Spatial Hashtags in Tweets

Qian Wu
University of Maryland at College Park
wuqian@cs.umd.edu

Hanan Samet
University of Maryland at College Park
hjs@cs.umd.edu

ABSTRACT

Twitter is an on-line social networking service which enables users to communicate by sending and reading up to 140 characters short text called ”tweets”. Users can attach any hashtag, starting with an ”#”, to tweets to indicate additional or summary information. tweets with explicit or implicit spatial intent can be classified as spatial tweets. Spatial tweets are commonly seen in practice, and it is important if we can discover such tweets and further understand the spatial information behind them. However, recognizing spatial tweets is very challenging since there is usually little spatial information in the short tweet text. In this paper, we design a Bootstrapping framework to automatically mine spatial hashtags from tweets corpus. This framework consists of three key components: PU Learning for classifier training, spatial tweets discovery and heuristic spatial hashtag extraction. Then we will show how to mine spatial hashtags from tweets with this framework and how to recommend spatial hashtags to tweets.

1. INTRODUCTION

Location based services are becoming increasingly important since mobile applications are everywhere today. For social media, such as Weibo [7] and Twitter, it is crucial to know where the event happens besides knowing what event happens. The location information associated with the tweets could be applied by many applications, such as recommending treads near the user. However, detecting the location associated with tweets is very challenging because that the tweet text is usually too short to contain location information of the event. Geotagger used in TwitterStand [6] tries to solve this problem by applying POS tagger and NER tagger, and extracting locations from a gazetteer. In this paper, we will investigate how to mine spatial hashtags, which can be used to enrich information in tweets. Thus other tools like geotagger could take advantage of this information for better geo location detection. In general, we propose to aggregate similar tweets in order to share spatial information among them, thus we can recommend spatial hashtags for tweets. To be specific, we focus on the following three aspects: recognizing spatial tweets, mining spatial hashtags and recommending spatial hashtags for tweets.

Users usually attach hashtags to tweets when they are tweeting. Although the hashtags are arbitrarily made up by different users, some popular hashtags are widely shared among lots of tweets. Thus we can use the user generated hashtags as the media of the spatial information. However, not all hashtags are spatial, and there are many random or topical hashtags. For instance, in the tweet “One of the 7 best screens to watch #StarWars is in #Seattle!”, there are two hashtags, #StarWars and #Seattle. It is obvious that #Seattle is a spatial hashtag while #StarWars is not. Besides, there are even more random hahstags, such as ”MakeAmericaGreatAgain” and ”justdoit”. Thus, distinguishing between spatial hashtags and non-spatial hashtags is one key point in our work.

We can easily find some spatial hashtags if the hashtag itself is a location name, such as #greenbelt, #collegepark, #seattle, etc. However, in practice there are also many spatial hashtags which may not appear in any gazetteer, such like #5thavenue, #GlenFalls and #umd. These spatial hashtags are not hard to recognize by a human. But it is too expensive to have human to extract all of these hashtags from a large tweets corpus. In this paper, we propose to use Bootstrapping technique to tackle this problem, that is, we start with a small set of manually found spatial hashtags, then use the set to enrich itself automatically by algorithms. Finally, we will show how to recommend the spatial hashtags for tweets.

The rest of this paper is organized as follows. In section 2, we introduce the tweets dataset we use in this work. In section 3, we describe our framework of mining spatial hashtags in detail. In section 4, we present and analyze our experimental results. In section 5, we conclude this work.

2. TWEETS DATASET

Our dataset contains 740,639 raw tweets from TwitterStand's tweets stream [6]. Some of the raw tweets associate with a geo location tagged by geotagger. The raw tweets are further normalized by the following procedure:

- Converting all the letters to lower case.
- Removing hyperlinks.
- Removing duplicate tweets according to tweet id and a hash function.
- Only keep letters, digits and #

Removing duplicate tweets based on tweet id is not enough

Hashtag	Count in the Corpus
#brussels	201
#india	135
#belgium	96
#syria	53
#paris	46
#sydney	38
#russia	30
#turkey	28
#america	22
#london	18

Table 1: Top 10 spatial hashtags by exact matching of the place names dataset

since there are lots of re-tweets with different tweet id but same content. Thus we design a hash function to further remove re-tweets in the dataset. This normalization procedure, we have we have 298,603 tweets left for further processing.

