ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 8(S7), 718–723, April 2015* DOI: 10.17485/ijst/2015/v8iS7/70475

Estimating Category of POIs Using Contextual Information

Su Jeong Choi, Seong-Bae Park* and Kweon-Yang Kim

School of Computer Science and Engineering, Kyungpook National University, Daegu 702-01, Korea; sjchoi@sejong.knu.ac.kr, seongbae@knu.ac.kr, kykim@kiu.ac.kr

Abstract

With the popularity of smart phones with a GPS function, location-based applications are widely used. The location-based applications require specific information of POIs (Points of Interest) such as name, category, and exact location as their basic information. However, it costs high to gather such information manually. Furthermore, POIs are located in large geographic areas. For these reasons, the information of POIs should be automatically gathered. In this paper, we propose a method to estimate the category of POIs automatically using two kinds of POI contexts. The two contexts are internal and external information of POIs. The category of POIs is sometimes exposed by their names. Thus, their name itself is used as internal context in estimating POI category. When the category can be determined by the names, the documents that describe POIs are used as external context. Such documents are widely available through internet review sites and contain various kinds of information for POIs such as location, service satisfaction, and menu. Thus, the category of POIs can be also estimated by analyzing this information. We train a machine learning algorithm, support vector machine for each kind of information, and then combine both SVMs. According to the experimental results, the proposed method shows accuracy of 70.35% for 20 POI categories.

Keywords: Classification, POI Category Estimation, Point of Interest

1. Introduction

With the popularity of smart phones embedded with a GPS, location-based applications are being used widely. Before the popularization of smart phones, the need to find good places that meet one's interest in unfamiliar areas was not easily fulfilled. However, due to the easy use of smart phone nowadays, it is not only fulfilled but some specific information such as menu, taste, and ambiance is also provided. Foursquare[1] is one of such location-based applications. It displays Points of interest (POIs) on a map and shows reviews on them written by visitors. Yelp[2] also provides various kinds of information of POIs including menu, price, location, and reviews.

A POI is a specific point location such as a café, a gallery, a shop, or a park. It consists of a name, a category, a location, and so on. For instance, "Kokkari Estiatorio" is a POI in "Restaurant" category. It is located in San Francisco and its popular menus are Baklava and Galaktoboureko. POIs are necessary information for location-based applications. However, it requires much efforts and costs to gather information of POIs. In addition, the number of POIs is flexible and unlimited. Some POIs are not everlasting. For instance, some restaurants close in the next year of its opening. Furthermore, there are temporal or periodic POIs like locations of annual festivals. New POIs appear continuously, and some POIs are located in large geographic areas. For these reasons, the information of POIs should be automatically collected or estimated.

There are a number of documents that describe POIs on the world wide web. They are mostly reviews on stores, theaters, and so on. These reviews contain usually information of a POI visited by reviewers including its location, service satisfaction, and menu. For instance, let us consider a review on Kokkari Estiatorio that is "We had a great meal, great service and the Kokkari is beautiful! It definitely is the most gorgeous Greek restaurant we've been to. We enjoyed two great appetizers and the lamb

*Author for correspondence

chop". After reading this review, one can find out that Kokkari Estiatorio is a Greek restaurant and provides lamb chops with a great service. In addition, the information of POIs is also exposed by their internal characteristics. That is, one can gather the information of POIs such as their main services and categories only from their names. For instance, the store name "Bill's Hot Dogs" delivers that the store sales hot dogs and is an eatery. In a summary, the information of POIs can be estimated with lower efforts and costs from both their names and documents mentioning them like reviews.

In this paper, we propose a method to estimate the category of POIs automatically using two kinds of POI contexts. The two contexts used are internal and external information of POIs. The internal context is a name of POIs. The documents that describe a POI are used as an external context. For automatic classification of POI category, both contexts are expressed with Bag-of-words representation. Once a POI is represented as a document vector, estimating category of POIs is considered as a multiclass classification problem. We solve this classification problem with Support Vector Machines (SVM). We evaluate the proposed method on Yelp data set. The Yelp data set is a set of reviews collected from Yelp and has 20 categories. According to our experimental results, the proposed method achieves 70.35% of accuracy for 914 POIs, which proves its effectiveness.

