Indian Journal of Science and Technology, *Vol 9(35), DOI: 10.17485/ijst/2016/v9i35/101763, September 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Location Disambiguation for POI Detection using Lingustic Clues in Social Media

Hyo-Jung Oh[1], Yong Kim[2] and Bo-Hyun Yun[3*]

[1]Graduate School of Archives and Records Management, Institute of Culture Convergence Archiving, Chonbuk National University, Korea; ohj@jbnu.ac.kr
[2]Department of Library and Information Science, Institute of Culture Convergence Archiving, Chonbuk National University, Korea; yk9118@jbnu.ac.kr
[3]Department of Computer Education, Mokwon University, Daejeon, Korea; ybh@mokwon.ac.kr

Abstract

Objectives: With the rise of locative media fostered by the growing ubiquity of mobile devises, many researches to detect users' preferred location information have been actively studied. **Methods/Statistical Analysis:** This paper proposes a method for hot place detection based on social media contents. Especially, we focus on POIs (Point-Of-Interests) disambiguation using linguistic clues based on social media content analysis. We try to combine implicit clues using linguistic analysis and geometric metadata which are embedded in tweet mentions. This feature can help to overcome the limitation of using explicit information only. **Findings:** By experiment results based on real tweet data, we show the effects and usage of our proposed method. **Improvements/Applications:** The POI method can also be enhanced by considering other efficiency metrics.

Keywords: Linguistic Clues, Location Disambiguation, POI, Social Media

1. Introduction

As mobile devices such as smartphone and tablet PC have become the latest range, social media is a growing trend explosively. On the other hand, as IoT (Internet of Things) technologies has been evolved, various practical researches focused on using Global Position System (GPS) information gathered from the wireless device have been tried[1]. In advance, more commercial application toward Location-based Service (LBS) has been developed[2].

Under the current ICT environment, social media has two attractive features: First, social media is the most common media to capture user's interests in the sense that generated by the users voluntarily. This nature in terms of 'spontaneity' is distinguishable feature from enforced motivation by traditional marketing or promotions. Second, social media usually have several implicit user's contexts including location and time information. Users' behaviors such as what they did or where they went melt into these information.

For commercial services such LBS or advertisement, the most valuable information is to know where user are interested and when they usually visit. In this paper, we define this information as "hot place". For adoption of a definition, we referred the definition of "hot spot", which is a physical location that offers Internet access over a Wireless Local Area Network (WLAN) through the use of a router connected to a link to an Internet service provider[3]. As similar with the role of hotspot is connecting physically, we try to detect hot place to connect user network semantically based on user's interest. This paper proposes a method for hot place detection based on social media contents. Especially, we focus on POIs (Point-Of-Interests) disambiguation using linguistic clues based on social media content analysis.

*Author for correspondence

2. Related Works

The research focusing on LBS is vast and a number of these services have been implemented and tested[4]. Tourist information systems are ideal examples for such applications in early days[5]. Location-tracking service is the second occasion of LBS application. The Location-tracking service system for the children or the elderly has been developed for safety purpose[6]. Currently wireless sensor networks have been given more attention in academia and industry, and have been taken as one of the most important technologies in 21st century[7,8]. Most of mobile phones have been integrated with a wireless communication module, a voice/recorder module, a camera module, as much as a GPS module. These modules are widely used sensors. Some scholars have proposed applications of mobile phones, such as the city management and monitoring[9], and the emergent communications in mines[10].

The research of collective behavior has attracted a lot of attention in recent years, which can empower various applications, such as recommendation systems and intelligent transportation systems. However, in traditional social science, it is practically difficult to collect large-scale user behavior data[11]. With the popularity of social networks, users leave a large volume of digital footprints online. For example, by analyzing repost behavior in social networks, In[12] studied predictability of the content dissemination trends. However, in traditional social networks, users′ behavior, such as posting blogs, sharing photos and uploading videos, does not necessarily reflect their daily activities. The analysis of collective behavior in LBSNs can also enable various applications. For example, by analyzing users′ check-in data in LBSNs, In[13] studied the personalized location based services such as POI recommendation.

Unfortunately many researches focused on POI or LBS are depends on explicitly information such as GPS code or other physical sensing results. To resolve this limitation, we analyze social media content using linguistic analysis and we refer to semantic meaning for disambiguation of location information in which users are interested.