We also use a dictionary of place names including city names, state names and country names. This dictionary is used as the initial spatial hashtags in our framework which contains 142,055 place names. However, among the 298K tweet corpus, we have only flitered out 102 place names, and top 10 hashtags are shown in Table 1.

From the result of Table 1, we can see that the spatial hashtags are extremely sparse in the tweet data set. Even if our corpus is as large as 298K, the tweets covered by spatial hashtag by exact matching are only aroud 2K. This result furthermore stimulates the need of discovering more spatial hashtags.

3. KEY STRATEGIES

In this section, I will introduce the proposed framework of mining spatial hashtags from tweets.

To discover more spatial hashtags automatically, we use a Bootstrapping algorithm inspired by the Snowball algorithm [1]. First let's define the spatial tweet.

Definition .1. *A tweet is a Spatial tweet if its content can be related with some specific location, and it can be associated with some spatial hashtag.*

In our algorithm, we need a starting set of spatial hashtags, and we will con use the 102 place names mentoned in last section as the initial set of spatial hashtags or seed spatial hashtags. First let's define some denotation as following.

- D : the full tweet corpus
- S : the set of spatial hashtags
- P : subset of D, all the spatial tweets from D

Our mining framework is given in Algorithm 1.

we first initialize S to the set of the seed tags as introduced in Section 2. Then we iteratively enrich S with newly found spatial hashtags until S does not increase anymore. As we can see from algorithm 1, there are three key components in this framework:

- **PU learning component**: train a spatial tweets classifier with the current spatial tweets. Please note that U denote for unlabled tweets from tweet corpus. (Ln.4 – Ln.6);
- **Spatial tweets discovery**: apply the spatial tweets classifier on the unlabeled tweets to find out potential spatial tweets with high confidence. (Ln.7);
- **Heuristic spatial hashtag extraction**: extract new spatial hashtags from the new spatial tweets. (Ln.8).

Algorithm 1 MiningSpatialHashtags(D)

```
1:  procedure MiningSpatialHashtags
2:      $S = SeedSpatialHashtags$
3:      DO
4:          $P = \{t | t \in D \wedge t.hasHashtagIn(S)\}$
5:          $U = D - P$
6:          $C = TrainClassifier(P, U)$
7:          $Q = MostLikelySpatialtweets(C, U)$
8:          $R = MostLikelySpatialHashtags(Q, U)$
9:          $S = S \cup R$
10:     While $R$ is not empty
11:     Return $S$
12: end procedure
```

The details of each key component is described in the following three subsections.

3.1 PU Learning Component

PU learning component aims to train a spatial tweets classifier. In this step, we are given a set of spatial tweets P and a set of unlabeled tweets U, and our goal is to learn a classifier which can predict whether a tweet is spatial or not. The problem itself is a binary classification task. However, since we do not have a negative set(non-spatial tweets), it is not straightforward to train the classifier on P and U directly. Thus we use PU Learning to solve this problem. To be specific, we apply the same strategy as described in [4]. At First, we build a Rocchio [5] classifier by a single pass of the data (including both P and U), by which we can compute two centroid tweets as following:

$$\vec{c}^{\,+} = \alpha \frac{1}{|P|} \sum_{\vec{t} \in P} \frac{\vec{t}}{\left\|\vec{t}\right\|} - \beta \frac{1}{|U|} \sum_{\vec{t} \in U} \frac{\vec{t}}{\left\|\vec{t}\right\|} \tag{1}$$

$$\vec{c}^{\,-} = \alpha \frac{1}{|U|} \sum_{\vec{t} \in U} \frac{\vec{t}}{\left\|\vec{t}\right\|} - \beta \frac{1}{|P|} \sum_{\vec{t} \in P} \frac{\vec{t}}{\left\|\vec{t}\right\|} \tag{2}$$

As suggested in their paper, we set $\alpha = 16$ and $\beta = 4$. Then for each tweet $\vec{t} \in U$, if $Sim(\vec{t}, \vec{c^+}) \leq Sim(\vec{t}, \vec{c^-})$, we add $\vec{t}$ into a set N(negative set) which is empty before running Rocchio. Note that the tweet is represented as a feature vector here, in our project, we use the tf-idf representation[1].

