Pattern Recognition Letters
journal homepage: www.elsevier.com

Tweet-SCAN: An event discovery technique for geo-located tweets

Joan Capdevila[a,b,**], Jesús Cerquides[c], Jordi Nin[a,b], Jordi Torres[a,b]

[a]*Department of Computer Architecture, Universitat Politècnica de Catalunya (UPC), Jordi Girona 1-3, Barcelona 08034, Spain*
[b]*Barcelona Supercomputing Center (BSC-CNS), Jordi Girona 1-3, Barcelona 08034, Spain*
[c]*Institut d'Investigació en Intel.ligència Artificial (IIIA-CSIC), Campus UAB, Cerdanyola 08193, Spain*

ABSTRACT

Twitter has become one of the most popular Location-Based Social Networks (LBSNs) that bridges physical and virtual worlds. Tweets, 140-character-long messages, are aimed to give answer to the *What's happening?* question. Occurrences and events in the real life (such as political protests, music concerts, natural disasters or terrorist acts) are usually reported through geo-located tweets by users on site. Uncovering event-related tweets from the rest is a challenging problem that necessarily requires exploiting different tweet features. With that in mind, we propose Tweet-SCAN, a novel event discovery technique based on the popular density-based clustering algorithm called DBSCAN. Tweet-SCAN takes into account four main features from a tweet, namely content, time, location and user to group together event-related tweets. The proposed technique models textual content through a probabilistic topic model called Hierarchical Dirichlet Process and introduces Jensen-Shannon distance for the task of neighborhood identification in the textual dimension. As a matter of fact, we show Tweet-SCAN performance in two real data sets of geo-located tweets posted during Barcelona local festivities in 2014 and 2015, for which some of the events were identified by domain experts beforehand. Through these tagged data sets, we are able to assess Tweet-SCAN capabilities to discover events, justify using a textual component and highlight the effects of several parameters.

1. Introduction

Twitter[1] is one of the most popular Social Networks and microblogging sites offering location-based services to identify the geographical location of social content, e.g. tweets. A tweet is a 140-character-long status message that responds to the question *What's happening?* This update message is associated with a user, a posting time and might contain some sort of geographical localization, among other metadata. In fact, Weidemann (2013) showed that one-in-five tweets is geo-located or its location can be inferred from user metadata. Given that about 500 millions tweets are generated per day[2], understanding some of the physical world behaviors from geo-located tweets seems now feasible.

There are numerous research papers supporting the use of Twitter in a broad range of fields from politics —Borge-Holthoefer et al. (2011) studied the dynamics of the Spanish political movement called 15M, epidemics —Kim et al. (2013) proposed to improve forecasting of human influenza infection, to seismology —Sakaki et al. (2010) presented a detection and monitoring system to track earthquakes. As a matter of fact, we can view Twitter as a rich source of data generated by millions of distributed users acting as sensors that report what is happening right now worldwide.

An event happening in a specific location (such as a demonstration, a music concert, an accident or a street fight) will be likely reported on Twitter by means of geo-located tweets posted by users close to the event location. Nonetheless, these events are usually masked by tweets which do not contribute to any particular pattern and which can be considered noise for the event detection task. Therefore, the problem of event discovery in Location-based Social Networks (LBSNs), and specifically in Twitter, consists in uncovering and determining these events while excluding the undesired observations (Zheng, 2012).

**Corresponding author: Tel.: +34-934-054-055;
e-mail: jc@ac.upc.edu (Joan Capdevila), cerquide@iiia.csic.es (Jesús Cerquides), nin@ac.upc.edu (Jordi Nin), torres@ac.upc.edu (Jordi Torres)

[1]http://www.twitter.com/
[2]https://blog.twitter.com/2013/new-tweets-per-second-record-and-how

In fact, we propose to frame the event discovery problem within a clustering type of problem in which clusters are dense groups of tweets posted by different users that talk about an event happening nearby their location. Tweets not related to the events are unwanted and we aim to group them together into a noise cluster. In a nutshell, the event discovery problem described here can be seen as an unsupervised machine learning problem which aims to group together similar event-related tweets.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) proposed by Ester et al. (1996), is a density-based clustering algorithm in which clusters are arbitrary shaped regions with higher density of spatial points. The algorithm defines three types of points depending whether they belong to a dense, a sparse or an intermediate region: *core points*, *noise points* and *border points*, respectively. GDBSCAN (Generalized DBSCAN) proposed by Sander et al. (1998), generalizes DBSCAN to use spatially extended objects instead of simply spatial points and more advanced predicates beyond euclidean proximity. This algorithm is a convenient framework to define a technique capable of uncovering clusters of event-related tweets from the rest.

Therefore, we present Tweet-SCAN, a novel event discovery technique which adapts the DBSCAN algorithm —or particularize GDBSCAN— to cope with Twitter objects, considering its spatial, temporal, textual and user dimensions. Tweet-SCAN considers spatially extended objects from DBSCAN and it implements independent neighborhood identification in each separate dimension to group close neighbors into a dense cluster which is finally associated to an event.

The textual part of a tweet is modeled through a probabilistic topic model (Blei, 2012), named Hierarchical Dirichlet Process (HDP) (Teh et al., 2006), which can be seen as the nonparametric extension of Latent Dirichlet Allocation (LDA) (Blei et al., 2003). This nonparametric topic model represents the textual dimension of each tweet as a Categorical probability distribution over topics. To assess similarity of tweet messages, we propose to use Jensen-Shannon distance (Endres and Schindelin, 2003), a proper and natural metric for probability distributions which outperforms other measures in terms of semantic similarity (Ljubešić et al., 2008) and categorization accuracy (Li et al., 2011)[3].

The algorithm capabilities to uncover events are assessed in a real data set composed of geo-located tweets from Barcelona during its local festivities in September of 2014 and 2015, called "La Mercè". This data set has been crawled through the Twitter Streaming API via a distributed system called Hermes (Cea et al., 2014). It has been shown that the Twitter Streaming API returns all geo-located tweets within the bounding box, instead of a sample (Morstatter et al., 2013). Furthermore, some tweets have been manually tagged and several events have been assigned to them based on our expert knowledge about the festivities. This tagging process allows to quantitatively evaluate the algorithm and to interpret the algorithm parameters.

[3]Other measures, such as cosine similarity, can be also applied here.

The rest of the paper is organized as follows: First, we present the digested background for this study in Section 2. Next, Tweet-SCAN technique is described in detail in Section 3. Section 4 contains a descriptive analysis of both data sets from "La Mercè" festivities. Then, we assess Tweet-SCAN discovering capabilities by studying different parameter settings, see Section 5. To conclude, we present the main conclusions for this work and identify future challenges in Section 6.

2. Background

There is a broad literature on event discovery in Location-based Social Networks which can be further classified based on the features used by the algorithm. Following the well structured background study presented by Yuan et al. (2013) with regards to geographical topic modeling, we attempt to classify the event discovery literature in a very similar manner. As stated above, we group the existing proposals in this area of research based on the combination of features used: time, content, location and user.

