Social Tagging Analytics for Processing Unlabeled Resources: A Case Study on Non-geotagged Photos

Tuong Tri Nguyen, Dosam Hwang, and Jason J. Jung*

Abstract. Social networking services (SNS) have been an important sources of geotagged resources. This paper proposes Naive Bayes method-based framework to predict the locations of non-geotagged resources on SNS. By computing TF-ICF weights (Term Frequency and Inverse Class Frequency) of tags, we discover meaningful associations between the tags and the classes (which refer to sets of locations of the resources). As the experimental result, we found that the proposed method has shown around 75% of accuracy, with respect to F1 measurement.

Keywords: Geotagging, Naive Bayes, Social tagging, Social networking services.

1 Introduction

Social tagging (also, called collaborative tagging) services can build a folksonomy which is a user-generated classification for resources. They have been increasingly regarded as an important research issue in social network services (SNS). There have been a number of SNS to employ the social tagging to a variety of resources (e.g., bookmarks, bibliographics, musics, and so on). Particularly, photos are the most popular resources that users want to share through SNS (e.g., Facebook, Photobucket, Instagram, Flickr and so on). In the social tagging from SNS, there have been many

Tuong Tri Nguyen · Dosam Hwang
Yeungnam University, Gyeongsan, Korea
e-mail: {tuongtringuyen,dosamhwang}@gmail.com

Jason J. Jung
Chung-Ang University, Seoul, Korea
e-mail: j2jung@gmail.com

* Corresponding author.

D. Camacho et al. (eds.), *Intelligent Distributed Computing VIII,*
Studies in Computational Intelligence 570, DOI: 10.1007/978-3-319-10422-5_37

studies which concentrate on the two main aspects; i) to understand collective behaviors among online users, and ii) to provide online services to users [6]. Most of these studies [4, 5, 10] have commonly introduced some methods to exploit the social tagging for extracting meaningful patterns and providing various services, e.g., information searching and recommendation.

In this work, we assume that the social tagging contains spatial knowledge to differentiate the tags related to geographical locations. Thus, a geotagged folksonomy from SNS is employed i) to discover meaningful patterns between the tags and geographical locations, and ii) to estimate the location of non-geotagged resources by using the patterns.

Particularly, this study focuses on Flickr[1] (which is a well-known photo-sharing SNS) to build a geotagged folksonomy. By using either i) the tags provided from the users or ii) the geographical locations of any particular topics (e.g., names of places, persons, and events), we can extract a number of various resources (e.g., photos and videos) related to the topics [5]. Hence, for building our testing bed, we have collected the social taggings from Flickr, and put all the geographical information into the database for analyzing the information and predicting the location of non-geotagged resources.

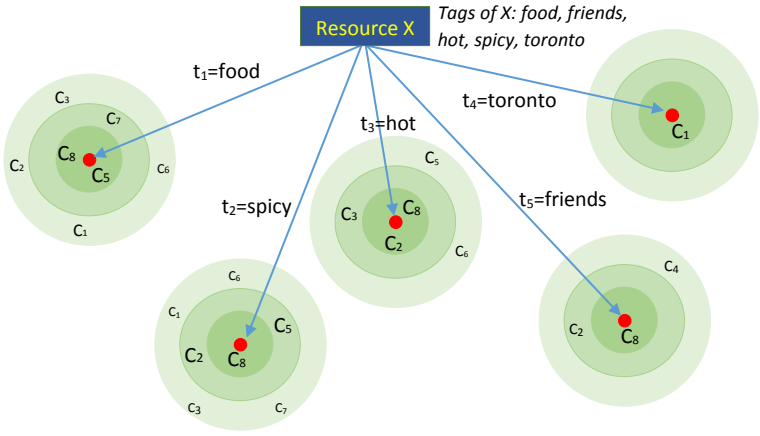

Fig. 1 Relationship between tags of resource X and a set of classes

As shown in Fig. 1, we assume that a non-geotagged resource X has 5 tags and there are 8 classes C_j (referred to as a location or a country) that X can belong to. Besides, we also know the number of occurrences of t_i in each class C_j. Thus, the question is "how can the location of X be determined?". We have to compute the weight of each tag and the probability of tags which occur in each class. In the next

[1] https://www.flickr.com

step, we determine the probability of each class and build a training set. We use the Naive Bayes method for classifying the data in the testing set.