The output of Rocchio method is the negative set N. With positive set P and negative set N, we now are able to learn the classifier using normal binary classification algorithms. In this project we use LibLinear [3] library to train the classifier. Another important strategy we take is Bagging [2]. Since in the following step we need to measure the confidence of a classification result, we actually trained 5 classifiers on different samples of P and N. So the ultimate result of this step is a set of classifiers $C = \{C_1, C_2, C_3, C_4, C_5\}$.

[1] https://en.wikipedia.org/wiki/Tf%E2%80%93idf

3.2 Spatial Tweets Discovery Component

In this step, we are given a set of classifiers C and a set of unlabeled data U, the objective is to find out some highly confident spatial tweets from U. Details is illustrated in Algorithm 2.

Algorithm 2 MostLikelySpatialtweets(C,U)

```
 1: procedure SpatialtweetsDiscovery
 2:     $Q = \emptyset$
 3:     For Each $t$ in $U$
 4:         $Vote = 0$
 5:         For Each $c$ in $C$
 6:             If $c$.classify($t$) == Positive
 7:                 $Vote = Vote + 1$
 8:             EndIf
 9:         EndFor
10:         If $Vote >= threshold$
11:             $Q = Q \cup \{t\}$
12:         EndIf
13:     EndFor
14:     Return $Q$
15: end procedure
```

Since we have 5 classifiers in C, we set the threshold in Ln.10 to the value of 4 in this project, which denotes a fairly high (80%) classification agreement/confidence. The output of this step is the set of new spatial tweets Q.

3.3 Heuristic Spatial Hashtag Extraction

The goal of the last component is to extract new spatial hashtags from the new spatial tweets Q. Recall the definition of spatial tweets, the intuition of this step is that if a spatial tweet contains a hashtag, it could be a spatial hashtag. Therefore, for each tweet t in Q, if t contains a hashtag h, we perform heuristic analysis on h. After analyzing all the h's, we summarize the results and returns the most confident h as the new spatial hashtags. The details are shown below in Algorithm 3, in which $Z_+[h]$ represents how many times hashtag h appears in spatial tweets and $Z_-[h]$ represents how many times hashtag h appears in non-spatial tweets.

As shown in algorithm 2, we first count the number of both spatial and non-spatial tweets for each hashtag h, after passing through the full dataset, our counts may collect statistically significant information to pick out high confident spatial hashtags. The confidence of a hashtag is calculated as in Ln.18, where we use a predefined threshold to filter out those hashtags with low confidence. This formula, can also be set to the following to reflect the quantity of positive appearance as discussed in [1].

$$Confidence\,(h) = \frac{Z_+\,[h]}{Z_+\,[h] + Z_-\,[h]} \log Z_+[h] \tag{3}$$

According to this confidence function, we can generate an reliable set R of the newly found spatial hashtags.

4. EXPERIMENTS

In this section, we will evaluate our spatial hashtag mining framework based on tweets datasets introduced in Section 2. Besides the analysis to each key component performance, we also implement a application to assign spatial hashtags to tweets inorder to enrich the information in tweets.

Algorithm 3 MostLikelySpatialHashtags(Q,U)

```
 1: procedure SpatialHashtagsDiscovery
 2:     $R = \emptyset$
 3:     $N = U - Q$
 4:     $Z_+[h] = 0$ For Any $h$
 5:     $Z_-[h] = 0$ For Any $h$
 6:     For Each $t$ in $Q$
 7:         For Each $h$ in $t$.hashTags
 8:             $Z_+[h] = Z_+[h] + 1$
 9:             $R = R \cup \{h\}$
10:         EndFor
11:     EndFor
12:     For Each $t$ in $N$
13:         For Each $h$ in $t$.hashTags
14:             $Z_-[h] = Z_-[h] + 1$
15:             $R = R \cup \{h\}$
16:         EndFor
17:     EndFor
18:     For Each $h$ in $R$
19:         $Confidence = Z_+[h]/(Z_+[h] + Z_-[h])$
20:         If $Confidence < threshold$
21:             $R = R - \{h\}$
22:         EndIf
23:     EndFor
24:     Return $R$
25: end procedure
```