- **Time:** Simply using time dimension to detect social events from LBSNs has shown encouraging results when exploiting diversity on multiple social networks. For example, Garcia-Gasulla et al. (2014) presented a model for the normal behavior of a city in order to then detect abnormal situations associated to real world events. They applied this model to a data set composed of aggregated observations from Twitter, Instagram and Foursquare. They have also considered Barcelona city as the test bed for their evaluation probably due to the high penetration of Social Networks and the vast number of events in the city.

- **Time, content:** Weng and Lee (2011) presented an algorithm to perform event detection in Twitter which tackles the problem by applying wavelet analysis on the frequency-based raw signals of words (content). Despite being defined in the frequency domain, this algorithm can be still included into this group since frequency is simply a transformation of time domain. An algorithm that was directly defined in the content-time domain was presented by Fung et al. (2005) in which they attempt to identify bursty words that explains real-life events in traditional media, instead of Twitter.

- **Time, location:** When searching for a specific event, existing proposals have relied on prior classification of tweets and then applying spatio-temporal techniques to detect and track the event from the set of classified tweets. For example, Sakaki et al. (2010) proposed an Earthquake detection system which detects and tracks an earthquake by applying Kalman filtering to Tweets semantically related to the phenomena. This approach requires to know the event beforehand in order to perform supervised classification of tweets and then uncover the event-related tweets from the rest.

- **Time, content, location:** Chen and Roy (2009) presented an event detection system from Flickr data[4] which exploits temporal, spatial and content information from photos. Instead of the image, they benefit from the photo annotations or *tags*. Authors employed here wavelet transform to suppress noisy observations and uncover events similarly to (Weng and Lee, 2011). A different approach within this group was presented by McInerney and Blei (2014), where they proposed a hierarchical probabilistic model to model newsworthy events from Twitter which accounts for all three dimensions. This probabilistic model assumes Gaussian shaped events within the spatial and temporal dimensions, which also limits the expressivity in real case scenarios. They also relied on a separate data set (New York times) for training the model and they performed transfer learning to detect the newsworthy tweets.
- **Time, content, location, user:** To our best knowledge, there is not yet any unsupervised event discovery algorithm or technique that considers all four dimensions. However, Yuan et al. (2013) presented a probabilistic model to discover geographical topics which it is not intended for event detection as they stated in the article. Hence, our proposed algorithm is a first attempt to consider all four features for the task of event discovery.

In the field of geographical topic modeling, Zhang et al. (2013) conducted a study of DBSCAN to consider the textual dimension. In their work, authors examined different setups to configure DBSCAN with LDA for discovering meaningful geographical topics. Although this work used similar state-of-the-art clustering and topic model techniques to Tweet-SCAN, we seek to detect abnormalities (e.g. events) in the data rather than finding a model that best explains geo-located data in terms of topics.

3. Tweet-SCAN technique

Tweet-SCAN is an event discovery technique for geo-located tweets which is based on the DBSCAN clustering algorithm presented by Sander et al. (1998). In the following section, we describe Tweet-SCAN by first introducing the main elements of DBSCAN (Ester et al., 1996) and then generalize them into GDBSCAN to make clustering of tweets possible. The proper definitions of GDBSCAN predicates will enable that the resulting Tweet-SCAN clusters matches real world events. Therefore, we define proper Tweet-SCAN predicates and present a text model for tweets.

3.1. Events as density-connected sets

DBSCAN (Ester et al., 1996) was proposed to uncover clusters with arbitrary shapes whose points configure a dense or packed group. This means that for each point in a cluster its neighborhood at a ϵ distance must contain at least a minimum number of points, $MinPts$. Formally, this implies the definition of two predicates:

1. $NPred(o, o') \equiv N_\epsilon(o, o') = |o - o'| \leq \epsilon$.
2. $MinWeight(o) \equiv |\{o' \in D \mid |o - o'| \leq \epsilon\}| \geq MinPts$.

The fulfillment of both predicates allows to define the notion of a point p being directly density-reachable from another point q, see (left) figure 1, where ϵ is given by the circle radius and $MinPts$ is set to 2. In this scenario, q is a *core point* because it satisfies both predicates and p is a *border point* since it breaks the second predicate. The notion of being direct reachable is extended to density-reachable points when p and q are far apart, but there is a chain of points in which each pair of consecutive points are directly density-reachable, as it is the case in (middle) figure 1. Finally, it might happen that p and q are not density-reachable, but there is a point o from which they are both density-reachable, that is when p and q are said to be density-connected, for example in (right) figure 1. Note that both points, p and q, are here *border points*, while o is a *core point*.

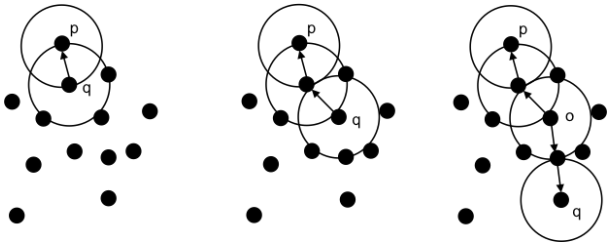


Fig. 1: Directly density-reachable (left), density-reachable (middle) and density-connected (right) points.

Consequently, a cluster in DBSCAN is defined to be a set of density-connected points that contains all possible density-reachable points. Furthermore, *noise points* can now be defined as those points which do not belong to any cluster since they are not density-connected to any.

GDBSCAN (Sander et al., 1998) generalizes DBSCAN by redefining the above-mentioned predicates to cope with spatially extended objects. For example, the neighborhood of a set of polygons is defined by the intersect predicate instead of a distance function. It is also the case for a set of points with financial income attributes within a region whose $MinWeight$ predicate is a weighted sum of incomes instead of mere point cardinality, so that clusters become regions with similar income. Therefore, both predicates can be generalized as follows:

1. $NPred(o, o')$ is binary, reflexive and symmetric.
2. $MinWeight(o) \equiv wCard(\{o' \in D \mid NPred(o, o')\}) \geq MinCard$, where $wCard$ is a function that $2^D \rightarrow \mathbb{R}^{\geq 0}$

These new predicates enable to extend the concept of density-connected points to objects and thus generalize density-based clustering to spatially extended objects. Particularly, this extension allows us to formulate the event discovery problem for tweets in the framework of GDBSCAN, which we call Tweet-SCAN. Tweet-SCAN specifies proper neighborhood and MinWeight predicates in order to associate density-connected sets of tweets to real world events. The following subsections deal with these specifications by defining the neighborhood and MinWeight predicates as well as explaining how the textual content from a tweet is modeled.

[4] http://www.flickr.com/

3.2. Tweet-SCAN neighborhood predicate

Most of the event-related tweets are generated throughout the course of an event within the area where it takes place. Although, it can happen that some users might be tweeting about an event remotely or even long after it has finished, these tweets cannot be considered part of it. Consequently, we need to find density-connected sets of tweets close in space and time, as well as similar in meaning. We also note that closeness in space is not comparable to time, nor to meaning.

