRI. The documents and user information available to the POI analysis are extracted from 1,188,457,555 tweet documents produced from January 2012 to February 2013, and the POI flow is analyzed according to gender, age, region and period. In the following steps, the necessary information for the POI analysis is extracted through the preprocessing and the only documents tagged with the GPS information are classified.

3.2. Processing of tweet documents

Subjects of the POI analysis are documents with the GPS information among all collected tweet documents. The subjects of analysis are limited to the tweets in which the POI includes the locations tweeters mentioned.

3.2.1. Filtering of tweets tagged with GPS information

If users write a tweet without using a GPS function of terminal devices, the GPS information is not included, and thus, the exact location of users cannot be confirmed. Additionally, it is necessary to select the tweets attached with the GPS information since it is difficult to conduct the POI analysis in some regions due to the different SNS user distributions according to region. Figure 2 is one of the GPS information-attached information.



```
{"poiStr":“BEXCO","time":14,"text":"BEXCO w/ swimzang)
http://t.co/izr8gKO2”,“cnt":3,"dateTime":2012060214,"utm_x":"114901
8","utm_y":"1687190","geoString":“Wu 2, Haewoondae-gu, Busan,
Republic of Korea","date":20120602,"geo_longitude":129.1353178,
"geo_latitude":35.16865149,"geo_code":"2635052000"}
```

Figure 2. Example of tweet attached with GPS information

3.2.2. Preprocessing of tweet documents

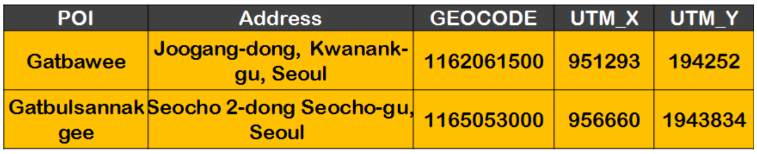
The collected tweet documents basically include various meta data and sometimes different meta data according to the tweet types and user settings. Since unnecessary meta data for the POI analysis is included, only the essential tweet text, date and time information, altitude, longitude, and address code are used after excluding a number of unnecessary data.

3.3. POI mapping of tweet documents

Mapping between tweet documents and POI requires a POI dictionary. This section explains the explanation of a POI dictionary and problems and solutions about the mapping.

3.3.1. POI dictionary

POI information consists of 49,246 Korean major POIs provided from ETRI and an example is presented in Figure 3.



POI	Address	GEOCODE	UTM_X	UTM_Y
Gatbawee	Joogang-dong, Kwanank-gu, Seoul	1162061500	951293	194252
Gatbulsannak gee	Seocho 2-dong Seocho-gu, Seoul	1165053000	956660	1943834

Figure 3. Example of POI dictionary

A POI dictionary consists of POI, address, GEOCODE, UTM_X, and UTM_Y. Since tweet documents include altitude and longitude as location information, these should be converted into UTM_X and UTM_Y formats and then mapped with the POI information.

3.3.2. POI and tweet mapping

When mapping tweet documents with POI, the two aspects should be considered: mapping with the POI that tweeters are located and mapping with the POI that is mentioned in tweets. The first method considers only tagged GPS information. not the texts. Since tweeters can mention unrelated places and write a post wherever they are, it is possible to have wrong POI. Additionally, mapping through the conversion of the tagged GPS information into UTM_X and UTM_Y as shown in Figure 3 does not result in the conversion into the addresses of specific places and includes surrounding places. Therefore, a problem of mapping a tweet with many POI information can occur. A problem in Figure 3 does not occur when mapping POIs mentioned in tweets, but it can be difficult to distinguish which region POIs are related because of the possibility of having same regional names. The present study mapped tweets with the POI information considering both GPS information and tweet texts. If different regions had a same POI, the POI regions were limited using the GPS information. If many POIs were mapped with the GPS information, only one POI was selected using the POI mentioned in the tweet texts.

3.4. Establishing user profiles

In the previous section through the filtering work of data set and the POI dictionary, tweet documents attached with the POIs were prepared. This section explains the establishment of user profile database necessary for the analysis of the POI flow based on the user profiles, and the analysis details of the POI flows. User profiles establish the database by inference of users' gender and age from meta data, "Description", in tweet documents. Description is a writing that users introduce somethings. Some users' gender and age can be inferred while some others cannot. In some cases, both gender and age can be inferred while in some others, only one of them can be inferred. We extracted the profiles of users in Seoul, Gyeonggi and Busan with the high user distribution and the extraction results are presented in Table 1.