The rest of this paper is organized as follows. We present an overview of the related work in Section 2. Section 3 gives a description of the proposed model to estimate the category of POIs. We show the experimental results in Section 4, and conclude our work in Section 5.

2. Related Work

As far as we know, there is no previous study of estimating the information of POIs. Most previous work focuses on POI itself. Rae et al.[3] addressed the question of whether it is possible to identify POIs automatically without manual intervention. They proposed a system for detecting POIs in unstructured texts that are a kind of unlabeled data. The system uses a Conditional Random Field (CRF) to detect mentions of POIs. The CRF is trained with not only the unlabeled data but also Wikipedia pages as a training data, and is applied to identifying POIs in free texts. It achieves higher precision and recall than those of the state-of-the-art POI identifier in Wikipedia dataset. However, the POIs in Wikipedia dataset are almost an official name for permanent structures. For overcoming this limitation, they bootstrapped training data with Foursquare and Gowalla checkins. According to their experimental results, their method improved precision up to 52% over the state-of-the-art identifier.

Web-a-Where[4] also identifies POI mentions in web pages. It assigns a geographical location to them and finds a geographical focus of the web page. The POI mentions are detected from web pages using a gazetteer. The gazetteer is a geographic dictionary which has a hierarchical view of the world, divided into continents, countries, states, and cities. The POI mentions are actually detected by occurrences of the names appearing in the gazetteer. Then, a geographic location is assigned to the detected POI mentions. The step to assign a geographical location of the POIs has two types of ambiguities: geo/non-geo and geo/geo. Geo/non-geo ambiguity occurs when a place name has another non-geographic meaning, while geo/geo ambiguity arises when two distinct places have the same name. In this work, these ambiguities are solved with heuristic rules. For the POIs of which ambiguity is resolved, they find a geographical focus of a web page, where a geographical focus is a POI that can cover geographically all POIs mentioned in a web page. For this, they proposed a focus algorithm. The focus algorithm is able to assign a geographic focus even to locations that the web page does not mention. In the experiments, 80% of precision is reported in identifying POI mentions, while just 38% of accuracy is achieved in finding geographical focuses. Zong et al.[5] also proposed a method to find geographical focus of web pages or page segments that describe POIs. The method uses the normalized counts of POI occurrences in the web page or page segments. It is similar to Web-a-Where except that it aims to POI mentions to web pages as a whole. They achieved 66% of accuracy at the web page level and 91% at the page segment level.

On the other hand, Wang et al.[6] proposed a novel probabilistic graphical model to learn the relationships between locations and words through latent topics. Assuming that each word in a document is labeled with a location, they incorporate the location information into Latent Dirichlet Allocation (LDA). The location information contributes to the latent topic discovery. This model provides a feasible way to mine the geographical knowledge from online activities including news and blogs. They showed comprehensive experimental results that demonstrate the effectiveness of their model.

KUSCO system[7] addressed the assignment of semantics to POIs. When a POI is given, KUSCO searches its related web pages. Afterwards, information extraction is applied to these web pages in order to identify semantic index of the POI. The semantic index consists of concepts contextualized in two distinct types: common concepts and specific concepts. The common concepts are assigned from common sense ontologies like WordNet. The specific concepts are generally proper nouns. They compare KUSCO with Yahoo!Term Extraction API (Yahoo!TE) to examine the diversity and richness of their module. As a result, KUSCO slightly outperforms Yahoo!TE.

3. Estimating POI Information

Our goal is to estimate the category of POIs automatically using two kinds of POI contexts. The two contexts used are internal and external information of POIs. These two contexts are represented using Bag-of-words. Once a POI is expressed as a document vector by the Bag-of-words, estimating POI category is considered as a multiclass classification problem.