3. Social Media Content Analysis

3.1 System Overview

Figure 1 illustrates overview of our system. Our ultimate goal is to propose a novel method of visualization for preferred location and moving patterns according to user groups based on contents analysis in social bigdata[14]. We collected tweet media, as known as a representative of social media, during one year and conducted preliminary investigation. We also distinguished user groups according to gender, ages, local area, and time and analyzed their preferred locations and moving pattern. Especially, this paper focuses on location disambiguation problem solving.

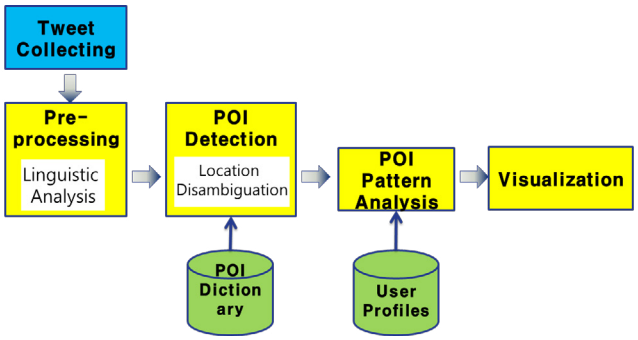

Figure 1. System Overview.

3.2 Social Media Contents

Traditional researches of location information detection are based on explicit information such as GPS code or RFID tags. Compared with this, we attempted to use implicit information in social media such as mention or metadata written in text. For target media, we chose Twitter© and collected real tweets data using open API[15]. In Twitter, many metadata are embedded such as user's id, publishing time, and tags. Figure 2 show an example of tweet in Korean.

```
{
        “datetime”:2012060214,
        "text”: “모터쇼 구경함 @롯데 http://t.co/
          izr8gKO2",
        “utm_x”:1149018, “utm_y”:1687190,
        "geoString":"대한민국 부산광역시 해운대구
          우2동",
        "geo_longitude":129.1353178,
        "geo_latitude":35.16865149,
        "geo_code":2635052000, ….
}
```

Figure 2. Example of a tweet in Korean.

As shown in that Figure, we can notice that the user was in "대한민국 부산광역시 해운대구 우2동 (U2-dong, Haeundae-gu, Busan, Korea)" and he attended an event on " 모터쇼 (motor show)" at 2 p.m. June 2nd, 2012.

3.3 Content Analysis

For preprocessing to detect hot places, we analyzed text contents in tweet mention and metadata. To understand users behaviors in tweets in Figure 1, we performed natural language processing (NLP) for "text" and "geo-String" tages and try to find related POI[16] information. We used NLP component and POI dictionary developed by ETRI (Electronics and Telecommunications Research Institute)[17]. The NLP component consists of morphological analysis, named entity recognition, and parsing. The ETRI POI dictionary consist of 49.246 entries with addresses, geo codes and UTM-K codes. Table 1 shows some examples of the POI dictionary.

Table 2 show results of tweet content analysis in Figure 1. We extracted that the user's action is 'sightseeing' and the specific event was 'motor show' from the tweet mention, "모터쇼 구경함 (sightseeing motor show)" by linguistic analysis. Moreover, we noticed that the address of user is "U-dong, Haeundae-gu, Busan, Korea" However, we don't know yet the exact information about 'Lotte' because there many POIs as known as 'Lotte' such as 'Lotte department store', 'Lotte mart', and 'Lotte world'[17]. We'll discuss this problem in the following section.

4. Location Disambiguation for POI Detection

There are two directions of POI detection method in tweets: 1) using GPS information automatically attached in tweets 2) using user mention POI names in text.

The first direction is more traditional way to refer current location when the user wrote the tweet. However, in case of that the tweet content is not related with current location or the user wrote the tweet while he were moving, the location information belongs to any particular POIs. The second method has also weakness in terms of disambiguation. As announced in Section 3.2, the POI mentioned by the user in Figure 1, 'Lotte', indicates several locations. Table 3 shows some examples of POI entries sharing same name 'Lotte department store'. Even in Busan, there are four branches of 'Lotte department store'. Thus we have to combine text information as well as geometric information to distinguish clearly which POIs are the user indicates. Even though the tweet in Figure 1 shows explicit GPS codes (geo_longitude = 129.1353178, geo_latitude =35.16865149), there are many POIs and landmarks in that area. Table 4 depicts various POIs while having same GPS code area in Busan. Among various POIs in Table 3, we have to clarify the specific location related with the tweet mention in Figure 1. Based on the analysis result in Table 1, we can figure out that the user attended ('sightseeing') a motor show in 'Lotte'. It means that the user was in 'Lotte department store Centum City branch'. We excluded 'Lotte Cinema' because people see a movie in theaters, not exhibition.

