Twitter data laid almost bare: An insightful exploratory analyser

Twitter data laid almost bare: an insightful exploratory analyser

Xin Xiao[a], Antonio Attanasio[a], Silvia Chiusano[a,*], Tania Cerquitelli[a]

[a] *Control and Computer Engineering Department, Politecnico di Torino, Corso Duca degli Abruzzi, 24 - 10129 Torino, Italy*

Abstract

In today's world, social networks and online communities continuously generate tons of data that reflect users' habits, personal interests, opinions and emotions. However, little profit can be gained from such huge raw data collections unless we are able to translate them into useful knowledge. Twitter, currently the leading microblogging social network, has attracted a great body of research works. Indeed, the rather heterogeneous dimensions characterizing Twitter data, such as space, time and text content, impose innovative methods in the data mining discovery process.

This paper presents TCharM, a data analytics methodology based on clustering and pattern discovery, to gain interesting knowledge from large complex collections of tweets. Cluster analysis is driven by a novel combined distance measure, named TASTE, to group tweets according to their spatio-temporal features and text content. In TASTE, the contributions of temporal and spatial distances are parametric and grounded on exponential proportionality. Each computed cluster is then locally characterized through association rules to ease the inspection of its Twitter messages. A categorization of rules into a few reference classes and topics is also proposed. TCharM exploits the computational advantages of distributed computing frameworks, as the current implementation runs on Apache Spark. The experimental evaluation performed on real datasets demonstrates the effectiveness of the proposed approach in discovering cohesive clusters and actionable

*Corresponding author. Tel.: +39 011 0907176.

Email addresses: `xin.xiao@polito.it` (Xin Xiao), `antonio.attanasio@polito.it` (Antonio Attanasio), `silvia.chiusano@polito.it` (Silvia Chiusano), `tania.cerquitelli@polito.it` (Tania Cerquitelli)

knowledge from Twitter data.

Keywords: Cluster analysis, Association rules, Text-spatio-temporal distance, Tweets, Social networks, Apache Spark

1. Introduction

Microblogs like Twitter have recently become a popular platform with millions of users and an impressive flow of messages (tweets) are published daily and spread by exchanges among users. The conciseness of their text messages (up to 140 characters) allows a very large number of tweets to be published at extremely low cost, thus making Twitter a timely and fresh source of data. Tweets can also be enriched with additional information describing their spatio-temporal publication context, such as when it was posted and the geographical location of the user.

The collection of tweets provides useful information to help understand peoples opinions and preferences on different *topics*, how peoples interests are spread *across geographical areas* and how they evolve *over time*. This better understanding of the collective dynamics of user interests can play a significant role in devising the most appropriate strategies and effective actions in various domains. From a business perspective, analyzing the trends of topics like sports, movies, and/or fashion, in different areas and time periods can help companies improve their services/products, the distribution of products as well as the planning of targeted promotional campaigns for specific services/products. In the Internet for instance, the analysis of social dynamics in different geographical areas helps characterize and predict the demand and supply of specific goods (Ikeda et al., 2013). On the other hand, policy makers can exploit microblogs in order to better understand peoples opinions regarding highly debated topics such as transport networks, taxes, healthcare systems, and public safety in different urban, regional or country areas and over time. The hidden knowledge in user messages allows policy makers to identify significant problems and devise targeted actions as well as evaluate how citizens perceive their effectiveness.

Although a large body of research focused on Twitter data analysis has already been proposed (e.g., (?Phelan et al., 2009; Steiger et al., 2016)), the potential impact of mining social data is still largely unexplored because various critical issues are yet to be addressed when analyzing tons of tweets to identify insightful nuggets. (i) Since a large number of tweets are continuously being posted worldwide, the size of tweet collections to be explored grows at an ever increasing rate. (ii) The collection of tweets generally tends to be scattered in spatio-temporal dimensions, and the conciseness of the tweet messages increases the brevity of their textual content (iii) Furthermore, the distribution of tweets can be characterized by different spatial and temporal granularities. (iv) Mined knowledge should be represented using concise and understandable patterns to enable its exploitation by domain experts. Thus, innovative data analytics solutions are needed to effectively and efficiently mine large Twitter data collections.

In this work we propose a novel exploratory analyser which enables end-users to gather insightful information, including a spatio-temporal-text viewpoint from tweet messages. Our data analytics methodology, named Tweets CHARacterization Methodology (TCHARM), explores large collections of Twitter data along the three dimensions characterizing tweets (i.e., text content, posting time and place) to support context-aware topic trend analysis.

TCHARM is based on two exploratory data mining techniques: (a) *Cluster analysis*, to identify cohesive groups of tweets with similar text content posted from nearby geographical areas and at close time instances, and (b) *Association rule analysis*, to find significant patterns that concisely describe each computed cluster. To make the proposed methodology scale up to larger datasets, TCHARM exploits the computational advantages of distributed computing frameworks since the current implementation runs on Apache Spark (Zaharia et al., 2010).

Unlike previous works (e.g.,(Kim et al., 2011; Lee, 2012; Cunha et al., 2014; Arcaini et al., 2016)), TCharM drives the clustering process by making joint use of the tweet spatio-temporal features and text content. A novel Text And Spatio-TEmporal distance measure, denoted by TASTE, is proposed in this study in order to combine the contributions of all three tweet features in one step. Through TASTE, spatial and temporal distances between tweets are used to modulate the text content distance. By taking into account both spatio-temporal features and text content in the clustering of tweets, TCHARM findings can provide useful insights to identify the users topics of interest in different areas and time periods. For instance, events such as sports, culture and politics, which have widespread visibility, can be useful to understand topics that are popular in different geographical areas. The information provided by the spatio-temporal distribution of such clusters may help characterize peoples involvement in different time frames. TCHARM has been currently integrated into the *K-means* clustering algorithm (Pang-Ning T. and Steinbach M. and Kumar V., 2006), to generate clusters of tweets that can be concisely represented by their centroids.

TCHARM then locally investigates each computed cluster to mine significant patterns which reveal underlying correlations among frequent topics, tweeting times and places that simultaneously emerge from cluster analysis.