Because of this, Tweet-SCAN is defined to use separate positive-valued ϵ_1, ϵ_2, ϵ_3 parameters for space, time and text, respectively. Moreover, specific metrics will be chosen for each dimension given that each feature contains different type of data. The neighborhood predicate for a point o in Tweet-SCAN can be expressed as follows,

$$NPred(o, o') \equiv |o_1 - o'_1| \leq \epsilon_1,\ |o_2 - o'_2| \leq \epsilon_2,\ |o_3 - o'_3| \leq \epsilon_3 \quad (1)$$

where $|o_i - o'_i|$ are distance functions defined for each dimension, namely space, time and text. The predicate symmetry and reflexivity are guaranteed as long as $|o_i - o'_i|$ are proper distances. Particularly, we propose to use the Euclidean distance for the spatial and temporal dimensions given that latitude and longitude coordinates as well as timestamps are real-valued features and the straight line distance seems a reasonable approximation in this scenario[5]. The metric for the textual component will be defined later once we present the text model for Tweet-SCAN.

If we scale each metric in the above predicate with its corresponding ϵ_i parameter, the predicate then must satisfy that the maximum scaled component is less or equal than 1. Each component being the distance for each separate dimension. Writing this down in term of the ∞-norm metric leads to the following expression,

$$NPred(o, o') \equiv \left\| \frac{|o_1 - o'_1|}{\epsilon_1},\ \frac{|o_2 - o'_2|}{\epsilon_2},\ \frac{|o_3 - o'_3|}{\epsilon_3} \right\|_\infty \leq 1 \quad (2)$$

which is equivalent to DBSCAN predicate expressed in terms of the metric scaled by ϵ parameter,

$$N_\epsilon(o, o') \equiv \left| \frac{o - o'}{\epsilon} \right| \leq 1 \quad (3)$$

Therefore, Tweet-SCAN can be seen as a particular case of DBSCAN which considers the ∞-norm of the scaled components as a metric function for the neighborhood predicate. This result enables to determine ϵ_i parameters through similar heuristics defined for DBSCAN, as we will show in Section 5.2.

3.3. Tweet-SCAN MinWeight predicate

Tweet-SCAN seeks to group closely related tweets generated by a diverse set of users instead of a reduced set of them. User diversity is imposed to avoid that a single user continuously posting tweets from nearby locations could generate an event in Tweet-SCAN. Forcing a certain level of user diversity within a cluster can be achieved through two conditions in the *MinWeight* predicate that must be satisfied at the same time,

$$MinWeight(o) \equiv |N_{NPred}(o)| \geq MinPts,\ UDiv(N_{NPred}(o)) \geq \mu \quad (4)$$

[5]Path distance could be more appropriate in the spatial dimension, but it would add more computational cost.

where $N_{NPred}(o)$ is the set of neighboring tweets of o such that $\{o' \in D \mid NPred(o, o')\}$ w.r.t. the previously defined Tweet-SCAN neighborhood predicate. The first condition from the MinWeight predicate establishes that neighboring tweets must have a minimum cardinality $MinPts$ as in DBSCAN. While in the second condition, the user diversity $UDiv()$ ratio, which is defined as the proportion of unique users within the set $N_{NPred}(o)$, must be higher than a given level μ of user diversity.

The combined predicate from equation (4) can be expressed as one single condition,

$$MinWeight(o) \equiv min\left(\frac{|N_{NPred}(o)|}{MinPts},\ \frac{UDiv(N_{NPred}(o))}{\mu} \right) \geq 1 \quad (5)$$

where $MinWeight(o)$ is now the minimum of two proportions which has to be greater or equal than 1. We note that the $MinWeight(o)$ predicate resembles the GDBSCAN one from Section 3.1 where $wCard()$ function corresponds to the minimum of both quotients and $MinCard$ is equal 1.

Note that if we set the user diversity level, μ, to 0 in equation (5), the MinWeight predicate simplifies to $|N_{NPred}(o)| \geq MinPts$ and this expression matches to the classical DBSCAN predicate. Under these conditions, we can use the same heuristics defined for DBSCAN to setup the $MinPts$ parameter as we will show in Section 5.2. Regarding μ, we propose to set empirically its proper value in Section 5.5 so that it maximizes the detection accuracy.

3.4. Tweet-SCAN text model

The text message in a tweet is a 140-character-long field in which users type freely their thoughts, experiences or conversations. The fact that users tweet in different languages, argots and styles dramatically increases the size of the vocabulary, making the use of simple Bag of Words (BoW) models (Salton et al., 1975) not viable. Therefore, we propose to use probabilistic topic models, which are common dimensionality reduction tools in text corpora (Blei, 2012). In this approach, a tweet message is encoded into a K-dimensional vector which corresponds to the Categorical probability distribution over the K topics. K is often much smaller than the vocabulary size and the resulting topics are represented by semantically similar words.

Nonparametric Bayesian models like Hierarchical Dirichlet Process (HDP) (Teh et al., 2006) can automatically infer the number of topics K, overcoming the limitation of their parametric counterparts like Latent Dirichlet Allocation (LDA) (Blei et al., 2003). The HDP topic model basically consists of two nested Dirichlet Process: G_o, with base distribution H and concentration parameter γ, and G_i, with base distribution G_o and concentration parameter α_o. Although the number of topics is automatically inferred, the hyperparameters γ and α_o might strongly influence the number of components. Because of this, vague informative gamma priors such as, $\gamma \sim Gamma(1, 0.1)$ and $\alpha_o \sim Gamma(1, 1)$ are usually considered (Escobar and West, 1995; Teh et al., 2006).

The straightforward use of HDP models on raw tweets does not provide meaningful topic distributions (Hong and Davison, 2010) due to the lack of word co-occurrence in short texts like tweets. Because of this, we propose the scheme from

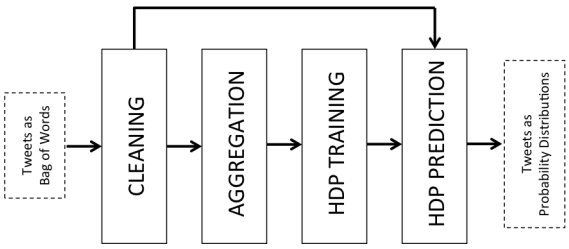


Fig. 2: Text model scheme. Stages are highlighted in bold in the text.

figure 2 which aims to alleviate these shortcomings. First, raw tweets, modeled as Bag of Words, are pre-processed and **cleaned** through classical data cleaning techniques from Natural Language Processing (NLP): lowering case, removing numbers and special characters, and stripping white-spaces. Then, processed tweets are **aggregated** to build longer training documents from a group of concatenated tweets. These aggregated documents are used to **train** the HDP model. Finally, the trained HDP model is employed to **predict** the topic distributions per each single tweet in order to obtain the Categorical probability distributions over the K topics that summarize each tweet message.

In the aggregation stage from figure 2, we consider two different strategies, although we will only use the first unless we say the contrary.

- By *hashtags*: it consists in creating a new training document per *hashtag*, which will concatenate all tweets that contains it. Therefore, there will be as many training documents as *hashtags*. One drawback of this aggregation approach is that tweets which do not have hashtags are aggregated together, although they might refer to completely different themes.