Table 1. User profile database

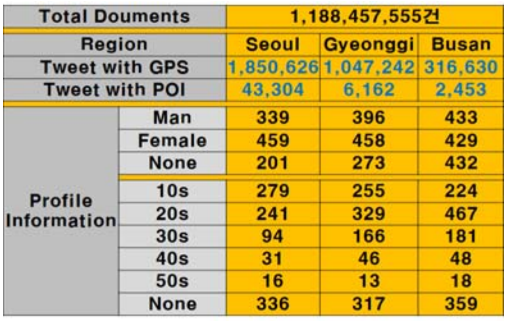
Region	*User*	*User profile*
Seoul	39,500	1,294
Kounggi	12,869	1,127
Busan	8,000	999

The following chapter analyzes seasonal, daily, and hourly hot places based on the tweets mapped with the POIs. The preference is also analyzed through the POI analysis according to age and gender, based on the existing user profile database.

4. POI Analysis

4.1. Analysis data

The results of document preprocessing and tagging the POI information with tweet documents and the extraction results of user profiles as mentioned in Chapter 4 are presented in Figure 4.



Total Douments		1,188,457,555건		
Region		Seoul	Gyeonggi	Busan
Tweet with GPS		1,850,626	1,047,242	316,630
Tweet with POI		43,304	6,162	2,453
Profile Information	Man	339	396	433
	Female	459	458	429
	None	201	273	432
	10s	279	255	224
	20s	241	329	467
	30s	94	166	181
	40s	31	46	48
	50s	16	13	18
	None	336	317	359

Figure 4. Analysis result of data

Although there are a number of documents collected for 14 months, there are relatively a few number of documents mapped with the POI information and including the GPS information. The rate of the user profile information is higher in women in terms of gender, while this rate is the highest in the 20s in terms of age.

4.2. Seasonal hot places

An analysis of seasonal hot places in each region is an yearly analysis of the POI flow for two preferred regions in each region, without considering user profiles. A representative seasonal POI is a place with the highest frequency in the relevant month.

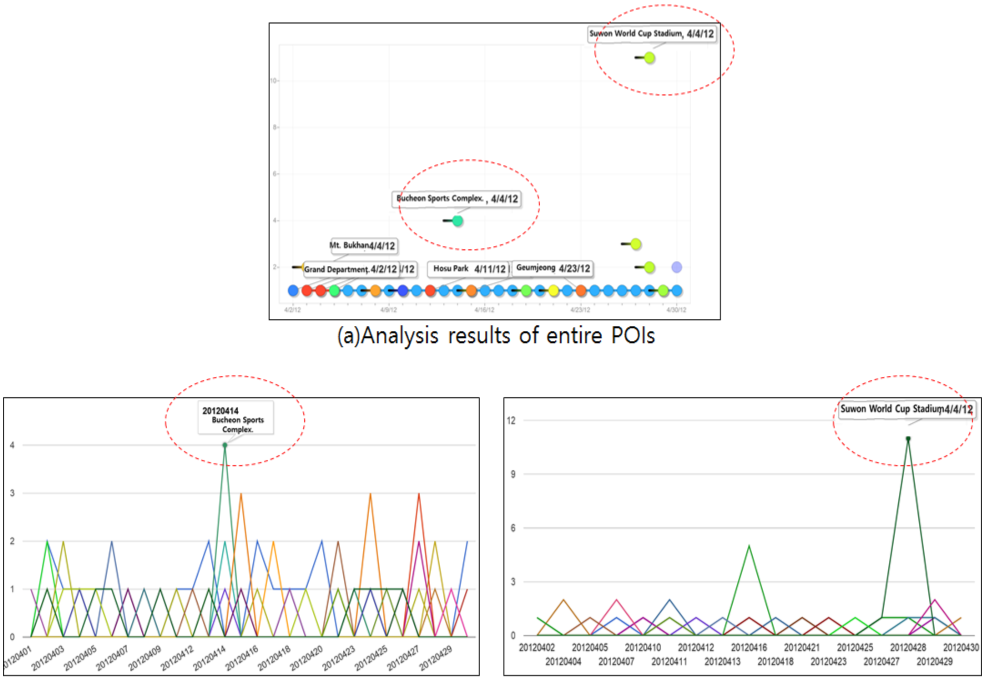
Seoul was not affected from seasons and its hot places were mostly located in Hongdae. Gyeonggi-do showed an opposite POI analysis result with Seoul. In April 2012 when sports games opened, Suwon World Cup Stadium and Bucheon Sports Complex were analyzed. The seasonal analysis showed that people visited Everland from May to June 2012, an exhibition venue, KINTEX in October 2012, and Imjingak in January in 2012. The analysis for Busan was similar. People visited Sajik Baseball Stadium from May to April when the professional baseball season started and Haeundae during the summer season from July to August. Seasonal popular areas were analyzed through the analysis of the monthly hot places. The following chapter explains which users visited the locations and the reason why they visited them.