This task has been carried out using association rule analysis (Pang-Ning T. and Steinbach M. and Kumar V., 2006), an exploratory data mining technique to extract correlations among data items. Quality indices (e.g., confidence, support, and lift) are used to distinguish the most significant correlations. Association rule analysis allows the extraction of the most recurrent spatio-temporal-text patterns in a systematic and structured way. These patterns describe the cluster content using a concise and clear knowledge representation.To further support the exploration of discovered patterns, four different classes of association rules have been defined. By "class" we mean a subset of patterns which determines significant relationships between tweet dimensions which can be used to perform a similar in-depth analysis. The identified patterns can provide domain experts with valuable support to identify which topics are most appealing to users in different areas and time periods.

It is worth mentioning that our methodology can be exploited to support knowledge discovery in different contexts, and in this study TCharM has been thoroughly evaluated using the large number of tweets collected during the 2014 FIFA World Cup championship. This football competition was selected as a representative case because it included a variety of events (e.g., football matches with different teams, ceremonies, celebrities statements) spread over a set time period. Moreover, as it is of worldwide interest, peoples interest in, and perceptions of, this kind of event may vary depending on their geographical location. The experimental evaluation demonstrates the effectiveness of TCharM in identifying interesting knowledge regarding the spatio-temporal distribution of peoples reactions to the events. The identified clusters provide useful findings regarding hot topics for users, in the different areas and time periods. Mined clusters are timely centered around an event related to the 2014 FIFA World Cup Championship and they mainly include messages about specific topics. Moreover, they show good spatio-temporal cohesion around their centroid.

The rest of the paper is organized as follows: Section 2 summarizes the related work regarding cluster analysis of Twitter data. Section 3 provides an in-depth description of the TCharM characteristics, while Section 4 discusses the experimental study conducted on the 2014 FIFA World Cup Championship dataset. Section 5 provides a theoretical and analytical comparison between TCharM and some previous works on tweet clustering. Section 6 discusses the significance of TCharM findings and their possible exploitation. Section 7 draws conclusions and future developments of the proposed approach.

2. Related Work

In the last few years the application of data mining techniques to discover relevant social knowledge from tweets collections has become an appealing research topic. Proposed approaches, mainly based on text processing and its extensions to heterogeneous data, can be classified into the following two main categories.

The first category refers to methods addressing the analysis of tweet textual content with the aim of (i) characterizing online communities (Rabiger & Spiliopoulou, 2015), (ii) performing spam detection (Thomas et al., 2011), (iii) detecting topics to analyse trends (Baralis et al., 2013; Vicient & Moreno, 2015; Yang & Rim, 2014), and (iv) addressing recommendation tasks (Phelan et al., 2009).

The second category includes methods considering spatio-temporal information in addition to tweet textual content. Different types of analysis have been addressed as (i) discovering regional social activities or nearby events using geo-tagged tweets (Kim et al., 2011), (ii) detecting events based on cluster analysis (Lee, 2012; Steiger et al., 2016), (iii) extracting insightful summaries of citizen perceptions from tweets (Bernabe-Moreno et al., 2015; Lee et al., 2015), (iv) discovering contrasting situations by means of generalized itemsets (Cagliero et al., 2014), (v) identifying the period in which a burst of information diffusion took place (Saito et al., 2015), and (vi) mining user opinions (Lloret et al., 2012).

Various approaches have been proposed to cluster tweets collections taking into account textual content and spatio-temporal information (Kim et al., 2011; Steiger et al., 2016), though such works do not jointly exploit all these features in the clustering process. Instead, they typically use a subset of features for clustering, while remaining features are considered either in the post-processing phase, for instance to refine or characterize discovered clusters, or in the preprocessing phase, for example to specify spatial or temporal segments in which tweets are locally clustered based on textual content. Kim et al. (2011) cluster tweets based on their GPS coordinates using the K-means algorithm, while Steiger et al. (2016) use a spatio-temporal clustering based on Self Organizing Maps (SOM). In both approaches, discovered clusters are then analysed to identify the main targeted topic. Density based clustering, mainly based on the DBSCAN algorithm, has been also adopted to detect high spatial concentrations or temporal bursts of tweets about specific topics (Arcaini et al., 2016; Lee, 2012; Lee et al., 2015; Sakai et al., 2015). For instance, Lee et al. (2015) group user trajectories derived from geo-tagged tweets and explore massive crowd movements, while Sakai et al. (2015) extract local bursty keywords and identify their dense areas to enhance local situation awareness.

Differently from all the works above, the TCharM **framework *jointly* exploits the spatio-temporal features and tweet textual content to drive the clustering process. Our main purpose is to discover cohesive clusters focused on single topics and, at the same time, with precise spatio-temporal references. Through the** TASTE **distance measure,** TCharM **explores the three dimensions characterizing tweets, to discover, in one step, groups of messages with similar content but posted in nearby time and space.**

As an additional contribution with respect to all the works mentioned above, TCharM performs a further step of clusters characterization through association rules extraction. The use of association rules to characterize clusters of tweets was proposed by Baralis et al. (2013). However, in TCharM rules are additionally categorized into few reference classes, according to their semantics, to ease the comprehension and exploitation of the extracted knowledge. Moreover, association rule analysis explores correlations not only in the textual content, but also between textual content and the time and location of tweet posting.

In this study, the TCharM framework has been deployed on Apache Spark. Several open source data mining platforms, like Scikit-learn, RapidMiner, Apache Mahout and Apache Spark have proposed their own scalability strategies to analyse the huge and rapidly growing amount of data. Such platforms include libraries implementing common machine learning algorithms which can be extended or modified by researchers. The adoption of Apache Spark in many research works (including but not limited to tweets) is mainly motivated by both the support for stream analysis (Dasgupta et al., 2015) and the scalable computing framework that makes it possible to speed up existing algorithms for different applications (Capdevila et al., 2016).