- By top keywords: it consist in first identifying a set of top keywords through the TF-IDF statistic (Salton and Buckley, 1988), and then aggregating tweets containing each of these top keywords. Thus, there will be as many training documents as top keywords and few tweets will be unassigned as long as we choose a reasonable number of top keywords.

Finally, we propose to use the Jensen-Shannon (JS) distance for the textual component in Tweet-SCAN neighborhood predicate. JS is a proper distance metric for probability distributions (Endres and Schindelin, 2003). It is defined as,

$$JS(p,q) = \sqrt{\frac{1}{2}D_{KL}(p||m) + \frac{1}{2}D_{KL}(q||m)} \tag{6}$$

where p, q and m are probability distributions and $D_{KL}(p||m)$ is the Kullback-Leibler divergence between probability distribution p and m written as,

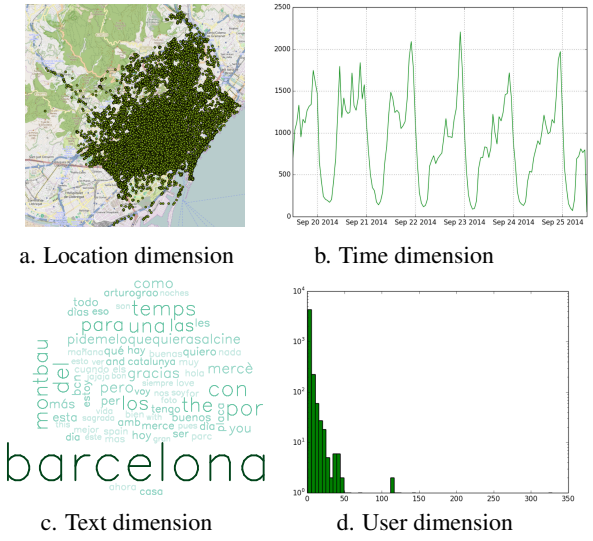
$$D_{KL}(p||m) = \sum_i p(i) log_2 \frac{p(i)}{m(i)} \quad m = \frac{1}{2}(p+q) \tag{7}$$

where m is the average of both distributions.

In Tweet-SCAN, p and q from equation (6) are two Categorical probability distributions over topics which are associated to two tweet messages. Given that Jensen-Shannon distance is defined through base 2 logarithms, JS distance will output a real value within the $[0, 1]$. Documents with the similar topic distribution will have a Jensen-Shannon distance close to 0 and those topic distributions which are very far apart, distance will tend to 1.

4. "La Mercè" data sets

In order to evaluate Tweet-SCAN, we have collected data through the Twitter streaming API[6] via Hermes (Cea et al., 2014). In particular, we have established a long standing connection to Twitter public stream which filters all tweets geolocated within the bounding box of Barcelona city[7]. This long standing connection was established during the local festivities of "La Mercè", that took place during few days in September 2014 and 2015[8].



a. Location dimension b. Time dimension

c. Text dimension d. User dimension

Fig. 3: Tweets dimensions from "La Mercè" 2014.

"La Mercè" festivities bring with several social, cultural and political events that happen in different locations within a considerably short period of time. This scenario is a suitable test bed for evaluating the accuracy of Tweet-SCAN on discovering these physical events from tweets. Moreover, the abundance of events during these days causes that some of them overlap in time and space, making text more relevant to distinguish them. However, these events are apparently not distinguishable by analyzing tweet dimensions separately as shown in figure 3, where event patterns are not visible. Figure 3a shows the spatial distribution of tweets within the borders of Barcelona city, where different tweet density levels can be appreciated in the map. Figure 3b represents the time series of tweets from the

[6] http://dev.twitter.com/streaming/overview

[7] Barcelona has 1,602,386 citizens in the 101.4 km^2 sized municipality according to Idescat

[8] Dataset published in https://github.com/jcapde87/Twitter-DS

19th to the 25th of September and daily cycles are recognizable. Figure 3c is a wordcloud in which more frequent words are drawn with larger font size, such as "Barcelona". The multilingualism at Twitter is also reflected at this wordcloud although this work does not considered translating between different languages. Last, figure 3d is a histogram of the number of tweets per user, which shows that most of the users tweet very few times, while there are a few, although non-negligible number of users, who tweet very often. All four dimensions play a key role in Tweet-SCAN to uncover events.

Table 1: "La Mercè" local festivities data sets

	Tweets	Tagged tweets	Tagged events
"La Mercè" 2014	43.572	511	14
"La Mercè" 2015	12.159	476	15

As shown in Table 1, we have also manually tagged several tweets with the corresponding events as per the agenda in "La Mercè" website[9] and our own expert knowledge as citizens. With this tagged subset of tweets, we will experimentally evaluate the goodness of Tweet-SCAN. We also note that the number of tweets collected in 2015 is much less than in 2014. This is because Twitter released new smart-phone apps in April 2015 for Android and IOS that enable to attach a location to a tweet (such as a city or place of interest) apart from the precise coordinates[10]. Since tweets generated during "La Mercè" 2014 data set did not contain this functionality, we only consider tweets whose location is specified through precise coordinates for "La Mercè" 2015 data set (12.159 tweets).

5. Tweet-SCAN assessment

In this section, we assess Tweet-SCAN for the task of event discovery in "La Mercè". Particularly, we aim to find evidence that proves the benefits of considering the textual component. We also seek to give insights on the role of each parameter, as well as to provide a scheme for determining them.

As a result, we first introduce some clustering measures to evaluate Tweet-SCAN performance against the tagged data set. Then, we present a heuristic to determine its spatio-temporal parameters (ϵ_1, ϵ_2) following intrinsic measures. With these parameters fixed, we analyze the textual component and the role of user diversity level through extrinsic clustering evaluation. Finally, we also assess its clustering performance when jointly varying the spatio-temporal and textual parameters.

5.1. Clustering extrinsic measures

Clustering evaluation metrics can be applied here since event discovery was defined to look for groups of tweets which are close in time, space and textual meaning. To evaluate clustering results against a tagged data set or *gold standard* is known as extrinsic cluster evaluation, in contrast to intrinsic evaluation, which is based on the closeness/farness of objects from the same/different clusters. Among extrinsic measures, we find out that purity, inverse purity and, specially, the combined F-measure have been extensively used for event discovery (Yang et al., 1998).

Purity and inverse purity are weighted averages of maximum precision and recall across clusters and labeled events, respectively. Because of that, both measures achieve trivial maximums when each object is set to a different cluster or when all tweets are grouped into a unique cluster. Van Rijsbergen (1974) combined both measures through the harmonic mean into the Van Rijsbergen's F-measure to mitigate the undesired trivial solutions from purity and inverse purity.

The F-measure score is defined as,

$$F = \sum_i \frac{|\,L_i\,|}{N} max_j\, 2 \cdot \frac{Rec(C_j, L_i) \cdot Prec(C_j, L_i)}{Rec(C_j, L_i) + Prec(C_j, L_i)} \tag{8}$$

where L_i is the set of tweets labeled as event i and C_j is the set of tweets clustered as j and N is the total number of tweets. Recall and precision are defined over these sets as the proportions $Rec(C_j, L_i) = \frac{|C_j \cap L_i|}{|L_i|}$ and $Prec(C_j, L_i) = \frac{|C_j \cap L_i|}{|C_j|}$.