The paper is organized as follows. Sect. 2 introduces related works. Sect. 3 shows basic knowledge and also give main steps in order to predicting location of non-geotagged resources. In the Sect. 4, experimentation has been conducted to evaluate the results of research and identify issues to be taken up for discussion. Sect. 5 draws a conclusion of this paper and indicates our future work.

2 Related Work

Several studies [1, 2] have tried to automatically collect and process the "big" data from SNS for providing the users of smart services. Some of these studies refer to tag analysis as text categorization methods [2], e.g., using tags for prediction [4] and discovering useful patterns and meaningful information from tags for recommendation. Particularly, Jung [7] exploits the tag matching to extend simple term-based queries and identify the lingual practice of each user for discovering the relationships between multilingual tags.

Also, Bischoff et al. [3] discover the associations across multiple domains and resource types and identify the gaps between the tag space and the querying vocabularies. Based on the findings of this analysis, it tries to bridge the identified gaps by focusing, in particular, on multimedia resources. By using geotagged photos, Feick and Rovertson [5] have found out a significant interaction between tag-space semantics and partial aggregation for exploring citizens sensing of urban environments (in Vancouver, Canada). Another approach [8] has used a set of geotagged photos on Flickr for extracting associative points-of-interest of a popular tourist destination in Queensland, Australia. Moreover, Clements et al. [4] is based on the Flickr geotags in the city where users have visited in order to predict a user's favorite locations in others cities and to recommend another places to the user.

Contrary to [4], we have used a set of geotagged photos with its country to determine the non-geotagged photos likely belong to the country. With this approach, we can expect to expand this research based on analyzing a set of tags of each featured country while the same refers to any problems.

3 Location Prediction by Geotagged Resources

3.1 Geotagged Folksonomy from Flickr

A folksonomy is a type of social tagging system in which the classification of data is done by users. It consists of three basic entities, which are users, tags, and resources [7]. Users create a set of tags to mark any resources, e.g., web pages,

photos, videos, and podcasts. These tags are used to manage, categorize and summarize online content. The folksonomy system also uses these tags as a way to index information, facilitate searching and navigate the resources. According to [7], a folksonomy generated by SNS is represented as $F = \langle U \times T \times R \rangle$ where R is a set of web resources described with a set of tags T by a set of users U.

Thus, as considering that some of resources are geotagged, F can be extended to the geotagged folksonomy $F^{\diamond}$.

Definition 1 (Geotagged Folksonomy). *A geotagged folksonomy is a quadruple* $F^{\diamond} = \langle U \times T \times R^{\diamond} \times \tau \rangle$, *where* $R^{\diamond} = R^{+} \cup R^{-}$ *is a set of resources. Some of them* R^{+} *are geotagged with* $\tau = \{lat, lon\}$ *which refers to the geographical coordination of the geotagged resource.*

We note that the problem of this study is to find the location τ of non-geotagged resources R^{-} by analyzing the set of geotagged resources R^{+} given from the users. For example, as shown in Fig. 1, we assume that there are 8 candidate classes (C_1 to C_8) which can potentially contain the resource $X \in R^{-}$, and we have to choose the single class as the real location of X. Thereby, the distribution of each tag of X needs to be found out.

3.2 Using TF-ICF Weight

With TF-IDF weight, we obtained the results according to what about discussion above, i.e., they will return class that has the highest probability (e.g., class 8). But, with ICF weight value, we achieved more accurate classifier. The value classified will not be class 8 such as a result of the TF-IDF. The classification by ICF weight returns class 1, and this is correct class. We can see the illustration in Fig. 2. Although only 3 tags belonging to class 1, but it is being correctly classified by TF-ICF.