In the viewpoint of machine learning, this problem is to learn a function $f(x) \rightarrow y$, where x is a vector representation of a POI context through Bag-of-words and y $\in$ Y is a category of the POI. Therefore, the goal corresponds to constructing an optimal function f(x). In this paper, the function is estimated in a supervised manner. Let D be a set of data for POIs, where $\mathrm{D} = \{(\bar{\mathrm{x}}_1, \mathrm{y}_1), (\bar{\mathrm{x}}_2, \mathrm{y}_2), \ldots, (\bar{\mathrm{x}}_n, \mathrm{y}_n)\}$, where $\bar{\mathrm{x}}_i$ denotes a context vector of the i-th POI, y_i is its category, and n is the number of POIs. The context vector $\bar{\mathrm{x}}_i$ is actually defined as $\bar{\mathrm{x}}_i = \langle nf_i, cf_i \rangle$. Here, nf_i is a vector represented by the internal context of a POI, while cf_i denotes a vector represented by its external context. $Y = \{c_1, c_2, \ldots, c_m\}$ is a set of POI categories whose cardinality is m.

Our task is to estimate the category y_i of a POI represented as $\bar{\mathrm{x}}_i$. In order to determine it, a score function $S(\bar{\mathrm{x}}, c)$ is defined as

$$S(\bar{\mathrm{x}}_i, c) = \alpha I(nf_i, c) + (1-\alpha) C(cf_i, c) \quad (1)$$

This function returns the score of a category c for a given $\bar{\mathrm{x}}_i$. Here, $I(nf_i, c)$ is a function which computes the score using nf_i of $\bar{\mathrm{x}}_i$, while $C(cf_i, c)$ computes the score using cf_i of $\bar{\mathrm{x}}_i$. All scores from $I(nf_i, c)$ and $C(cf_i, c)$ are summed up with linear weights. $0 \le \alpha \le 1$ in Equation (1) is the mixing ratio between internal and external contexts.

When $S(\bar{\mathrm{x}}_i, c)$'s are given for all $c \in Y$, the category of a POI, c^* is determined as the category with the highest score. That is, it is chosen by

$$c^* = \arg\max_{c \in Y} S(\bar{\mathrm{x}}, c) \quad (2)$$

3.1 Internal Context

The internal context is POI names. The name of a POI often contains information of the POI such as its main service, category, or location. For instance, from the name of "Austin Moto Academy", we can find out that it is a place to learn something just because of the word "Academy". That is, some particular words are representative to a category. Therefore, the category of a POI can be determined by considering such particular words rather than a word sequence.

Thus, we express the internal context in the Bag-of-words representation. That is, the internal context nf_i is defined as

$$nf_i = \langle p_i^1, p_i^2, \ldots, p_i^k \rangle$$

Here, nf_i is a word set of length k, where each p_i^j is the j-th word of $\bar{\mathrm{x}}_i$.

Once nf_i is expressed as a nf_i vector, estimating a category is considered as a multiclass classification problem. Thus, $I(nf_i, c)$ is constructed from SVMs. In this paper, we use the multiclass SVMs proposed by Crammer et al. [8] The multiclass SVMs find the optimal hyperplane that maximizes margins among all classes. This goal of the multiclass SVMs can be represented as the optimization problem as follows.

$$\min_{M, \xi} \frac{1}{2} \beta \|M\|_2^2 + \sum_{i=1}^{m} \xi_i \quad (3)$$

with constraints

$$\forall i, r, \ \bar{M}_c \cdot nf_i + \delta_{y_i c} - \bar{M}_r \cdot nf_i \ge 1 - \xi_i, \quad (4)$$

where β is a regularization constant and ξ_i is a slack variable. $\bar{M}_c$ in Equation (4) is a parameter matrix to be estimated for category c. $\delta_{p,q}$ is equal to 1 if $p = q$, and 0 otherwise. Then, the margin for category c is computed by Equation (5), and it is used as the internal score.

$$I(nf_i, c) = \bar{M}_c \cdot nf_i \quad (5)$$

In order to know for how many POIs people can identify their category only by looking their names, we conduct a simple survey. The survey is done by seven persons and

380 randomly sampled POIs from Yelp. According to the survey results, people identify only 59% of the POIs correctly. For instance, people fail in identifying the category of a POI "Flipp". It is a furniture store, but its category is not determined by its name. Therefore, we require other information to estimate the category of POIs.