#delhi hc adjourns hearing on #kanhaiyakumar s bail cancellation to apr 28 read
#belgium zaventem airport suicide attackers identified brothers bakraoui 3ed man on the run is najim laachraoui
belgian media reports #brussels airport bombers named as brothers linked to paris suspect #abdeslam
#india pm modi pays tribute to bhagat singh rajguru and sukhdev on martyrs day read more
manamohana #arizona voters are my heroes you did not yield you stood your ground you cast your votes #feelthebern

Table 2: Positive Spatial tweets Sample

4.1 Evaluation on PU Learning Component

4.1.1 Creating Positive Dataset

In the PU learning component, the positive Spatial tweets are generated by filtering tweets with seed hashtags. We started the first iteration of our algorithm with a predefined place name list as introduced in Section 2. The positive dataset is obtained by collecting the tweets containing the seed hashtags. The resulting positive dataset contains 4,745 tweets in total which is far less than the full corpus size of 298,603. This demonstrates the severe problem of sparsity in both spatial tweets and spatial hashtags. Table 2 lists some randomly selected postive data obtained in this step. As we can see from table 2, by carefully selecting the seed hashtags, we can obtain a trustful positive dataset P.

4.1.2 Creating Negative Dataset

As stated in last subsection, Rocchio method in PU learning will generate a negative set of documents. We represent a tweet by its tf-idf feature vector, which is then fed into Rocchio to determine if its a negative tweet or not. Finally, we obtained 256,604 negative tweets. Which is far more than the positive tweets. This further introduces the class imbalance problem – if we train a classifier on the positive and negative datasets directly, the majority negative datasets will dominate the performance of classification. To alleviate this problem, we use upsampling technique on the positive dataset to make them balanced before we train the classifier.

4.1.3 *Training Classifiers*

To enable measuring confidence level of classification and make the classification more stable, we use Bagging technique to aggregate multiple classifiers together. We do random sub-sampling on both P and N for 5 times, which generates 5 sub-samples of P and N, denoted by $\{P_1,N_1\}$, $\{P_2,N_2\}$, ..., $\{P_5,N_5\}$. Then we train 5 binary classifiers on the 5 sub-samples. The classifiers aggregated as an ensemble model, are then sent to the Spatial tweet Discovery component.

It is very important to introduce confidence measurements into the tweet classification since the performance of a single classifier is usually not good enough. Figure 1 illustrates the F1 measure during a training process.

As we can see, the classifier performs poorly on the testing set. The F1 measure achieved its peak value around 0.7 very quickly, even though F1 score increases constantly on the training set. The reason behind this may be two folds – Firstly, the classification of tweets is a well-known hard problem, since the length limit of 140 characters limits the amount of information in the tweet. Secondly, as spatial tweets are extremely sparse, the scale of our positive dataset is about only 1.5% of the full dataset. With such a small scale, it is too hard to collect enough information about spatial tweets. This difficulty further emphasises the need of classification confidence. Considering the poor prediction performance, we can only trust the results that have high confidence levels.

4.2 Evaluation on Spatial Tweets Discovery

As we just discussed, the Spatial Tweets Discovery component will take an input of an ensemble classifier which consists of 5 separate models. The main reason is that we need a confidence level to measure how trustful the classification result is. Here since we have 5 models, for each tweet being classified, we will have 5 predictions from each model. We take these predictions as voting. If the total number of positive predictions is at least K (a predefined threshold), we will trust the classification result and accept the tweet as a spatial tweet. By applying this process on all the unlabeled tweets in set U, we can generate a list of candidate spatial tweets. Table 3 shows some random samples of the tweets which are accepted to be Spatial tweets.