Tweets about the 2014 FIFA World Cup has been considered as a reference case study for the validation of the proposed framework. Various studies have addressed the analysis of tweets related to this event, with different targeted analyses devoted to (i) performing sentiment analysis to characterize U.S. soccer fans' emotional responses (Yu & Wang, 2015); (ii) addressing topic detection through a combined approach based on the DBSCAN algorithm and Non-Negative matrix (Godfrey et al., 2014); (iii) tracking user behavior through Latent Dirichlet Allocation (LDA) (Kim et al., 2015). All these approaches analyse the textual content only, while TCharM clusters the tweet collection besides characterizing the cluster content based on textual and spatio-temporal dimensions.

3. TCharM architecture

The main components of the Tweets CHARacterization Methodology (TCharM) architecture are shown in Figure 1. The components are briefly introduced below while a more thorough description of each of them is given in the following subsections.

The first activity is *data collection and preprocessing*. All information about tweets, including text content, publication time and user geographical location, are retrieved through the Twitter Stream Application Programming Interfaces (APIs) specifying a set of filter parameters (e.g., keywords, hashtags). The collected data then undergo a preprocessing phase to be represented in a format suitable for the subsequent clustering analysis. The adopted data model is described in Section 3.1. The output of the preprocessing is a dataset where each record corresponds to a single tweet and contains basically three features: *text content*, *time* of tweet posting and *location* of the user when posting the tweet.

Once the dataset is ready, the *cluster analysis* elaborates its records in order to partition the tweets collection into cohesive groups (clusters). For this activity, a novel combined distance measure, called Text And Spatio-TEmporal (TASTE), is used to cluster Twitter messages considering their spatio-temporal information and the text content as well.

Finally, TCharM analyses each discovered cluster to mine a set of patterns describing the cluster content. Specifically, through *association rule analysis*, patterns of relevant correlations among tweets text contents, posting times and geographical areas are extracted for each cluster. Extracted rules are then categorized into four classes defined according to the types of correlation among the tweets attributes while, to ease their semantic interpretation, the same rules are associated with one of the few reference topic families according to the word set they contain.

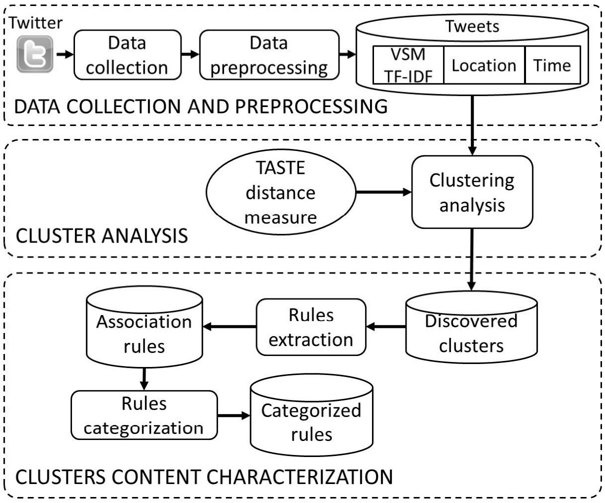


Figure 1: The TCharM architecture

3.1. Twitter data representation

This study aims at the characterization of groups of tweets with similar text and posted in close geographical areas and time instants. To support this analysis, the following three features have been considered for the representation of Twitter data: (i) *tweet text content* (ii) *tweet temporal feature*, i.e. tweet posting time, and (iii) *tweet spatial feature*, i.e., user geographical position at posting time.

Tweet text content. Tweets are posts published by Twitter users that include also *text messages* 140 characters long at most. Such messages represent the text content used in our analysis. Due to the limited size of the single message and to the high dimensionality of many text content representations, the represented samples are inherently sparse. This property leads to higher levels of noise in the tweet collection, thus adding complexity to the clustering process, which requires an adequate treatment (Jing et al., 2007). Moreover, Twitter messages are usually extremely impure because they include a wide variety of Unicode data, symbols, numbers and links. They therefore need to be properly cleaned and prepared before the analysis.

Tweet temporal feature corresponds to the *timestamp* including date and time instant when the tweet was posted. In this study, we omit the temporal information possibly appearing in the tweet message, since it is considered less relevant for discovering tweets posted in nearby time.

Text	England 2-0 I still believe
Time	Friday June 20 09:26:53 +0000 2014
Location (latitude, longitude)	52.076, -1.363

Table 1: Example tweet including text content and spatio-temporal features

Tweet spatial feature can be acquired as *geographical coordinates* of the user when she/he posted the tweet, with the location specified in the user profile, and location mentioned in the tweet text content. Geo-coordinates (i.e., latitude and longitude) are available when GPS enabled devices are used and localization is enabled. They specify the spatial position of people right when posting the tweet. Instead, the location reported in the user profile is free-text information provided by the same user. It usually corresponds to the place (such as city, state or country) where people come from. Similarly, locations mentioned in the tweet message do not necessarily correspond to the user position when the tweet was sent. Since our aim is to discover tweets with similar text content but posted in nearby geographical areas (and time periods), we focused mainly on the spatial information provided through geo-coordinates.

Table 1 reports an example tweet including the three features. The tweet refers to the 2014 FIFA World Cup, considered as a reference case study in this paper. The tweet was posted on Friday morning, June 20th 2014, at 9:26 a.m. from Banbury City (UK), according to geo-coordinate values.

In TCharM the tweets collection is represented as a dataset where each record corresponds to a single tweet and contains basically three attributes, corresponding to the three features above, i.e., tweet text content, and tweet temporal and spatial features. For the purposes of this study, the text content has been represented using the *Bag-of-Words* (BOW) model usually adopted in text mining (Steinbach et al., 2000). The message is represented as the multiset of its words, disregarding grammar and even word order, but keeping word multiplicity. A more formal definition of the adopted representation for tweet data is the following one.

Definition 3.1 (Tweet data representation). *Let $\mathcal{D}$ be a set of tweets and $\Sigma = \{w_1, \ldots, w_k\}$ the set of words appearing in at least one tweet in $\mathcal{D}$. An arbitrary tweet $\tau_i \in \mathcal{D}$ is represented as a triplet $\tau_i = (t_i, s_i, W_i)$ where t_i and s_i are respectively the temporal and spatial features of τ_i, while $W_i \subseteq \Sigma$*

is the tweet text content.