Table 2. Tweet Content Analysis Results

Tweet contents	Text	Morpheme / Meaning
Text	Motor show	Noun / Event
	Sightseeing	Verb / Action
	Lotte	Noun
GeoString	Republic of Korea	Noun/ Country
	Busan	Noun/ City
	Haeundae-gu	Noun/ District
	U2-dong	Noun/ Village

As described before, we dealt with metadata embedded in tweet directly, meanwhile, some tweets does not provide GPS code in real situation. For these cases, we extracted location information which the user explicitly mentioned in main messages. For example, the mention of "부산 서면 웨어하우스 마시썽!! (WAREHOUSE in Busan Seomen Delicious!!) Indicates the hot place

Table 1. Examples of the POI dictionary

POI	Address	GEO CODE	UTM_X	UTM_Y
(NamDaeMoon)	(Namdaumoon-ro 4 Gong-gu, Seoul)	1114011700	953662	1951316
(COEX)	(Samsung-dong, Kangnam-gu, Seoul)	1168010500	961076	1945855
(Lotte World)	(Jamsil3-dong, Songpa-gu, Seoul)	1171068000	964463	1945845
(Jungbang Fall)	(Jungbang-dong Seoguipo-si, Jeju)	5013052000	913506	1473008

Table 3. POIs sharing same name, 'Lotte department store'

POI	Address	GEO CODE	UTM_X	UTM_Y
Lotte Department Store Gangnam branch	(Daechi 1-dong, Gangnam-gu, Seoul)	960509	960509	1944280
Lotte Department Store Jeonju branch	(Seosin-dong, Wansan-gu, Jeonju, Jeolabuk=-do)	4511112900	965852	1759834
Lotte Department Store Dalseo branch	(Sangin 1-dong, Dalseo-gu, Daegu)	2729062400	1093844	1758532
Lotte Department Store Centum City branch	(U 2-dong, Haeundae-gu, Busan l)	2635052000	1148524	1687277

is a restaurant and it is located in "Jeonpodaero 209, Busanjin-gu, Busan, Korea"

Table 4. POIs sharing same GPS code

Category of business	POIs
Coffee/Bar	(12Bar)
	(GamSalon)
	(Caffebene)
	(Hollys Coffee) …
Restaurant / cafeteria	405kitchen
	(NamicoRamen)
	CafeMichaya …
Exhibition / Theater	BEXCO
	Lotte Cinema …
Shopping	(Lotte Department Centum City)
	(Centum City)
	(homeplus)…
Transportation	(subway line No. 2)
	Subway Line 2 Station Museum of Art
	…

5. Experimental Results

To evaluate our proposed method, we collected tweets from 2012 through March 2013. The total number of tweets 1,002,240,097, while only 0.41% (4,107,420) tweets among them have GPS information. Besides, over 26% (250,560,024) tweets included location information with text. This aspect supports our claim that only using explicit information has limitations.

On the other hands, among 49,246 entries in our POI dictionary, almost 30% (15,904) entries are ambiguity. This result endorses the importance of our proposed method. To measure the precision our location disambiguation method, we selected tweets which were generated in Busan area because we have to manually examine correct answer and our system results. The test collection consists of 258,952 tweets and 24,530 tweets have ambiguous location information. Table 5 shows the specification of our test collection.

Table 5. Test Collection

Specification	Number	Comments
Total of collected tweets	1,002,240,097	tweets from 2012 through March 2013
Initial collection	4,107,420	tweets including GPS information
Test Collection	258,952	tweets in Busan
Experiment Target tweets	24,530	tweets including ambiguous location
Ambiguous POIs in our test collection	19,243	Multiple POIs

Table 6. Experiment Results

Specification		Number
Ambiguous POIs in our test collection		19,243
Disambiguated POIs		15,424
Precision		0.805
Failed POIs		3,819
	POI dictionary errors	2,413
	Mismatch GPS/Geo codes	802
	Linguistic analysis errors	604

In our test collection, users mentioned 24,530 locations information, while 19,243 POIs have ambiguous. Using our method, we distinguished 15,424 exact POIs

and we finally got 0.805 of precision as shown in Table 6. Based on failure analysis of 3,819 POIs in Table 6, 2,413 POIs are not appeared in our POI dictionary and 802 POIs are mismatched between GPS codes in tweets and our POI dictionary Geo codes. The others come from our linguistic analysis errors.