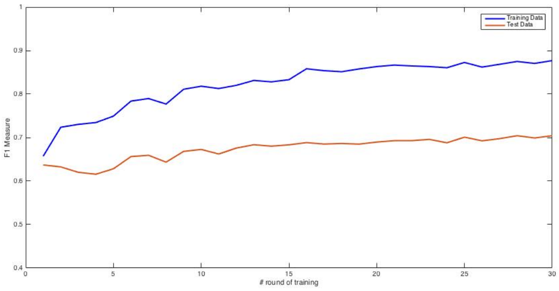

Figure 1: Classifier Training Performance

spring blizzard shuts down denver airport affects mccarran #denverinternationalairport
johnspatricc this is belgium s new normal after the paris attacks #whereisbelgium
portiafox5 spotted this belgian flag tribute to #brusselsattacks on smoke stack near downtown atl connector gooddayatlanta
mali releases photos of gunmen killed in hotel attack state media broadcast photos monday of the two attacker
#pakistan news london holds vigil for brussels terror attack victims
cnnbrk this was leuven for a moment of silence main university library we also applauded #brusselsattacks
brussels attack investigation 1 arrested and what are they planning to do next those are some #parisattacks
obama on terror attacks we stand in solidarity #belgiumbombing #whereisbelgium
#jesuisbruxelles trending reactions to tuesday s attacks on the belgian capital
authorities search #indigo flights after bomb threat
world ex bosnian serb leader karadzic sentenced to 40 years for genocide #japantoday #news
premier li extends deep condolences to victims of #brusselsattacks says #china opposes all forms of terrorism

Table 3: Positive Spatial tweets Predicted by the Ensemble Classifier

As we can see, among the 12 spatial tweets in Table 3, each of them contains some spatial hashtag. Despite that hashtags are rare, these hashtags are mostly spatial – such as #whereisbelgium, #parisattacks and #pakistan, which contain spatial information. Only #news is clearly not spatial. This insight is the motivation of our next step.

4.3 Evaluation on Spatial Hashtags Extraction

We use K=4 in the last step to filter the high confidence spatial tweets. As a result, we got about 1.7% of the unlabeled set U classified as positive. We put these classified positive tweets into a set Q, then we apply Algorithm 3 on classified positive set Q and unlabeled set U.

Algorithm 3 will generate a list of spatial hashtags from the input spatial tweets. To make the result selectable, it also attaches a confidence value with each hashtag. We denote this as Hashtag Confidence. The normal formula of computing hashtag confidence is given in Ln.19 of Algorithm 3. However, in practice we used the formula of equation (3) since it takes the quantity into account. The top 10 spatial hashtags in terms of hashtag confidence is show in Table 4.

From the result of Table 4, we can see that the leading results look pretty good. In the top 10 hashtags, only 3 out of 10 are not spatial, and they are "#news", "#topstories" and "#breaking". However, "news" seems to appear frequently in spatial tweets, thus it is resonable for "#news" to be mined. Looking at the spatial hashtags we mined, such as "#brusselsattacks" and "#parisattacks", they are clearly spatial hashtags. And it is reasonable that the sparsity problem caused non-spatial hashtags such as "#news" being high confident – this hashtag almost always appeared together

Hashtag	Z_+	Z_-	Confidence
#news	782	561	3.88
#brusselsattacks	1252	1119	3.77
#topstories	52	7	3.48
#parisattacks	147	65	3.46
#stopislam	68	21	3.22
#brusselsairport	48	13	3.05
#najimlaachraoui	21	0	3.04
#breaking	521	560	3.01
#brusselsattack	73	42	2.72
#schaerbeek	12	0	2.48

Table 4: New Spatial Hashtags by descending order in hashtag confidence

with other popular spatial hashtag.

4.4 Overall Performance Evaluation

In the previous experiments, we have successfully mined some new spatial hashtags in a single iteration. Since our framework is an iterative algorithm, next we will run through the whole process for 3 iterations. And at the end of each iteration, we will expand the seed hashtags set by adding the mined hashtags with confidence >= 1.7. The new spatial hashtags we add to the seed hashtags set in each iteration are shown in Table 5.