The *temporal feature* t_i is the *timestamp* indicating *when* tweet τ_i was posted, while the *spatial feature* s_i is the pair of *geo-coordinates* reporting from *where* tweet τ_i was posted. The *text content* W_i is given by the subset of words w_j ($w_j \in \Sigma$) appearing in tweet τ_i, with their respective frequencies.

Unweighted word frequencies do not properly characterize tweet text content, since words related to more specific events may appear with lower frequency than common words. In this study, the Term Frequency (TF) - Inverse Document Frequency (IDF) scheme (Manning et al., 2008), usually used in text mining, has been adopted to highlight the relevance of specific words for each tweet, while reducing the importance of common terms in the collection. The adoption of the TF-IDF scheme in the message representation makes it possible to focus the tweet matching in the subsequent clustering phase on words specific for each subset of tweets rather than on words common to most tweets. To weight word relevance based on the TF-IDF scheme, the tweet text content is transformed using the Vector Space Model (VSM) representation (Salton et al., 1975). Each tweet text content is a vector in the word space. Each vector element corresponds to a different word and is associated with the TF-IDF weight describing the word relevance for the tweet, as in the following Definition 3.2.

Definition 3.2 (Tweet text content representation). *Let* $\tau_i = (t_i, s_i, W_i)$ *be an arbitrary tweet in collection* $\mathcal{D}$. *The tweet text content* W_i *is a vector of* k *elements corresponding to words in* Σ *(i.e.,* $k = |\Sigma|$*). Each vector element* $W_i[j]$ *contains the TF-IDF weight of word* w_j *for tweet* τ_i. $W_i[j]$ *is computed as* $W_i[j] = TF(\tau_i, w_j) \cdot IDF(w_j)$, *where terms* $TF(\tau_i, w_j)$ *and* $IDF(w_j)$ *are defined as follows:*

1. $TF(\tau_i, w_j)$ *is the relative frequency of word* w_j *for tweet* τ_i. $TF(\tau_i, w_j) = f(\tau_i, w_j) / \sum_{l=1}^{k} f(\tau_i, w_l)$, *where* $f(\tau_i, w_j)$ *is the number of times word* w_j *appeared in tweet* τ_i *and* $\sum_{l=1}^{k} f(\tau_i, w_l)$ *is the total number of words contained in* τ_i.
2. $IDF(w_j)$ *is the relative frequency of word* w_j *in* $\mathcal{D}$. $IDF(w_j) = log(|\mathcal{D}|/|\mathcal{D}_j|)$ *where* $|\mathcal{D}|$ *is the number of tweets in* $\mathcal{D}$ *and* $|\mathcal{D}_j|$, $\mathcal{D}_j = \{\tau_i \in \mathcal{D} : f(\tau_i, w_j) > 0\} \subseteq \mathcal{D}$, *is the number of tweets in* $\mathcal{D}$ *which contain (at least once) word* w_j.

Mathematically, the base of the *log* function for IDF computation in Definition 3.2 does not influence the overall results as it constitutes a constant multiplicative factor (Robertson, 2004). The TF-IDF weight $W_i[j]$ for word w_j in tweet τ_i is high when w_j appears with high frequency in tweet τ_i but low frequency in tweets in the collection $\mathcal{D}$. When word w_j appears in more tweets, the ratio inside the IDF *log* function approaches 1, and both the IDF(w_j) value and the TF-IDF weight $W_i[j]$ become close to 0. Hence, the approach tends to filter out common words. In short messages like tweets, the TF-IDF weighting score could actually be reduced to a pure IDF scheme due to the limited word frequency within each tweet. Nevertheless, we preserved the TF-IDF approach to consider also possible word repetitions.

3.2. Twitter data collection and preprocessing

Twitter data for the TCharM framework are retrieved through the Twitter Stream Application Programming Interfaces (APIs) by specifying a set of filter parameters (e.g., keywords, hashtags). Collected data include all information characterizing tweets useful for the subsequent data analysis phase, i.e., tweet message, publication time and geographical location of the user. Of the tweets collected, only those in English are considered.

To enable the subsequent data analysis process on crawled tweets, the following data preparation steps are applied. Tweet messages are cleaned by removing numbers, usernames and URLs. After converting the letters into lowercase, messages are further cleaned by eliminating stop words (such as "is", "at", "the", etc.). Finally, the text content is represented using the data model described in Section 3.1, i.e., the BOW data model is applied and the TF-IDF score schema is used to weight word relevance.

3.3. Cluster analysis of tweets

Cluster analysis partitions objects into groups so that objects within the same group are more similar to each other than to the ones assigned to different groups. Different kinds of clustering algorithms are available, like partitional (e.g., K-means, K-medoids), density-based (e.g., DBSCAN), and hierarchical (e.g., agglomerative) (Pang-Ning T. and Steinbach M. and Kumar V., 2006).

In TCharM, the K-means algorithm is used for clustering tweet data collections. K-means has been widely used in different applications domains, including tweets analysis, providing good quality solutions. The K-means algorithm discovers K clusters modeled by their representatives, named *centroids*, given by the mean value of the objects in the clusters. Initially, K tweets of the tweet collection $\mathcal{D}$ are randomly chosen as centroids. Then each tweet $\tau_i \in \mathcal{D}$ is assigned to the cluster of the nearest centroid. Finally, the centroids are relocated by computing the mean of the tweets features within each cluster. The process iterates until a convergence criterion is met, i.e., the centroids do not change or some objective functions are achieved.

The K-means algorithm used in TCHARM exploits the novel distance measure TASTE to discover clusters with similar content but also posted in nearby geographical areas and close time periods. The TASTE measure takes into account the three tweet features at once to determine an overall distance between tweets.

3.3.1. The TASTE *distance measure*

The proposed Text And Spatio-TEmporal (TASTE) distance measure is formally defined as follows.