Using TF-IDF to compute the term weight based on two statistics, term frequency (TF) and inverse document frequency (IDF) which are very popular in fields such

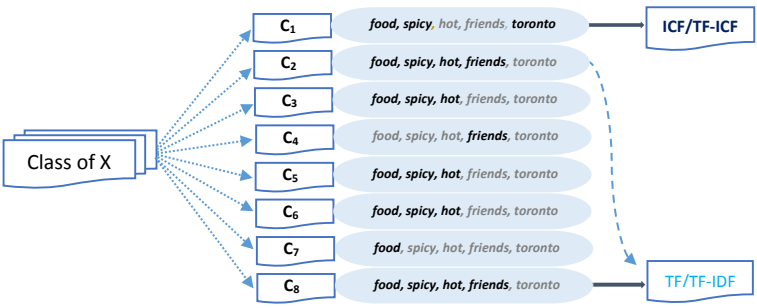

Fig. 2 Predicting location from a set of tags of the resource by TF-ICF

as classifying document [12]. They could determine the exact values of class of the document by using term frequency from set of words in that document.

For this purpose, we use geotagged resources to reach an new method for classifying the data resources from social network system. We use the location of the geotagged resources to determine the locations of the non-geotagged ones. We consider to use each location as a class for classifying. Thus, the classification problem is specified on determining location (or country) of non-geotagged photos on Flickr.

We denoted the terms as follows:

- $P(C)$ is a set of photos $\{X_1, X_2, .., X_n\}$;
- $P(C_i)$ is a set of photos of C_i;
- $P(t)$ is a set of photos tagged by t;
- $P(t, C_i)$ is a set of photos of C_i tagged by t;
- C is a set of classes (which are referred to as a set of countries);
- C_i is a class i-th, $i \in [1, m]$;
- m is the total number of classes;
- T is a set of tags;
- $T(C_i)$ is a set of tags of C_i;
- X is a photo with n tags, $X = \{t_1, t_2, .., t_n\}$;
- $\wp$ is a probability function

While the value of TF weight $\in [0, 1]$ in [11,12], this work uses TF weight $\in [0.5, 1]$ (since it is convenient for combining with Naive Bayes method). TF can be is computed by

$$tf(t_i, C_j) = 0.5 + 0.5 \times (\frac{P(t_i, C_j)}{max\{P(t_k, C_j)|t_k \in T_{C_j}\}}) \quad (1)$$

and, IDF is determined as

$$idf(t_i, C) = 1 + log(\frac{|C|}{1 + |\{C_j \in C|t_i \in T_{C_j}\}|}) \quad (2)$$

where $|C|$ is cardinality of C, or total of class in training set, $|\{C_j \in C : t_i \in T_{C_j}\}|$ is number of class contains tag t_i. TF-IDF can be computed as

$$tfidf(t_i, C_j) = tf(t_i, C_j) \times idf(t_i, C) \quad (3)$$

While we use IDF for reducing the value of tags occurred in many classes, we use formula ICF for determining exactly class frequency with tag. From IDF and coefficient, ICF can be represented as

$$icf(t_i, C_j) = (1 + log(\frac{P(t_i, C_j) + 1}{(|\{C_k \in C : t_i \in T_{C_k}\}| + 1)})) \times idf(t_i, C) \quad (4)$$

The ICF value of tags which only occur in one class (e.g., the tag 'Toronto' as shown in Fig. 1 get high ICF value as shown in Tab. 4) became useful for classifying and predicting process. Finally, TF-ICF can be computed by

$$tficf(t_i, C_j) = tf(t_i, C_j) \times icf(t_i, C_j) \tag{5}$$

3.3 Using Naive Bayes Method for Classification Problem

According to Naive Bayes theorem [11], we compute the probability of resource X is contained by class C_i by

$$\wp(C_i|X) = \frac{\wp(X|C_i)\wp(C_i)}{\wp(X)} \tag{6}$$

where $\wp(C_i|X)$ is probability of class i, contains resource X, $\wp(C_i)$ is probability of class i, $\wp(t_k|C_i)$ is probability of tag t_k in class i, $\wp(X|C_i)$ is probability of resource X in class i, and $\wp(t_k)$ is probability of tag t_k.