5.2. Determining spatio-temporal parameters

Ester et al. (1996) proposed a rather simple but effective heuristic approach to determine DBSCAN parameters. However, Sander et al. (1998) argued that setting GDBSCAN parameters is extremely application dependent and the same heuristic could not necessarily apply there. As we have shown, Tweet-SCAN predicates matches those from DBSCAN under certain conditions and hence, the heuristic can be reused here.

The heuristic for DBSCAN is based on the following procedure. First, we define the distance of the ($MinPts$-1)-th nearest neighbor object from p as $d_{MinPts-1}(p)$. Note that within the neighborhood of p at distance $d_{MinPts-1}(p)$, there are at least $MinPts$ objects. By computing $d_{MinPts-1}(p)$ for each object in the data set and ordering them in descending order, we obtain a graph of sorted distances. From this graph, we can visually locate the point that determines the first "valley", whose $d_{MinPts-1}(p)$ corresponds to the proper ϵ parameter for a given $MinPts$. All objects p' with shorter $d_{MinPts-1}(p')$ distances than ϵ will be grouped into clusters, while those with larger distances, into noise.

We have previously shown that Tweet-SCAN neighborhood predicate is equivalent to DBSCAN predicate. Moreover, we have similarly highlighted that Tweet-SCAN MinWeight predicate matches DBSCAN for a null user diversity level. Consequently, the heuristic proposed for setting parameters in DBSCAN applies also here when $\mu = 0$.

The $d_{MinPts-1}(p)$ distance from tweet p for the spatio-temporal dimensions can be rewritten in terms of the ∞-norm as follows,

$$d_{MinPts-1}(p) = \left\| \frac{|p_1 - p'_1|}{\epsilon_1}, \frac{|p_2 - p'_2|}{\epsilon_2} \right\|_\infty = \left\| \frac{|p_1 - p'_1|, \epsilon_1/\epsilon_2 \cdot |p_2 - p'_2|}{\epsilon_1} \right\|_\infty \tag{9}$$

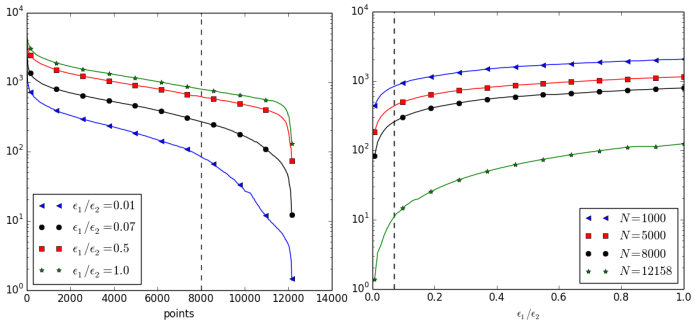
where p' is the ($MinPts$-1)-th nearest neighbor object from p and ϵ_1 can be factorized to obtain a matching expression with DBSCAN in which ϵ_1 plays now the role of ϵ. Through this

[9] http://lameva.barcelona.cat/merce/en/

[10] https://support.twitter.com/articles/78525

formula, we can now compute the $d_{MinPts-1}(p)$ for each tweet in order to get ϵ_1 parameter for different ϵ_1/ϵ_2 values. Note that this distance could be also extended to consider the textual component.

We now compute the $d_{MinPts-1}(p)$ for each tweet in "La Mercè" 2015 for different ϵ_1/ϵ_2 values ranging from 0.01(m/s) to 1(m/s), which seems a reasonable proportion range for events happening in Barcelona city. We also impose $MinPts = 10$, which means that there will be at least 10 tweets per event. This value makes sense for the type of events in "La Mercè" data set. Moreover, the change in the $d_{MinPts-1}(p)$ distance curves for Tweet-SCAN should be within short distance range, in contrast to DBSCAN, given that event discovery looks for rarely patterns in the data and thus, noise predominates over clusters.



a. Threshold point ≈ 8000 b. Threshold proportion $\epsilon_1/\epsilon_2 \approx 0.07$

Fig. 4: Sorted $d_{MinPts-1}(p)$ distances in "La Mercè" 2015

Figure 4a plots the ordered $d_{MinPts-1}(p)$ distance for each tweet in "La Mercè" 2015 and for different values of ϵ_1/ϵ_2. We visually locate the change in curvature in the short distance range around the 8000th point for all setups of ϵ_1/ϵ_2. From figure 4b, we can also visually identify the smallest ϵ_1/ϵ_2 value for which the $d_{MinPts-1}(p)$ distance slightly varies as $\epsilon_1/\epsilon_2 \approx 0.07$. With this ϵ_1/ϵ_2 proportion, we can determine $\epsilon_1 \approx 250m$ from the corresponding $d_{MinPts-1}(p)$ distance in figure 4a and $\epsilon_2 \approx 3600s$ from the ϵ's proportion out of figure 4b.

5.3. Understanding textual component

Next, we aim to find out which textual parameter ϵ_3 optimizes the F-measure for a fix set of parameters $\epsilon_1 = 250m$, $\epsilon_2 = 3600s$, $MinPts = 10$ determined as per the above heuristic. To do that, we plot the F-measure score as a function of ϵ_3 for "La Mercè" 2014 and 2015, and identify the ϵ_3 that maximizes F-measure.

From figure 5, it is clear that F-measure achieves a global maximum when ϵ_3 is within the range 0.8-0.9 for both data sets. Given that the maximum is achieved for $\epsilon_3 < 1$, the textual component clearly improves clustering performance. If the maximum would have been achieved at $\epsilon_3 = 1$, this would mean that the optimal Tweet-SCAN setup disregards the textual component. Besides, Tweet-SCAN performs slightly better in "La Mercè" 2015 than in 2014, what can be explained from the proportion of tagged tweets in "La Mercè" 2015 is greater than in 2014.

Table 2 shows F-measures per each tagged event in "La Mercè" 2014 and 2015. We observe that "fireworks" event has

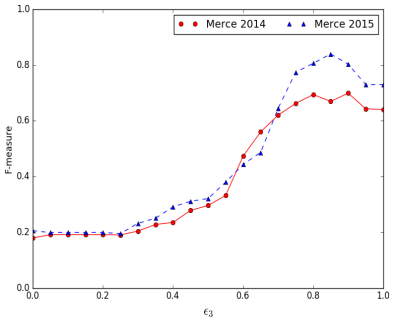
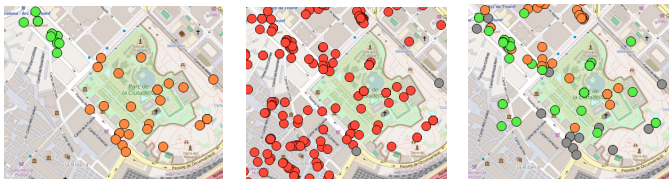


Fig. 5: F-measure as a function of ϵ_3 for $\epsilon_1 = 250m$, $\epsilon_2 = 3600s$, $MinPts = 10$, $\mu = 0.5$.