Definition 3.3 (TASTE **distance measure).** *Let $\tau_i = (t_i, s_i, W_i)$ and $\tau_j = (t_j, s_j, W_j)$ be two arbitrary tweets in collection $\mathcal{D}$. The* TASTE *distance measure between tweets τ_i and τ_j is defined as*

$$d_{TASTE}(\tau_i, \tau_j) = d_W(W_i, W_j) \cdot (k_s \cdot \mathrm{e}^{p_s \cdot d_s(s_i, s_j)} + k_t \cdot \mathrm{e}^{p_t \cdot d_t(t_i, t_j)}) \qquad (1)$$

where parameters $k_s, k_t, p_s, p_t \in \mathbb{R}$; $k_s, k_t \in [0, 1]$ and $k_s + k_t = 1$. Terms $d_W(W_i, W_j)$, $d_s(s_i, s_j)$, and $d_t(t_i, t_j)$ measure the distance on tweet text content, spatial feature, and temporal feature, respectively. These distances have been normalized in the range $[0, 1]$ using the *min-max* normalization method (Pang-Ning T. and Steinbach M. and Kumar V., 2006).

TASTE is defined as a measure of dissimilarity. Given tweets τ_i and τ_j, lower values of $d_{TASTE}(\tau_i, \tau_j)$ denote a higher similarity between τ_i and τ_j, while higher values of $d_{TASTE}(\tau_i, \tau_j)$ denote a lower similarity.

In the TASTE measure, spatial and temporal distances ($d_s(s_i, s_j)$ and $d_t(t_i, t_j)$) modulate the text content distance ($d_W(W_i, W_j)$) to determine the overall value of $d_{TASTE}(\tau_i, \tau_j)$. The exponential form is used for $d_s(s_i, s_j)$ and $d_t(t_i, t_j)$ to significantly penalize tweets with a large space and/or time distance.

The parameters of the TASTE measure can be conveniently tuned to fit scenarios with different spatial and temporal scales. Parameters k_s and k_t

weight the relevance of spatial and temporal distances in modulating the text content distance. Parameters p_s and p_t are included as exponents to adjust the (possibly differentiated) growth rates of exponential terms of spatial and temporal distances. For instance, to discover clusters of tweets with a high temporal cohesion, but possibly spread over a large geographical area, suitably higher values should be assigned to parameter p_t to penalize distances in time.

In TASTE, three different measures are used to compute $d_W(W_i, W_j)$, $d_s(s_i, s_j)$, and $d_t(t_i, t_j)$ based on the data type describing tweet text content, spatial feature and temporal feature.

Text content distance measure ($d_W(W_i, W_j)$). The distance between the weighted word frequency vectors W_i and W_j of tweets τ_i and τ_j is evaluated using the *cosine distance measure* (Pang-Ning T. and Steinbach M. and Kumar V., 2006), which has often been used to compare documents in text mining (Steinbach et al., 2000). We define the text content distance measure $d_W(W_i, W_j)$ as

$$d_W(W_i, W_j) = \arccos\left(cos(W_i, W_j)\right). \tag{2}$$

Term $cos(W_i, W_j)$ in Equation 2 represents the *cosine similarity* between W_i and W_j, and it is computed as

$$cos(W_i, W_j) = \frac{\sum\limits_{l=1}^{k} W_i[l] W_j[l]}{\sqrt{\sum\limits_{l=1}^{k} W_i[l]^2} \cdot \sqrt{\sum\limits_{l=1}^{k} W_j[l]^2}} \tag{3}$$

where k is the cardinality of the word set Σ in collection $\mathcal{D}$ ($k = |\Sigma|$).

The value range is $[0, 1]$ for the cosine similarity $cos(W_i, W_j)$, while the value range for the content distance measure $d_W(W_i, W_j)$ is $[0, \pi/2]$. When $cos(W_i, W_j) = 1$, then $d_W(W_i, W_j) = 0$ which describes the exact similarity of text content for tweets τ_i and τ_j. When $cos(W_i, W_j) = 0$, then $d_W(W_i, W_j) = \pi/2$ which points out that tweets τ_i and τ_j have completely different texts.

Temporal distance measure ($d_t(t_i, t_j)$). The tweet temporal feature is an integer number representing the time instant when the tweet was posted. The *Euclidean distance* (Pang-Ning T. and Steinbach M. and Kumar V., 2006) is adopted here as the distance on temporal features t_i and t_j of tweets

τ_i and τ_j. As t_i and t_j are expressed as time instants, the Euclidean distance is computed as the absolute value of their difference, i.e.,

$$d_t(t_i, t_j) = |t_i - t_j|. \tag{4}$$

Spatial distance measure ($d_s(s_i, s_j)$). Both Haversine and Euclidean distance measures have been used in other works to calculate the spatial distance between two geographical points (Lee, 2012). However, the Haversine distance is usually considered as more appropriate and precise especially when the distance between two points gets larger and it cannot be approximated as a straight line. For this reason, in this study the *Haversine distance* is used for computing the spatial distance between tweets. The Haversine distance corresponds to the great-circle distance between two points, i.e., their shortest distance over the earth's surface. Hence, the spatial distance between s_i and s_j for tweets τ_i and τ_j is computed as

$$d_s(s_i, s_j) = 2 \cdot R \cdot \arcsin(\sqrt{h}) \tag{5}$$

$$h = \sin^2(\Delta\varphi/2) + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \sin^2(\Delta\lambda/2) \tag{6}$$

where $\Delta\varphi$ and $\Delta\lambda$ are latitudinal and longitudinal differences between the tweets and R is a constant value equal to the Earth's mean radius (6,371 km).

The content, spatial and temporal distance measures defined above satisfy the positivity, symmetry, and triangle inequality properties that characterize a metric (Pang-Ning T. and Steinbach M. and Kumar V., 2006). It easily follows that the TASTE measure also verifies these properties. Specifically, the following properties hold. (i) *Positivity*: $d_{TASTE}(\tau_i, \tau_j) \geq 0$ for all $\tau_j, \tau_i \in \mathcal{D}$, while $d_{TASTE}(\tau_i, \tau_j) = 0$ only if $\tau_i = \tau_j$. (ii) *Symmetry*: $d_{TASTE}(\tau_i, \tau_j) = d_{TASTE}(\tau_j, \tau_i)$ for all $\tau_j, \tau_i \in \mathcal{D}$. (iii) *Triangle inequality*: $d_{TASTE}(\tau_i, \tau_j) \leq d_{TASTE}(\tau_i, \tau_k) + d_{TASTE}(\tau_k, \tau_j)$ for all $\tau_i, \tau_k, \tau_j \in \mathcal{D}$.