Here, we only consider $X = \{t_1, t_2, .., t_n\}$ and each t_j is independent with each other. Thus, $\wp(X|C_i) = \sum_{j=1}^{n} \wp(t_j|C_i)$ where n is the number of tags of resource X.

Using Equ. 6, we compute the class of resource X by getting the max value of $\wp(C_i|X)$, with $i \in [1, m]$, where m is the number of classes in training dataset. Besides, since probability value of each class has to be divided by the same value ($\wp(X)$), we can omit the denominator. We show the value of classifying probability of resource X as

$$classOf(X) = \arg\max\{\wp(X|C_i)\wp(C_i)\}. \tag{7}$$

3.4 Proposed Algorithm

To propose the classifying algorithm, we have considered using TF-ICF weight of tags by using the Naive Bayes-based classification method. By comparing to the similar work [3, 11], we propose a novel classification algorithm as follows:

In this algorithm, the training set is used in order to compute the tag weight (TF, ICF, TF-ICF and TF-IDF) for the Input set. The testing set is used to predict the location of resource and to evaluate the results. The algorithm used the Naive Bayes method to compute the probability of each class.

Algorithm 6. Algorithms for classification

Data: Training set, Testing set
Result: Geotags for Testing set
initialization;
Compute $\wp(C_i)$;
$\wp(C_i|X) = 0$;
while *photo* $X \in P_{Testing}$ **do**
  while *tag* $t_j \in T_X$, *and* $t_j \in T_{Training}$ **do**
    if $t_j \in T(C_i)$ **then**
      $\wp(C_i|X) = \wp(C_i|X) + w(t_j) \times \wp(t_j|C_i)$
    else
      $\wp(C_i|X) = \wp(C_i|X) + w(t_j) \times (1 - \wp(t_j|C_i))$
    end
  end
  Class.of(X) $\Leftarrow$ arg-max$\{\wp(C_i)\wp(C_i|X))\}$
end
{with $w(t_j)$ is TF, ICF or TF-ICF value of tag t_j;}

4 Experimentation

4.1 Dataset

We collected data from Flickr, and performed some basic data processing to obtain the data, as the basis for the experiments. As shown in Tab. 1, 4 keywords were selected to collect the dataset. On average, more than 12 tags per each photo and less than 20% of the collected photos have been geotagged.

Moreover, we also used the threshold for removing geotagged photos, if they can not create a new class (we assume that each class has more than 10 photos).

Table 1 Collecting dataset

Keyword	# photos on Flickr	# photos collected	# Geotagged photos	# Tags
kimchi	16,143	8,490	1,136	100,144
noodle	49,128	10,779	1,527	143,766
samsung	442,372	1,142	254	13,808
tower	1,413,010	6,499	2,528	114,768

After collecting data, we split them into two sets (70% in training set, 30% in testing set). We analyze the data in training set. As an example with keyword 'kimchi', we could determine 8 classes in Tab. 2.

We implemented a simple preprocessing in order to remove some tags which have no meaning (e.g., stop words) [7] and counted the number of tags for each class (some popular tags are showed as in Tab. 3).

Table 2 Extracting data with 'kimchi'

Class	# Photos	# Tags	Class	# Photos	# Tags
Canada (CA)	12	90	South Korea (KR)	248	2917
China (CN)	18	300	Taiwan (TW)	11	162
Japan (JP)	22	271	United Kingdom (UK)	52	426
North Korea (NK)	149	1520	United States (US)	283	3552

Table 3 Popular tags with 'kimchi'

Tag/Class	Canada	China	Japan	NorthKorea	SouthKorea	Taiwan	UnitedKingdom	UnitedStates
korean	6	9	9	0	96	1	37	178
hot	0	7	5	0	1	1	0	7
spicy	2	5	2	0	7	1	0	13
food	7	9	12	0	138	5	47	131
friends	0	9	0	2	0	0	0	32

4.2 Experimental Results

We compute the the value of all tags which include probability, TF, ICF, TF-IDF, TF-ICF and put them into the dataset as shown in Tab. 4.

On the following step, we use the Alg. 6 to implement. The classified results are computed the precision, recall values. Here, we use equation in [9] to calculate the F-measure values.