Table 2: F-measure per event for $\epsilon_1 = 250m$, $\epsilon_2 = 3600s$, $\epsilon_3 = 0.8$ $MinPts = 10$, $\mu = 0.5$

Event	"La Mercè" 2014	"La Mercè" 2015
Food market - day 1	0.41	0.93
Food market - day 2	0.27	0.94
Wine tasting - day 1	0.14	0.21
Wine tasting - day 2	0.33	0.34
Human towers - day 1	0.88	0.54
Human towers - day 2	-	0.74
Giants and Bigheads	-	0.34
Firerun	-	0.62
Fireworks	0.98	0.94
Projections	0.57	
MACBA concerts	0.45	-
Fabrica Damm concerts	0.97	-
Maria Cristina concerts	0.59	0.78
Bogatell concerts	0.96	0.84
OBC Sagrada Familia concert	-	0.85
CaixaForum conference	0.63	-
Atypl conference	0.40	-
Drupal conference	-	0.88
Referendum demonstration	0.94	-
Political meeting	-	0.27
Barça football game	-	0.99

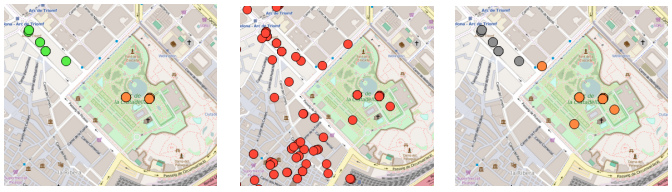
been successfully identified both years, whilst "wine tasting" has been poorly classified. The situation is that "fireworks" occurred during the closure of the festivities and happened in isolation of events. On the contrary, "wine tasting" took place near the "food market" event (Passeig Lluís Companys and Parc de la Ciutadella, respectively) during the same hours (all day events). Since both events were close in space and time, Tweet-SCAN has to use the textual component to distinguish among events.



a. Tagged Events b. $\epsilon_3 = 1$ c. $\epsilon_3 = 0.8$

Fig. 6: "Wine tasting" and "food market" events in "La Mercè" 2014

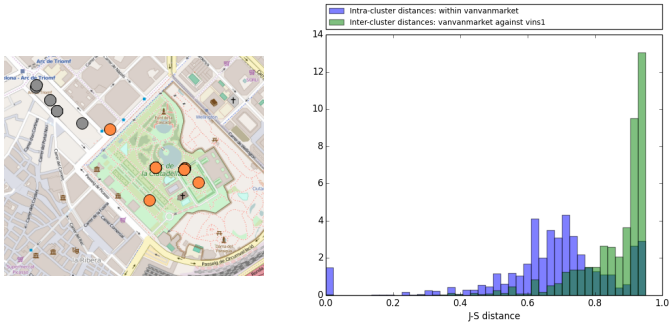
To show that, figure 6 and 7 plot the geo-located tweets in the surrounding area of both events during "La Mercè" 2014 and 2015. Both figures on the left show the tagged tweets associated with these events, where green dots belong to the "wine tasting" exposition and oranges, to the "food market". Figures



a. Tagged Events b. $\epsilon_3 = 1$ c. $\epsilon_3 = 0.8$

Fig. 7: "Wine tasting" and "food market" events in "La Mercè" 2015

on the middle plot the result of Tweet-SCAN without textual component $\epsilon_3 = 1$. It does not only merge both events into the same, but other tweets in the surrounding area are clustered all together. Figures on the right show the results of Tweet-SCAN for a $\epsilon_3 = 0.8$, which discovers both events in "La Mercè" 2014 and one in "La Mercè" 2015 despite mismatching several tweets and events. Note that grey dots represent tweets clustered as noise.

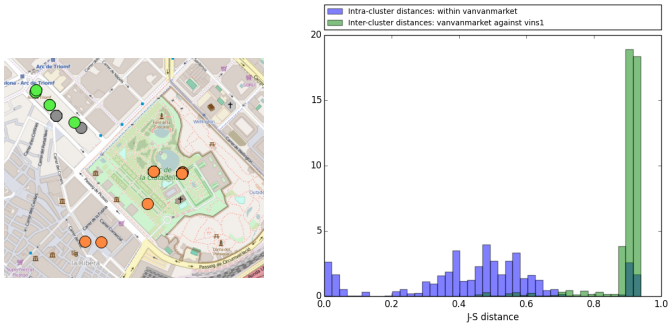


a. $\epsilon_3 = 0.8$ b. Inter- Intra- cluster distances

Fig. 8: Aggregation by hashtag in "La Mercè" 2015

Figure 8 shows Tweet-SCAN results for a $\epsilon_3 = 0.8$ (left) and the probability density function of inter- and intra- Jensen Shannon distances among the textual message of tweets in the "wine tasting" exposition and "food market" (right) in "La Mercè" 2015. Inter-class distances are among tweets from the same class, while intra-class distances, among tweets from different classes. Therefore, we look for minimum inter-class distances and maximum intra-class. Given that we simply consider the textual component, J-S distances plotted in figure 8b are bounded in the interval [0,1]. The probability density functions of these inter- and intra- distances show that tweets from both tagged classes overlap and the mean of intra-class distances is far from 0. As a result, we can state that text model performs poorly in discriminating these events.

In figure 9, we perform the same analysis by considering the HDP text model that aggregates by top keyword instead. As it can be seen from the left figure 9, we are now capable to correctly distinguish tweets related to the wine tasting event thanks to the fact that the probability distribution functions of inter- and intra- Jensen Shannon distance are less overlapped. This result encourages to aggregate by top keywords instead and to explore more robust tweet models with richer textual representation.



a. $\epsilon_3 = 0.8$ b. Inter- intra- cluster distances

Fig. 9: Aggregation by top keyword in "La Mercè" 2015

5.4. *Discovering untagged events*

The unsupervised Tweet-SCAN algorithm does not only discover tagged events, but it also identifies untagged events. Table 3 shows tweets grouped within an unknown event that took place during several hours in "La Mercè" 2014. From the text column, we can associate these tweets with "F.C. Barcelona" and we can also locate the event in the surroundings of "Camp Nou" stadium from their coordinates.

Moreover, we observe that it consists of several sub-events: a football game played by "Barça B vs. Llagostera", a "handball game" and "visits to Camp Nou stadium". Although all of them are in one way or another related to F.C. Barcelona, one might wish to obtain each sub-event by setting more restrictive Tweet-SCAN parameters. However, all three sub-events happen in the surrounding area of the stadium during similar hours. Here is where a more fine-grained text model for the textual component might make the difference to identify them.

From this event, we also note that "visits to Camp Nou stadium" are not strictly a sub-event, since they tend to happen quite often in time. To avoid identifying this type of clusters as events but as landmarks, Tweet-SCAN should consider inter-day/week patterns, which is out of the scope for this study.