As an example, Figure 2 reports four sample tweets (τ_1 to τ_4) with their text content, temporal and spatial features. The values of d_{TASTE} between tweet τ_1 and the other tweets are also specified. Tweets are about the 2014 FIFA World Cup. Aimed at easing the comprehension of the results, the figure shows the original text messages, in place of the corresponding data model based on both BOW representation and TF-IDF score. It is worth noting that tweets τ_2 and τ_3 have a higher similarity with τ_1 than with τ_4. Tweets τ_1, τ_2 and τ_3 have a similar text content as they all talk about the

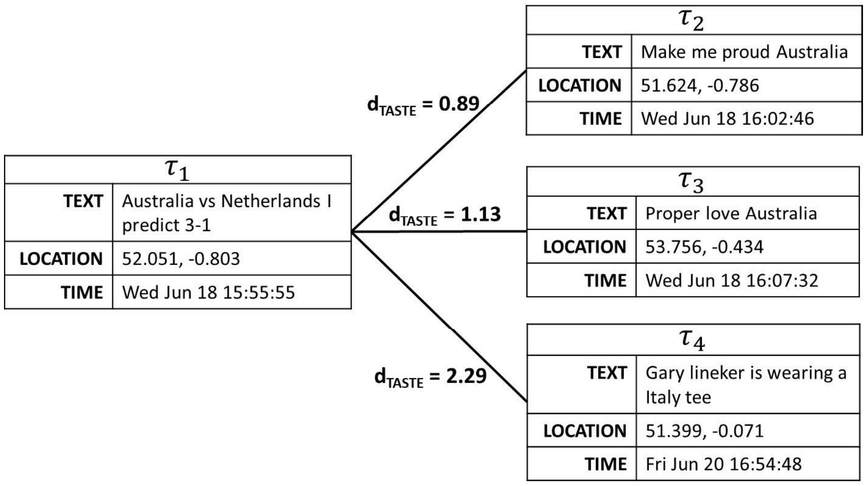


Figure 2: Sample tweets about 2014 FIFA World Cup with TASTE distance values

Australia football team. Tweets τ_2 and τ_3 were posted almost at the same time as τ_1, but τ_3 exhibits a farther geographical location from τ_1 than τ_2. This larger spatial distance penalizes the similarity on the text content and finally provides a higher value of d_{TASTE} for tweet τ_3. Conversely, tweet τ_4 exhibits a significantly higher TASTE distance from τ_1 even though it was posted in the neighbourhood, as τ_4 has a completely different content from τ_1 and it was posted two days later.

3.3.2. Clustering Evaluation

For the (internal) validation of clustering results, TCharM adopts the *Sum of Squared Errors* (*SSE*) quality index, usually adopted for evaluating the quality of a cluster set computed with the K-means algorithm (García-Gavilanes et al., 2014). The *SSE* index measures the cluster cohesion in prototype-based clusters, i.e., how objects in a cluster are closely related to the corresponding centroid. SSE is defined as the sum of the squared distances between each member of the cluster and its centroid. In TCharM, the SSE index is computed as

$$SSE = \sum_{i=1}^{K} \sum_{\tau_j \in C_i} d_{TASTE}(\tau_j, c_i)^2 \tag{7}$$

where c_i is the centroid of cluster C_i, and C_i is included in a cluster set with

K clusters. $d_{TASTE}(\tau_j, c_i)$ is the TASTE distance between a tweet $\tau_j \in C_i$ and the centroid c_i of C_i.

3.4. Clusters content characterization

After the cluster set is generated, in TCharM each cluster is then locally explored to characterize its content. Specifically, each cluster is analysed to discover underlying correlations in the text content, and between text content and the spatial and temporal features characterizing tweets. Cluster characterization makes use of *association rules* as reference pattern type (Agrawal et al., 1993).

3.4.1. Association rules extraction

Association rules analysis is an exploratory data mining technique to mine correlations among data items (Agrawal et al., 1993). To enable the association analysis process, tweets contained in the cluster under analysis are tailored to a transactional data format.

Consider an arbitrary cluster C included in the cluster set computed on tweet collection $\mathcal{D}$. The *transactional tweet dataset* $\mathcal{D}_{\mathcal{T}}(C)$ for cluster C is a set of transactions. Each *transaction* $\mathcal{T}_i$ corresponds to a tweet $\tau_i \in C$ and it consists of a set of tweet features called *items*, represented in the form $\{attribute : value\}$. The items of the generic transaction $\mathcal{T}_i$ are (i) each single *word* $w \in W_i$ appearing in the text content of tweet τ_i, (ii) the value of the *spatial feature* s_i of τ_i, and (iii) the value of the *temporal feature* t_i of τ_i.

An *association rule* is an implication in the form $r : X \Rightarrow Y$, where X and Y are disjoint *itemsets* (i.e., sets of items). X and Y are denoted as *rule antecedent* and *consequent*, respectively. Association rules extraction is commonly driven by rule support and confidence quality indexes. Whereas the *support* index represents the observed frequency of occurrence of rule r in the transactional dataset, the *confidence* index represents the rule strength. Consider the transactional tweet dataset $\mathcal{D}_{\mathcal{T}}(C)$ for cluster C; let $r : X \Rightarrow Y$ be a rule mined from $\mathcal{D}_{\mathcal{T}}(C)$. Rule support (*supp*) is the percentage of tweets in cluster C that contain both X and Y. Rule confidence ($conf$) is the percentage of tweets in cluster C containing X that also contain Y.

Consider, for example, association rule r : {*start, world, cup*} $\Rightarrow$ {*love*} (supp = 1.1%, conf = 60%) mined from cluster C. Rule r talks about people's feelings on the World Cup game. The rule represents relationships that emerge from tweets messages contained in C, i.e., the correlation between subset of words included in these messages. According to the rule support and confidence values, 1.1% of tweets in cluster C contain all the words appearing in the rule (i.e., *start, world, cup* and *love*), but the word *love* appears in 60% of tweets including the words *start, world* and *cup*.