We implemente on the dataset with 10 iterations for the input data 10%, 20%, .. to 100% (of training set). For each iteration, we use testing set to predict location and to compute the results. The process are conducted following 3 steps (illustrations by computing value of resource X on the Fig. 1).

1. Computing $\wp(C_i)$, $\wp(t_k)$ and $\wp(t_k|C_i)$: We compute the value for classifying with TF/ICF/TF-ICF weight for resource X, given by Tab. 4 as follows:

Table 4 The results for classification

Class	food		friends		hot		spicy		toronto		Results		
	TF	ICF	TF	ICF	TF	ICF	TF	ICF	TF	ICF	TF	ICF	TF-ICF
CA	0.017	0.044	0.010	0.036	0.019	0.049	0.048	0.147	0.750	9.085	0.816	9.278	6.935
CN	0.021	0.059	0.163	0.979	0.213	0.856	0.106	0.362	0.071	0.340	0.504	2.257	1.605
JP	0.028	0.081	0.010	0.036	0.141	0.594	0.045	0.147	0.071	0.340	0.215	0.823	0.50
NK	0.001	0.002	0.033	0.197	0.019	0.049	0.013	0.031	0.071	0.340	0.033	0.197	0.101
KR	0.303	1.316	0.010	0.036	0.038	0.137	0.114	0.521	0.071	0.340	0.456	1.975	1.361
TW	0.012	0.031	0.010	0.036	0.041	0.137	0.030	0.085	0.071	0.340	0.084	0.254	0.144
UK	0.128	0.385	0.010	0.036	0.019	0.049	0.013	0.031	0.071	0.340	0.128	0.385	0.367
US	0.271	1.241	0.399	4.278	0.157	0.856	0.203	1.038	0.071	0.340	1.031	7.414	4.270

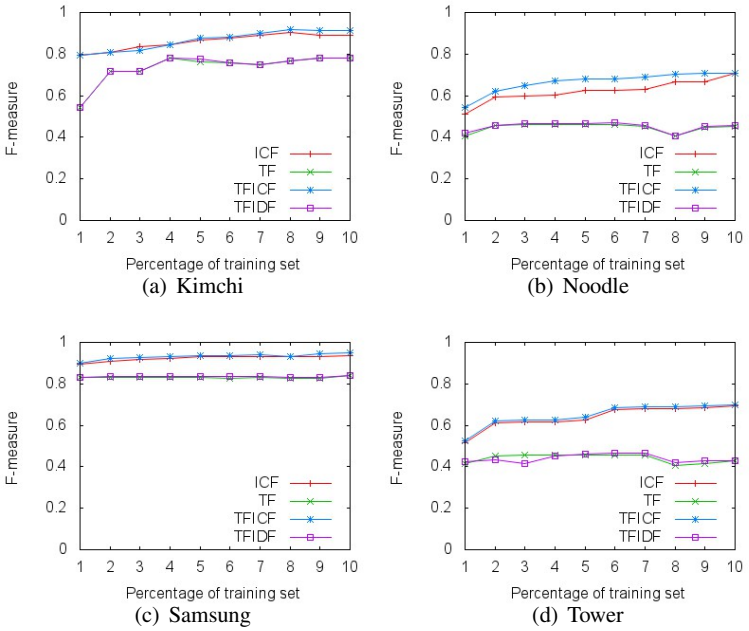

Fig. 7 Compare $F_{measure}$ with some keywords

2. Classifying for new resource $X = (t_1, t_2, \ldots, t_n)$: We calculate the probability of each class to know the country of resource X. The classified value of X is computed by using the Equ. 7 and the results are showed in Tab. 4.
3. Computing $F_{measure}$: It is used on [9], $F_{measure} = \frac{2PR}{P+R}$, where P is the precision and R is the recall.

We show the implementation results in the Fig. 7 and Fig. 8.