3.2 External Context

In order to overcome limitation of the internal context, the external context of POIs is used to estimate their category. The external context is a set of documents that describe a POI, and the documents are usually reviews on the POI. These review documents are usually written by reviewers of the POI and deliver information of the POI such as location, service satisfaction, or menu. For instance, let us consider Figure 1 which is a review on Flipp. From the part of this review "*This compact space is full of cool furniture and interesting design pieces, and Flipp hits that style perfectly*", we can recognize that Flipp is a furniture store. That is, for many POIs, we can estimate the category of POIs from their review contents. Since reviews are used as data, estimating POI category can be regarded as a document classification problem.

The external context is represented as a document vector by the bag-of-word. That is, the external context cf_i is defined as

$$cf_i = \left\langle w_i^1, w_i^2, \ldots, w_i^l \right\rangle,$$

where W_i^j is the tf-idf value of the j-th word in the document for $\bar{x}_i$. After that, estimating a category using cf_i is considered also as a multiclass classification problem. That is, $C(cf_i, c)$ is also constructed with multiclass SVMs. Therefore, the margin for category c in the external context is computed in the same way with the internal context. It is computed by Equation (6), and used as the external score.

$$C\left(cf_i, c\right) = \bar{M}_c \cdot cf_i \tag{6}$$

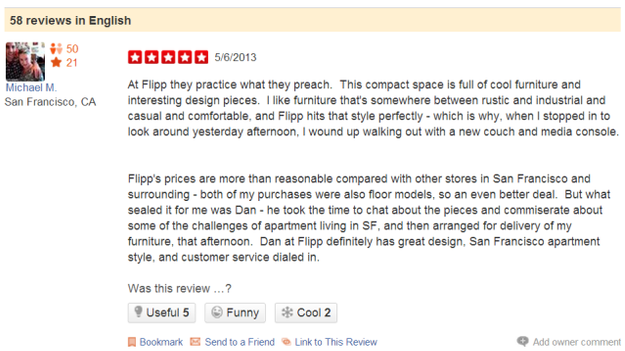

Figure 1. A review on Flipp in Yelp.

4. Experiments

4.1 Experimental Settings

The evaluation of the proposed method is performed on the Yelp dataset. The Yelp dataset is a collection of reviews in Yelp.com. Yelp provides a list of POIs and their reviews. The dataset is collected automatically by crawling reviews on POIs through Yelp API. Table 1 shows a simple statistics of the dataset. 4,613 POIs are distributed in 20 categories. A POI has 5.42 reviews on average. We use 3,499 POIs as training data, 914 POIs as test data, and the remaining 200 POIs as development data.

We estimate the weight α in Equation (1) using the development dataset. Figure 2 shows the accuracies on the development dataset according to various α's. $\alpha = 1$ implies that only internal context is used, while $\alpha = 0$ means that only external context is used. According to this table, the accuracy is lowest, when $\alpha = 1$. Its accuracy is just 0.56. The proposed method achieves the best accuracy of 0.72 with $\alpha = 0.1$. Thus, we use 0.1 as the value of α

Table 1. Categories of Yelp dataset

Category	The number of POI
Active	272
Arts	272
Auto	263
Beauty and spa	258
Education	217
Event services	252
Financial services	135
Health	190
Home Services	273
Hotels travel	237
Local services	269
Mass media	134
Nightlife	291
Pets	276
Professional	202
Public service	207
Real estate	182
Religious orgs	118
Restaurants	281
Shopping	284
Overall	4,613

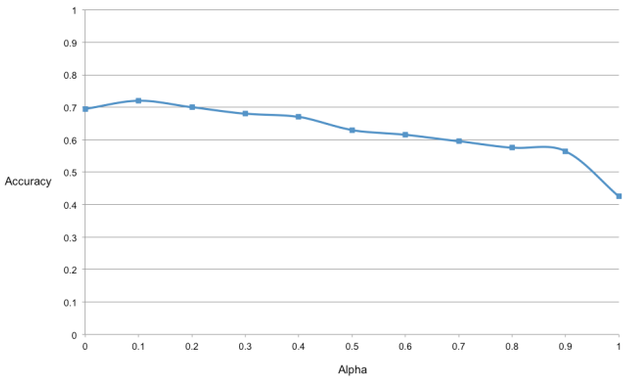

Figure 2. The accuracy changes according to the values of α.

in the following experiments. In addition, the multiclass SVM[9] is used as a classifier for all experiments.