Iteration 1			
#brusselsattacks	1252	1119	3.77
#parisattacks	147	65	3.46
#stopislam	68	21	3.22
#brusselsairport	48	13	3.05
#najimlaachraoui	21	0	3.04
#brusselsattack	73	42	2.72
#schaerbeek	12	0	2.48
#sudarsanpattnaik	9	0	2.20
#brusselsblasts	18	6	2.17
#belgiumflag	17	6	2.09
#france24	17	6	2.09
#belgian	8	0	2.08
#laachraoui	8	0	2.08
#japantimes	13	5	1.85
#jesuisbruxelles	14	7	1.76
Iteration 2			
#bruxelles	21	3	2.67
#siachen	10	0	2.30
#brusselsterrorattack2016	6	0	1.79
Iteration 3			
#europe	42	31	2.15
#prayforbrussels	7	0	1.95
#daesh	15	7	1.85

Table 5: New Spatial Hashtags Discovered in 3 Iterations

As shown in Table 4, the leading hashtags in the first iteration looks promising, and we obtained 15 new spatial hashtags. However, as we proceeds to later iterations, the leading hashtags become less reliable, and the confidence value drops more quickly. We argue that this could be alleviated by using more tweet data (i.e. using larger corpus D) since most incorrect hashtags are due to their frequent co-occurances with other spatial hashtags.

spatial hashtag	tweet
#belgian	video belgian prosecutor briefing on attacks news briefing on the terror attacks in brussels on tuesday whi
#holland	spacekatgal this is a highly respected professor and co inventor of vlsi she literally cowrote the book on microchip design in the 70s

Table 6: Spatial Hashtag Recommendation Results

4.5 Spatial Hashtag Recommendation

With a set of spatial hashtags, we can do some interesting stuff. One useful application is to recommend the most appropriate spatial hashtag to a spatial tweet that contains no hashtags.

Our approach works like this: Firstly, for each spatial hashtag h, we compute a centroid vector representation for it. It is computed by taking the average vector over all the spatial tweets in set P which contains hashtag h. Then, for a tweet containing no hashtag, we use our current classifier to predict that if the tweet is spatial or not. If it is not a spatial tweet, we do not recommend any spatial hashtag to it. If it is, we find the nearest spatial hashtag to it using the distance between its vector representation and the hashtag's centroid vector. Finally, we check the distance to the nearest spatial hashtag, if it is below a threshold γ, we will recommend the spatial hashtag to the tweet. We have rwo interesting examples shown in Table 6.

From Table 6, we can see that the spatial hashtag recommendation is a pretty tough problem. Clearly, the first spatial hashtag "#belgian" is highly relevent to the tweet content and the recommendation seems very resonable. However, from the text of the tweet, we cannot tell whether the second spatial hashtag "#holland" is relevant to the tweet or not. Thus our future work includes how to automatically evaluate the recommendation of spatial hashtag. In summary, recommend spatial hashtag to tweets is just one application of spatial hashtags, which could help to enrich the information in tweets.

5. CONCLUSION

In this paper, we proposed a framework of mining spatial hashtags from totally unlabeled tweet corpus. Our framework starts with a seed set of spatial hashtags which can be easily obtained from a gazetteer. Then the framework iteratively discover new spatial hashtags in an automatic fashion. The seed of spatial hashtags keep increasing as more and more spatial hashtags are mined. We use PU Learning with Rocchio method to tackle the problem of lacking negative examples. Besides, we use Bagging technique to make our internal prediction model more reliable. Finally, we observed the effectiveness of our approach via various experiments. At the same time, we also noticed the following aspects to improve in the future investigation: 1) Better feature representation should be conducted to handle the information shortage in tweet text. 2) More data is always desirable.

6. REFERENCES

[1] E. Agichtein and L. Gravano. Snowball: Extractingrelations from large plain-text collections. In *International Conference on Digital Libraries*, 2000.

[2] L. Breiman. Bagging predictors. In *Technical Report 421,Department of Statistics, University of California at Berkeley*, 1994.

[3] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. In *Journal of Machine Learning Research*, 2008.

[4] X. Li and B. Liu. Learning to classify text using positive and unlabeled data. In *IJCAI*, 2003.

[5] J. Rocchio. The smart retrieval system: experiments in automatic document processing. In *Englewood Cliff NJ*, 1971.

[6] J. Sankaranarayanan, H. Samet, B. Teitler, and J. Lieberman, M. Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 42–51, 2009.

[7] L. Yu, S. Asur, and B. A. Huberman. What trends in chinese social media. In *The 5th SNA-KDD Workshop*, 2011.