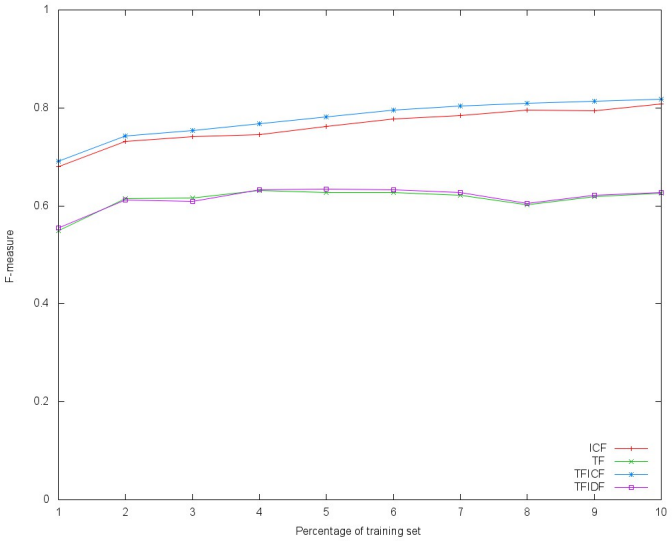

Fig. 8 Average of $F_{measure}$ with all keywords

5 Conclusion and Future Work

Predicting location of resources based on its tags might be used for classifying, categories or clustering with each place, country or city. In this paper, we implement with different keywords such as “kimchi”, “noodle”, “samsung” and “tower”. We got the results approximate more than 0.75 with F-measure value. However, we also known that the results have depended on the data of each searching keyword and also depended on the variability of the tags in training set.

Through this work, we have found out there are many issues that need further consideration, such as the construction process of the training data set, collecting data should be using multi-lingual search. Besides, the issue of handling the selected tags for the classifying should be also considered.

We improve our research by expanding the set parameters such as user data and some others collected attributes. On the other hand, we will use collected data by searching multi-language keyword same as the method which is used in [6]. We are planning

1. to combine more folksonomies which are available on the web (e.g user, owner),
2. to consider proposing new approach to recommend on SNS based on our results and
3. to rank the location through set of tags.

Acknowledgements. This work was supported under the framework of international cooperation program managed by National Research Foundation of Korea (NRF-2013K2A1A205-5213). Also, this work is supported by BK21+ of National Research Foundation of Korea.

References

1. Atzori, L., Iera, A., Morabito, G.: The internet of things: A survey. Computer Networks 54(15), 2787–2805 (2010)
2. Atzori, L., Iera, A., Morabito, G., Nitti, M.: The social internet of things (siot) when social networks meet the internet of things concept, architecture and network characterization. Computer Networks 56(16), 3594–3608 (2012)
3. Bischoff, K., Firan, C.S., Nejdl, W., Paiu, R.: Bridging the gap between tagging and querying vocabularies: Analyses and applications for enhancing multimedia {IR}. Web Semantics: Science, Services and Agents on the World Wide Web 8(2-3), 97–109 (2010)
4. Clements, M., Serdyukov, P., de Vries, A.P., Reinders, M.J.: Using flickr geotags to predict user travel behaviour. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, pp. 851–852. ACM (2010)
5. Feick, R., Robertson, C.: A multi-scale approach to exploring urban places in geotagged photographs. Computers, Environment and Urban Systems (2014)
6. Jung, J.J.: Discovering community of lingual practice for matching multilingual tags from folksonomies. The Computer Journal 55(3), 337–346 (2012)
7. Jung, J.J.: Cross-lingual query expansion in multilingual folksonomies: A case study on flickr. Knowledge-Based Systems 42(0), 60–67 (2013)
8. Lee, I., Cai, G., Lee, K.: Exploration of geo-tagged photos through data mining approaches. Expert Systems with Applications 41(2), 397–405 (2014)
9. Manning, C.D.: Foundations of statistical natural language processing, vol. 999. MIT Press
10. Morrison, P.: Tagging and searching: Search retrieval effectiveness of folksonomies on the world wide web. Information Processing and Management 44(4), 1562–1579 (2008)
11. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys (CSUR) 34(1), 1–47 (2002)
12. Zhang, W., Yoshida, T., Tang, X.: Tfidf, lsi and multi-word in information retrieval and text categorization. In: IEEE International Conference on Systems, Man and Cybernetics, SMC 2008, pp. 108–113. IEEE (2008)