4.2 Experimental Results

Table 2 shows the experimental results. Classifying the category of POIs using the internal context achieves an accuracy of 45.84% for 20 categories. The accuracy is satisfactory for how many POIs humans can identify their category only by looking their names. Most errors occur when the internal context does not contain any information on the POI. These errors are, however, what even human experts can't give a correct answer for. For instance, the category of "Charles Schweb" or "Komi" is really difficult to be determined by its name.

On the other hand, classifying POI category using the external context shows the accuracy of 69.39%. This is much higher than the internal context. The external context comes from a great amount of documents so that its information is larger than that of internal context. The most important reason for the errors is the length of the review documents. When a review document is short, it does not deliver information enough to estimate the category of a POI. Most portion of such short reviews is the feeling of the reviewer to the POI. For instance, let us consider a review "*My husband found here on yelp and I'm glad he did! Tyrone is the best! He contacted us back quickly and his quote was about 60 less than the phone calls I made. I can tell he's a pro in the business and knows his trade*". Note that this this review does not have any evidence for the category of a POI. Thus, it is impossible even for human experts to estimate its category only with this review.

Table 2. A Result classified categories of POIs

	Accuracy
Internal context	45.84%
External context	69.39%
The proposed method	70.35%

The proposed method achieves the highest accuracy of 70.35% by combing both the internal and external information. Even if it outperforms two base methods that uses only one context, there still exist some POIs whose category is not correctly determined even by using two kinds of information together. Such an incorrect case is a POI "Cosmopawlitan". The proposed method fails in classifying it with this name. Its review also delivers no information about its category. The only Yelp review for the POI is "*This review is slightly overdue, but my experience today at a competitor reminded me I should share my love of Cosmopawlitan*". Since the review does not contain any information on its category, the proposed method fails in classifying it correctly with the review.

5. Conclusion

We have proposed a method to estimate the category of POIs automatically using two kinds of contexts. The two kinds of contexts are the internal and the external contexts. The internal context of a POI is its name, while the external context is the documents that describe a POI. The proposed method combines the contexts in order to reflect them all into POI estimation. The proposed method achieved higher accuracy than any single context. The main source of error is that two contexts don't have any information for the category of a POI.

In the future work, we will employ a semi-supervised or an unsupervised learning into this task. In general, more data are required for better performance of machine learning techniques. However, it requires often too much costs and efforts to gather more labeled data. On the other hand, there are tremendous documents that describe POIs and the documents can be regarded as unlabeled data. If we can use such unlabeled data in training a classifier, even higher accuracy can be achieved with less labeled data. In addition, the bag-of-words representation of documents suffers from the sparseness problem. To solve this problem, we will consider other document representation rather than the bag-of-words in our future work.

6. Acknowledgement

This work was supported by the IT R and D program of MSIP/KEIT. [10044494, WiseKB: Big data based self-evolving knowledge base and reasoning platform].

7. References

1. Foursquare. Available from: http://www.foursquare.com
2. Yelp. Available from: http://www.yelp.com
3. Rae A, Murdock V, Popescu A, Bouchard H. Mining the web for points of interest. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2012. p. 711–20.
4. Amitay E, Har'El N, Sivan R, Soffer A. Web-a-where: Geotagging web content. Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval, 273-280, 2004.
5. Zong W, Wu D, Sun A, Lim E, Goh D. On assigning place names to geography related web pages. Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries; 2005. p. 354–62.
6. Wang C, Wang J, Xie X, Ma W. Mining geographic knowledge using location aware topic model. Proceedings of the 4th ACM Workshop on Geographical Information Retrieval; 2007. p. 65–70.
7. Alves A, Pereira F, Biderman A, Ratti C. Place Enrichment by mining the web. Proceedings of the European Conference on Ambient Intelligence; 2009. p. 66–77.
8. Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines. J Mach Learn Res. 2002; 2:265–92.
9. Joachims T. Making large-Scale SVM learning practical. Advances in kernel methods - support vector learning. Cambridge, Massachusetts: MIT-Press; 1999. p. 169–84.