4.3. Daily hot places

We analyzed which user (gender, age) visited the relevant POIs, by selecting time periods in several regions based on the analysis results of the seasonal POIs. Figure 5 shows an analysis result of daily POIs obtained by distinguishing genders from user profiles, and major hot places in Gyeonggi in April 2013. Figure 5(a) shows an analysis result of the entire POIs without applying user profiles, confirming the high frequencies in Suwon World Cup Stadium and Bucheon Sports Complex, which are the places where professional soccer games were held. The analysis through the application of user profiles to these results is presented in Figure 5(b) and (c). Figure 5(b) is a result of the POI analysis focusing on male users, while Figure 5(c) is a result of the POI analysis focusing on female users. It was confirmed that the male users visited Bucheon Sports Complex more than female users and female users visited Suwon World Cup Stadium more than male users. Through these analysis results, it can be inferred that Bucheon Sports Complex has many male fans and Suwon World Cup Stadium has many female fans.

Another example can be confirmed in Figure 6. Figure 6 is a POI analysis result for Gyeonggi-do in October 2012. Through Figure 6(a), it can be confirmed that KINTEX is the most visited place in October.

Through Figure 6(b) and (c) which are the POI analysis results through the application of user genders, it can be confirmed to be visited by both male and female users.

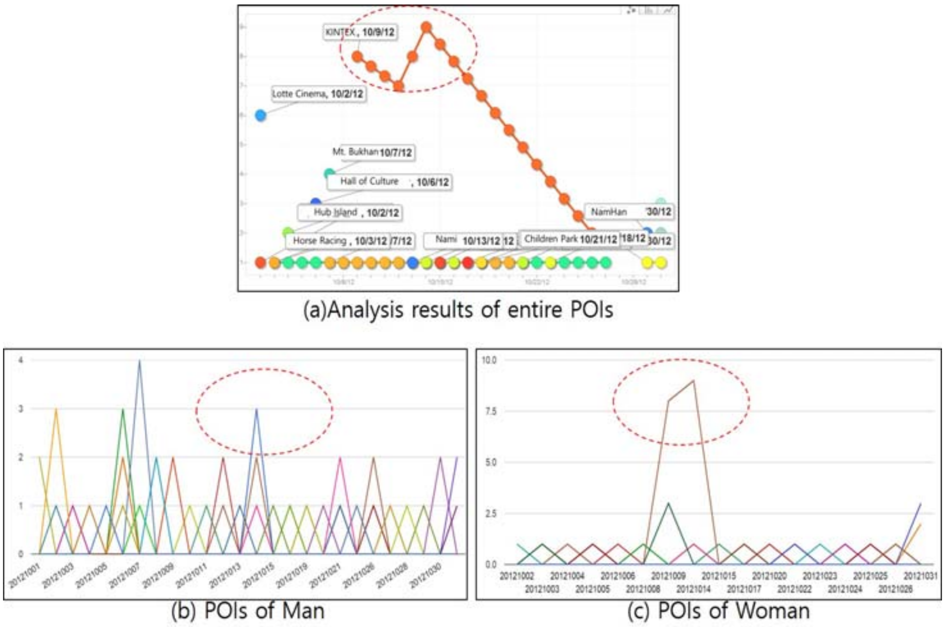


(a)Analysis results of entire POIs

(b) POIs of Man

(c) POIs of Woman

Figure 5. Daily POI analysis result for Gyeonggi-do in April 2012



(a)Analysis results of entire POIs

(b) POIs of Man

(c) POIs of Woman

Figure 6. Daily POI analysis results for Gyeonggi-do in October 2012

4.4. Hourly hot places

Since the seasonal and daily POI analyses are related to the most popular POI according to the relevant month and day, it is impossible to confirm at what time users actually visited what place. This section analyzes the time when users visited a certain place on the basis of specific dates. For the hourly POI analysis, a analysis was conducted for Gyeonggi-do and Busan based on December 31, 2012, and the analysis results are presented in Figure 7 and 8.