In some cases, measuring the strength of a rule in terms of support and confidence values may be misleading. When the rule consequent has a high support value, the rule may be characterized by a high confidence value even if its actual strength is relatively low. To overcome this issue, the *lift* (or correlation) index (Pang-Ning T. and Steinbach M. and Kumar V., 2006) may be used, beyond the confidence index, to measure the (symmetric) correlation between sets X and Y. Lift values below 1 show a negative correlation between sets X and Y, while values above 1 indicate a positive correlation. In this study, to mine patterns representing strong correlations among features characterizing tweets, the selection of association rules is based on confidence and lift values.

3.4.2. Association rule categorization

Although association rules are a powerful method to discover data correlations, analyzing the (usually) large number of extracted rules is not a trivial task. To support the exploration of the mined rule set, TCharM exploits a categorization of rules into few *classes*, built upon the attributes characterizing Twitter data, i.e., tweet spatial feature (denoted *Location* (L)), tweet temporal feature ($Time$ (T)), and text content of the tweet message (*TextContent* (TC)). Each class refers to correlations among a subset of the above attributes. Specifically, four classes of rules have been defined which are aimed at progressively providing more detailed information about the cluster content. Classes are described below while an example rule is reported for each of them in Table 2.

1. *TextContent class (TC).* This class focuses on tweet text content. Patterns model correlations between words in tweet messages and these are aimed at capturing the peculiar characteristics of messages in the cluster (i.e., which topics attract/involve users). This class omits both spatial and temporal details on *when* and *where* each tweet was posted. Instead, this information is concisely represented by the location and time values of the cluster centroid, considered as representative points of the cluster.

2. *Location-TextContent class (L-TC).* This class analyses the correlations between the words in tweet messages and the locations where tweets have been posted. It makes it possible to identify the topics attracting/involving users in a given location.
3. *Time-TextContent class (T-TC).* This class analyses the correlation between words in tweet messages and the time when tweets have been posted so as to discover the topics attracting/involving users in a given time frame.
4. *Location-Time-TextContent class (L-T-TC).* This class considers all the properties characterizing tweets in order to analyse the correlation between the words in tweet messages together with the time when, and the location where, the tweets were posted. It makes it possible to discover the topics attracting/involving users in a given time frame and location.

		Example pattern	
Class	**Example question**	**Association Rule**	**Meaning**
TC	What are the topics attracting/involving users?	{world,final} $\Rightarrow$ {cup} *centroid*(T = y, L = x)	Users talked about world final cup event (reference time frame y and geographical area x)
L-TC	Given a spatial location, what are the topics attracting/involving users?	{L = x} $\Rightarrow$ {TC = (german, win, argentina)}	Users talked about the match Germany-Argentina in the geographical area x
T-TC	Given a time frame, what are the topics attracting/involving users?	{T = y} $\Rightarrow$ {TC = (best, player, PlayerName)}	Users talked about PlayerName as the best player in time frame y
L-T-TC	Given a time frame and a geographical area, what are the topics attracting/involving users?	{T = y, L = x} $\Rightarrow$ {TC = (good, performance, PlayerName)}	Users talked about the good performance of PlayerName in time frame y and geographical area x

Table 2: Reference rule classes with example rules about 2014 FIFA World Cup tweets

Topic family ID	Family description
T1	emotional states
T2	events
T3	points of interest
T4	celebrities

Table 3: List of topic families for the 2014 FIFA World Cup use case

To facilitate the semantic interpretation of the rules discovered, TCharM employs a list of reference *topic families*. A dictionary of the words characterizing each topic family is used to associate each rule with the proper family, based on the word set appearing in the rule. For instance, Table 3 reports an example list of reference *topic families* when targeting the analysis of tweets about the 2014 FIFA World Cup. The *events* family includes events such as the football matches and the opening and the closing ceremony. The *points of interest* family concerns where the events take place. Instead, the *celebrities* family regards players, coaches or other famous people somehow involved with the 2014 FIFA World Cup events.

Before applying the rule extraction process, the spatial and temporal features of tweets are processed to map their initial values into new ones with a coarse granularity in order to discover a limited but frequent set of rules. Indeed, too fine a granularity in the representation of spatio-temporal features can produce a fragmented rule set which may negatively affect the rule quality evaluation. For example, the geographical location of the user can be specified in terms of city, region, or country instead of using geo-coordinates. Similarly, the information about tweet posting time can be described with hourly or daily time slots instead of using the entire timestamp value.

3.5. TCharM *implementation*

The entire data analysis process (preprocessing, clustering, and association rules extraction) in TCharM has been implemented as a Scala application in the open source computing framework *Apache Spark* (version 1.5) (Zaharia et al., 2010). This framework was selected because it is currently one of the leading platforms for data analytics and provides a Machine Learning library (MLlib) which has been exploited and extended in this study to support all the functionalities of TCharM.

Available packages in MLlib are used for the TF-IDF weighting score calculation in the data preprocessing phase. For the subsequent cluster analysis, the K-means algorithm available in MLlib has been extended by integrating the TASTE measure. Moreover, to evaluate the quality of the generated cluster set, the computation of the *Sum of Squared Error* (SSE) index was implemented, based on TASTE and integrated in K-means too. For association rule analysis, the FP-growth algorithm (Han et al., 2000) available in MLlib was adopted to generate association rules from the computed clusters. To point out relevant association rules in clusters, we used the formulas of *support* and *confidence* values available in Apache Spark, but we also integrated the calculation of the *lift* value.

The preliminary data collection step relies on Twitter's Streaming Application Programming Interfaces (APIs) to retrieve tweets data. The Streaming APIs provide low latency access to Twitter's global stream of tweets data by establishing and maintaining a continuous connection with the stream endpoint. A Java crawler is used to collect and parse tweets in real time based on a predefined set of keywords (e.g., "worldcup2014", "fifaworldcup" in our case study), with a case-insensitive search.