5.5. *Analyzing user diversity level*

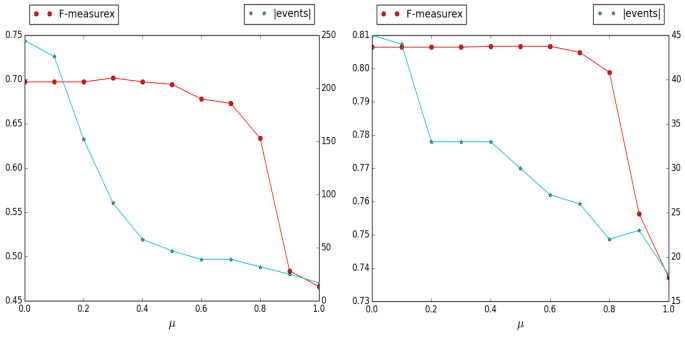
The role of user diversity level, μ, is two-fold. On the one hand, to guarantee that an event is created by a broad range of Twitter users; on the other hand, to filter out Twitter bots or accounts that does not reflect the spirit behind *user as a sensor* to report what's is happening nearby.

In what follows, we examine the effect of different user diversity levels μ to the clustering results in terms of F-measure. To do that, we fix all parameters ($\epsilon_1 = 250m$, $\epsilon_2 = 3600s$, $\epsilon_3 = 0.8$, $MinPts = 10$) except μ and we plot F-measure as function of the user diversity level. Lower user diversity level causes that few users could generate an event in Tweet-SCAN, while higher values will entail that events are generated by many different users. Since different μ values influences the number of discovered clusters by Tweet-SCAN, we will also add the number of events into the figure.

Figure 10 plots the F-measure and number of clusters as a function of μ for both data sets. It is clear from the figures that F-measure starts decreasing after a level of μ around 0.6. Similarly, the number of discovered clusters decreases but much

Table 3: 10 tweets from an untagged event during "La Mercè" 2014 for $\epsilon_1 = 250m$, $\epsilon_2 = 3600s$, $\epsilon_3 = 0.8$, $MinPts = 10$, $\mu = 0.5$

timestamp	coordinates	text
2014-09-20 15:11:56	2.123949 : 41.378228	@valds_jaycar a tope!! Vindras?
2014-09-20 15:57:19	2.121559 : 41.379579	Camp Nou http://t.co/ZjImGpOsuS
2014-09-20 17:48:04	2.118119 : 41.379819	Barça B-Llagostera (@ Mini Estadi in Barcelona, CT) https://t.co/cOfuwJnCon
2014-09-20 17:49:57	2.118119 : 41.379819	La Masía en vena (@ Mini Estadi for Barcelona B vs Llagostera in Barcelona, CT w/ @_muzki_) https://t.co/hoM1Pap45m
2014-09-20 18:49:03	2.120434 : 41.38028	@BMGranollers gran resultat noies!
2014-09-20 18:58:02	2.117992 : 41.379657	Llarga vida a la Masia!!! Visca el Bara des del Mini!! http://t.co/GSFCS8tryH
2014-09-20 19:32:31	2.120434 : 41.380264	Com apreta en Baena! #handbol http://t.co/etyKdVxBRV
2014-09-20 20:49:50	2.121351 : 41.380448	araeslhora fcbarcelona_es #campnou @ Camp Nou (FC Barcelona) http://t.co/c33qmi2Mx2
2014-09-20 21:01:55	2.121351 : 41.380448	#barcelona #fcb #mesqueunclub #campnou @ Camp Nou (FC Barcelona) http://t.co/nIk2kQ8WDF
2014-09-20 21:05:50	2.121351 : 41.380448	Tak for autografen 3jesper @ Camp Nou (FC Barcelona) http://t.co/AJe5tildAc



a. "La Mercè" 2014 b. "La Mercè" 2015

Fig. 10: F-measure for different μ values.

faster and sooner than F-measure. We observe that a user diversity level of 50% ($\mu = 0.5$) gives high figures of F-measure and reasonable number of events (≈50 events for "La Mercè" 2014 and ≈30 events for "La Mercè" 2015). Given that the size of "La Mercè" 2015 data set is four times smaller, make sense to obtain less number of events for the same μ level.

5.6. Analyzing spatio-temporal components

The aim of this section is to understand the impact of different neighborhood sizes through the spatio-temporal and textual parameters. Thus, we assess Tweet-SCAN in terms of F-measure scores when varying ϵ_1, ϵ_2 and ϵ_3. Figure 11 shows four possible ϵ_1, ϵ_2 configurations as function of ϵ_3 for both data sets.

A Tweet-SCAN setup with shorter distances in time and space ($\epsilon_1 = 250m$, $\epsilon_2 = 1800s$) than values obtained in Section 5.2, optimizes F-measure for $\epsilon_3 = 1$. This means that Tweet-SCAN disregards the textual component and it can be explained by the fact that these $\epsilon_1\epsilon_2$-neighborhoods are too restrictive for the tagged events.

For the spatio-temporal values determined in Section 5.2 ($\epsilon_1 = 250m$, $\epsilon_2 = 3600s$), we have seen that the optimum value for ϵ_3 is achieved within the range 0.8-0.9 in both data sets. Now, we can also see that this spatio-temporal configuration performs much better than others.

If we increase the spatial component to $\epsilon_1 = 500m$, but we keep the temporal short $\epsilon_2 = 1800s$, F-measure score is lower in both data sets, but the optimum value for ϵ_3 is attained within 0.8-0.9 in "La Mercè" 2014, and $\epsilon_3 = 1$ in "La Mercè" 2015. In fact, the curves for "La Mercè" 2015 are very similar to those given by $\epsilon_1 = 250m$, $\epsilon_2 = 1800s$.

Last, we increase both dimensions to $\epsilon_1 = 500m$ and to $\epsilon_2 = 3600s$. Although the optimum F-measure score for this setup is lower than the best performing configuration ($\epsilon_1 = 250m$, $\epsilon_2 = 3600s$), we observe that the textual component becomes more relevant. This is due to the fact that large $\epsilon_1\epsilon_2$-neighborhoods need textual discrimination to identify meaningful events.

6. Conclusions and future work

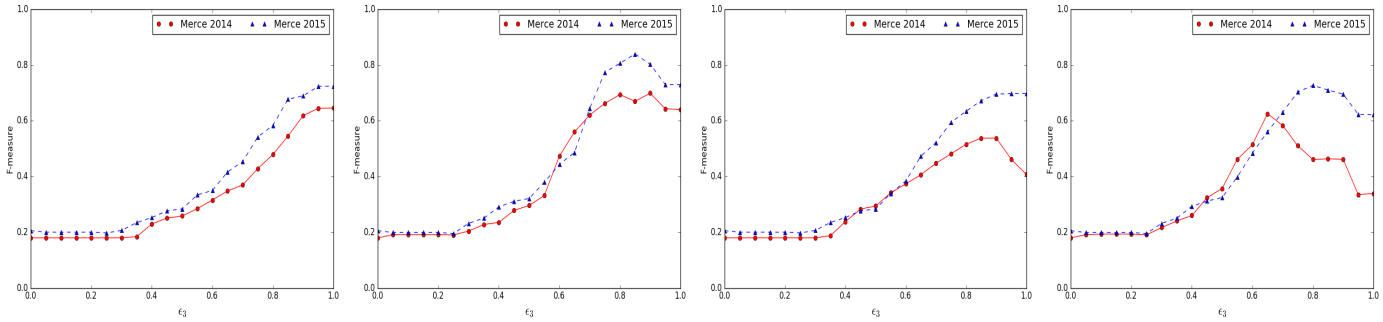
To our best knowledge, Tweet-SCAN provides a first step in using spatial, temporal, textual and user features for the purpose of uncovering real world events unsupervisedly from Location-based Social Networks like Twitter. The formulation of Tweet-SCAN within the framework of DBSCAN enables to understand events as density-connected set of tweets, as well as to use similar heuristics for determining some of the parameters.