When analyzing Gyeonggi-do in Figure 7, it can be confirmed that users in their 10s tended to greet the New Year at Imjingak, and users in their 30s at the Korean Folk Village. Also, the analysis for Busan in Figure 8 shows that users in their 10s preferred to greet the New Year at Seomyeon and Haewoondae in Busan, and users in their 30s preferred to greet the New Year at Igidae Cliff.

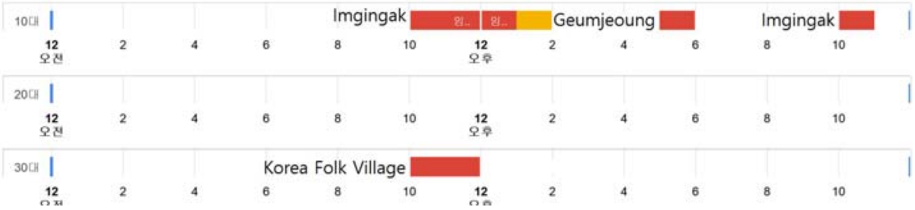


Figure 7. Hourly POI flow analysis for Gyeonggi-do on December 31, 2012

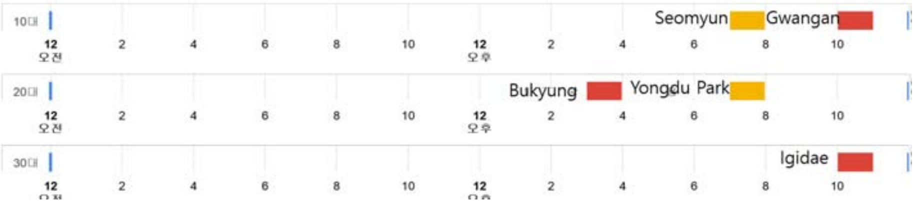


Figure 8. Hourly POI flow analysis for Busan on December 31, 2012

5. Conclusion

The present study suggested an analysis system for the POI flow by using user profiles from the representative big data, Tweeter, and analyzed the POI flows based on the actual SNS data. Even though studies on providing location-based services to users have been actively conducted, these studies have some flaws to provide incorrect information because only location is considered. To resolve this problem, the present study suggested a system providing services according to gender and age by using user profile information as well as location information.

Through the analysis results of the POI flow, monthly, daily, and hourly hot places and the POI preferences according to gender and age were confirmed. Future studies on a method to automatically extract user profiles are planned to be conducted.

6. References

[1] Chunghee Lee, et al. “Friend Suggestion through User Preference and Transfer Patterns for Mobile Social Networks”, *Journal of the Korea Society of Computer and Information: Database* Vol. 40. No. 1, pp. 79-87, Feb. 2013.

[2] X. Xiao, Y. Zheng, Q. Luo, and X. Xie, “Finding Similar Users Using Category-Based Location History”, Proc. ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, NY, USA, 2010, pp. 442-445.

[3] Jinpeng C., Zhenyu W, Hongbo G., Changjie Zhang, Xuejun C., and Deyi L., “Recommending Interesting Landmarks Based on Geo-tags from Photo Sharing Sites”, Proc. 14th International Conference, 2013, pp. 151-159.

[4] Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee. “Exploiting Geographical Influence for Collaborative Point-of-interest Recommendation”, Proc. SIGIR'11 - Research and Development in Information Retrieval, NY, USA, pp. 325-334, 2011.

[5] Chen C, Haiqin Y, Michael R. L, and Irwin K, “Where you like to go next: successive point-of-interest recommendation”, Proc. IJCAI'13 Artificial Intelligence, 2013, pp. 2605-2611.

[6] Pombinho, P., Afonso, A. P., and Carmo, M. B., “Point of Interest Awareness Using Indoor Positioning with a Mobile Phone”, Proc. of the PECCS 2011, pp. 5-14.

[7] Beomsu Kim, et al. “Discovery Techniques for Users’ Point of Interest and Realization of a POI Recommendation System, for indoor location-based services”, *Journal of the Korea Society of Computer and Information*, Vol. 17, No. 5, pp. 81-91, 2012.