6. Conclusions

With the rise of locative media fostered by the growing ubiquity of smartphones, the ways in which place, intimacy and location are visualized is changing. In particular, hot place detection is useful in many ways for commercial services. For instance, advertiser can decide target marketing area based on hot places. This paper proposed the location disambiguation method based on social media content analysis. We also combined implicit clues using linguistic analysis and geometric metadata which are embedded in tweet mentions. This feature can help to overcome the limitation of using explicit information only.

Our future works will be extended to using other sophisticated metadata. There are various applications to recommend popular restaurant or tourist destination using LBS, whereas, most of them can not reflect the users' characteristics such as age or gender. To overcome the lack that previous works only depend on the users' current position, we will device a novel method to analyze the correlation of POI based on the user profile constructed by using the content and metadata in SNS data.

7. References

1. Jinpeng C, Zhenyu W, Hongbo G, Zhang C, Xuejun C, Deyi L. Recommending Interesting Landmarks Based on Geotags from Photo Sharing Sites. Lecture Notes in Computer Science. 2013 Oct; 8181:151 – 9.
2. Hjorth L. The place of the emplaced mobile: A case study into gendered locative media practices. Mobile Media & Communication. 2013 Jan; 1(1):110–5.
3. Hotspot (Wi-Fi).Available from: https://en.wikipedia.org/wiki/Hotspot-(Wi-Fi). Date Accessed: 28/07/2016.
4. Zhu X, Zhou C. POI Inquiries and data update based on LBS. Proceedings of International Symposium on Information Engineering and Electronic Commerce. 2009; 730–4.
5. Roth J. Context-aware Web Applications Using the Pin Point Infrastructure. Proceedings of International Conference WWW/Internet. 2002 Nov, pp. 1–8.
6. Marmasse NC Schmandt, Safe & sound - a wireless leash. Proceeding CHI '03 Extended Abstracts on Human Factors in Computing System. 2003; 726–7
7. Warneke B. Smart Dust: Communicating with a Cubic Millimeter Computer. IEEE Computer. 2001 Jan; 34 (1):44–51.
8. Li Y. Sensor Network Measurement Technologies Based on Endto-end Measurement. Doctoral Thesis of Northwest Industrial University. 2007.
9. Li D. The Construction and Application of Wuhan Urban Grid Management and Service System. Bulletin of Surveying and Mapping. 2007.
10. Zhang Y. A mine emergency communication system based on wireless sensor networks. Industry and Automation. 2008; 4: 71–3.
11. Yang D, Zhang D, Chen L, Qu B. NationTelescope: Monitoring and visualizing large-scale collective behavior in LBSNs. Journal of Network and Computer Applications. 2015 Sep; 55:170–80.
12. Lu X, Yu Z, Guo B, Zhou X. Predicting the content dissemination trends by repost behavior modeling in mobile social networks. Journal of Network and Computer Applications. 2014 Jun; 42:197–207.
13. Yang D, Zhang D, Yu Z, Yu Z. Fine-grained preference-aware location search leveraging crowdsourced digital footprints from lbsns. Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2013 Sep. p. 479–88.
14. Oh HJ, Yun BH, Chio NH, Yoo CJ, Kim Y. Visualization for Preferred Locations and Moving Patterns According to User Groups based on Contents Analysis in Social Big Data. Journal of Kalinga Institute of Industrial Technology. 2014 Dec; 12(12):195–203.
15. The Streaming APIs Overview. Available from: https://dev.twitter.com/streaming/overview. Date Accessed: 2016.
16. Min KJ, Young-Tack P. POI Detection and Route Identification for Building Route Models for Smartphone users. Journal of KIISE: Software and Application.2013; 40(12):799–808.
17. Oh HJ, Yun BH. Hot Place Detetion based on Social Media Content Analysis. Proceedings of the 1st International Conference on Internet of Things and Convergence. 2015. p. 231–2.