The results of Tweet-SCAN points out to the benefits of using text, when uncovering events from geo-located tweets, specially for large spatial and temporal neighborhoods. We have also shown that more fine-grained text models could help to discriminate overlapping events in space and time, as we have seen for the "wine tasting" and "food market" events. Moreover, we have seen that imposing a user diversity condition does not worsen discovery performance, but it avoids clusters of non-event tweets posted by a very limited set of users.

Finally, we have observed that some clusters discovered by Tweet-SCAN were not exactly physical events, but popular places, a.k.a. landmarks, that usually have a huge attendance, e.g. "visits to Camp Nou stadium". To avoid this type of events, we should incorporate inter-day/week patterns to Tweet-SCAN, which will be part of future work. Not least, tuning up Tweet-SCAN parameters optimally for event detection tasks will require a cross-validation approach which maximizes F-measure in a training data set in such a way it then generalizes well in an independent test or validation data set.

Acknowledgments

This work is partially supported by Obra Social "la Caixa", by the Spanish Ministry of Economy and Competitivity under contract TIN2012-34557, by the BSC-CNS Severo Ochoa program (SEV-2011-00067), by the SGR programme (2014-SGR-1051) of the Catalan Government and by COR (TIN2012-38876-C02-01) project. We would like to also acknowledge reviewers for their constructive feedback.



a. $\epsilon_1 = 250m$, $\epsilon_2 = 1800s$ b. $\epsilon_1 = 250m$, $\epsilon_2 = 3600s$ c. $\epsilon_1 = 500m$, $\epsilon_2 = 1800s$ d. $\epsilon_1 = 500m$, $\epsilon_2 = 3600s$

Fig. 11: F-measure for different ϵ_1, ϵ_2, ϵ_3 and $MinPts = 10$, $\mu = 0.5$.

References

Blei, D.M., 2012. Probabilistic topic models. Communications of the ACM 55, 77–84.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. the Journal of machine Learning research 3, 993–1022.

Borge-Holthoefer, J., Rivero, A., García, I., Cauhé, E., Ferrer, A., Ferrer, D., Francos, D., Iñiguez, D., Pérez, M.P., Ruiz, G., et al., 2011. Structural and dynamical patterns on online social networks: the spanish may 15th movement as a case study. PloS one 6.

Cea, D., Nin, J., Tous, R., Torres, J., Ayguadé, E., 2014. Towards the cloudification of the social networks analytics, in: Modeling Decisions for Artificial Intelligence. Springer, pp. 192–203.

Chen, L., Roy, A., 2009. Event detection from flickr data through wavelet-based spatial analysis, in: Proceedings of the 18th ACM conference on Information and knowledge management, ACM. pp. 523–532.

Endres, D.M., Schindelin, J.E., 2003. A new metric for probability distributions. IEEE Transactions on Information theory .

Escobar, M.D., West, M., 1995. Bayesian density estimation and inference using mixtures. Journal of the american statistical association 90, 577–588.

Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise., in: Kdd, pp. 226–231.

Fung, G.P.C., Yu, J.X., Yu, P.S., Lu, H., 2005. Parameter free bursty events detection in text streams, in: Proceedings of the 31st international conference on Very large data bases, VLDB Endowment. pp. 181–192.

Garcia-Gasulla, D., Alvarez-Napagao, S., Tejeda-Gómez, A., Oliva-Felipe, L., Vázquez-Salceda, J., Gómez-Sebastià, I., Bejar, J., 2014. Social network data analysis for event detection, in: 21st European Conference on Artificial Intelligence (ECAI2014), IOS Press. pp. 1009–1010.

Hong, L., Davison, B.D., 2010. Empirical study of topic modeling in twitter, in: Proceedings of the First Workshop on Social Media Analytics, ACM. pp. 80–88.

Kim, E.K., Seok, J.H., Oh, J.S., Lee, H.W., Kim, K.H., 2013. Use of hangeul twitter to track and predict human influenza infection. PloS one 8, e69305.

Li, X., Liu, H., Jia, H., Huang, L., 2011. Research on the categorization accuracy of different similarity measures on chinese texts, in: Business Management and Electronic Information (BMEI), 2011 International Conference on, IEEE. pp. 224–227.

Ljubešić, N., Boras, D., Bakarić, N., Njavro, J., 2008. Comparing measures of semantic similarity, in: Lužar-Stiffler, V., Hljuz Dobrić, V., Bekić, Z. (Eds.), Proceedings of the 30th International Conference on Information Technology Interfaces, SRCE University Computing Centre, Zagreb. pp. 675–682.

McInerney, J., Blei, D.M., 2014. Discovering newsworthy tweets with a geographical topic model. NewsKDD: Data Science for News Publishing workshop. Workshop in conjunction with KDD2014 the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining .

Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M., 2013. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. arXiv preprint arXiv:1306.5204 .

Sakaki, T., Okazaki, M., Matsuo, Y., 2010. Earthquake shakes twitter users: real-time event detection by social sensors, in: Proceedings of the 19th international conference on World wide web, ACM. pp. 851–860.

Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. 24, 513–523.

Salton, G., Wong, A., Yang, C.S., 1975. A vector space model for automatic indexing. Communications of the ACM 18, 613–620.

Sander, J., Ester, M., Kriegel, H.P., Xu, X., 1998. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. Data Mining and Knowledge Discovery 2, 169–194.

Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M., 2006. Hierarchical dirichlet processes. Journal of the american statistical association 101.

Van Rijsbergen, C.J., 1974. Foundation of evaluation. Journal of Documentation 30, 365–373.

Weidemann, C., 2013. Social media location intelligence: The next privacy battle-an arcgis add-in and analysis of geospatial data collected from twitter. com. International Journal of Geoinformatics 9.

Weng, J., Lee, B.S., 2011. Event detection in twitter. ICWSM 11, 401–408.

Yang, Y., Pierce, T., Carbonell, J., 1998. A study of retrospective and online event detection, in: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM. pp. 28–36.

Yuan, Q., Cong, G., Ma, Z., Sun, A., Thalmann, N.M., 2013. Who, where, when and what: Discover spatio-temporal topics for twitter users, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA. pp. 605–613.

Zhang, L., Sun, X., Zhuge, H., 2013. Location-driven geographical topic discovery, in: Semantics, Knowledge and Grids (SKG), 2013 Ninth International Conference on, IEEE. pp. 210–213.

Zheng, Y., 2012. Tutorial on location-based social networks, in: Proceedings of the 21st international conference on World wide web, WWW, ACM.