4. Experimental Results

This section presents the results of the experiments with TCHARMimplementation, regarding (i) *geographical and temporal distribution* of the computed cluster sets, (ii) *clusters content characterization* through association rules analysis, and (iii) *performance evaluation* in terms of overall execution time and scalability.

The experimental evaluation was conducted on a real collection of Twitter data related to the FIFA World Cup held in Brazil in 2014. Experiments were executed on a cluster of 3 master nodes (DELL PowerEdge R620 with 128GB of RAM) and 30 worker nodes (18 DELL PowerEdge R720XD with 96GB of RAM, 2 SuperMicro with 64GB of RAM, and 10 SuperMicro with 32GB of RAM). Each node runs Cloudera distribution based on Apache Hadoop including HDFS and Apache Spark (version 1.5) for Big Data distributed applications on Linux Ubuntu 14.04.02 LTS.

4.1. Datasets

The public stream endpoint offered by the Twitter APIs was monitored over a time period of 27 days from June 18th to July 14th 2014, by tracking a selection of keywords related to the 2014 FIFA World Cup (e.g., "worldcup2014", "fifaworldcup"). Tweets in English and with the exact GPS coordinates of the user location were extracted. The resulting collection includes 302,052 tweets. To ease the computation of temporal distances between tweets in the clustering phase, all timestamps have been converted according to the reference time zone of *America/Sao Paulo*, in Brazil, where the 2014 FIFA World Cup was held.

Since the collected tweets were widely spread over both time and space, the tweets collection was partitioned into subsets referred to disjoint spatio-temporal segments before applying the cluster analysis, as follows.

To analyse how the tweet text content developed over time, the tweet collection was partitioned according to three *time windows* following the official time schedule of the football matches. *Time window #1* and *time window #2* cover respectively the first and the second stage time period (i.e., from June 18th to June 27th and from June 28th to July 3rd), while *time-window #3* covers the remaining time period from the quarter-finals to the end (i.e., from July 4th to July 14th). The number of tweets is comparable in the three windows.

The tweet spatial distribution was then locally analysed within each of the three time windows based on tweet geo-coordinates. In each time window tweets appeared to be widely dispersed and geographically partitioned into different areas. English speaking countries like the United Kingdom (UK), USA, and Central America show higher tweets concentrations than other areas. Following this evaluation of tweet spatial distribution, we selected two *spatial partitions*, corresponding to *UK* and *USA*, for each time window. Table 4 summarizes the main characteristics of the six resulting datasets which are used as reference case studies for the experimental evaluation. Each dataset was named using the corresponding spatio-temporal segment. For example, dataset $\mathcal{D}_{(TW1,UK)}$ contains tweets posted during time window #1 in UK.

4.2. Parameters configuration for cluster analysis

We set the parameters for the clustering analysis to best fit the use case considered, the 2014 FIFA World Cup, which involves people worldwide. Aimed at discovering clusters including tweets about the same topics but posted in nearby locations and time periods, we assigned the same relevance to spatial and temporal terms in modulating the text distance, i.e., we set $k_s = k_t = 0.5$. On the other hand, as usually happens on Twitter, we expect

Dataset	Time window	Geographical partition	Number of tweets	Average tweets length
$\mathcal{D}_{(TW1,UK)}$	1	UK	29,864	8.10
$\mathcal{D}_{(TW1,USA)}$	1	USA	26,447	8.02
$\mathcal{D}_{(TW2,UK)}$	2	UK	15,175	8.43
$\mathcal{D}_{(TW2,USA)}$	2	USA	19,828	8.27
$\mathcal{D}_{(TW3,UK)}$	3	UK	34,392	8.46
$\mathcal{D}_{(TW3,USA)}$	3	USA	50,028	8.06

Table 4: Main characteristics of selected reference datasets from 2014 FIFA World Cup tweets collection

most reactions to a given event (e.g., a football match) to be published as soon as the same event occurs (or within a short delay), even from quite distant locations. Indeed, while users interested in the same event can be also located in different areas, it is unlikely that they tweet at completely different times. Therefore, to group tweets with very close temporal distances, we set the weight of the temporal exponent p_t to a higher value than the spatial one p_s. We empirically found that $p_s = 3$ and $p_t = 6$ provide the lowest variability of SSE among clusters for different values of K (number of clusters) on datasets in Table 4. For each dataset, we evaluated the average SSE among the resulting clusters for a range of values of K. K was then set to 200 as a good trade-off to minimize SSE and to limit the number of clusters as well. As an example, Figure 3 plots the decrease of the average SSE for dataset $\mathcal{D}_{(TW1,UK)}$ when increasing the value of K. SSE abruptly decreases until $K = 150$, after which it goes down at a lower rate. Since we needed to limit both the desired number of clusters and the expected value of SSE, we assumed that $K = 200$ was a good trade off between these two objectives.

To address the problem of centroid initialization in K-means, a common approach was adopted. We performed multiple runs, each with a set of randomly chosen initial centroids, then we selected the cluster set with minimum SSE.

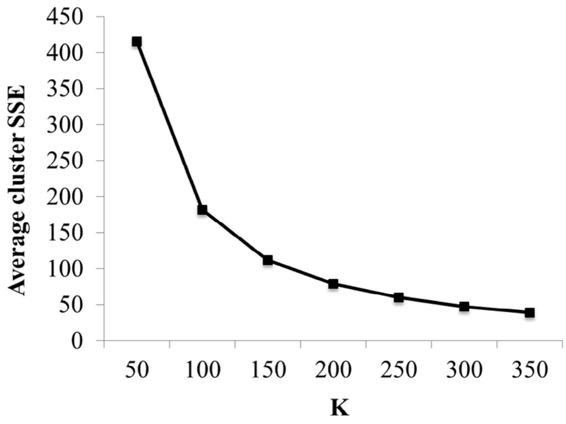


Figure 3: Variation of average cluster SSE with respect to the number of clusters (K) for dataset $\mathcal{D}_{(TW1,UK)}$ ($p_s = 3$, $p_t = 6$, $k_s = k_t = 0.5$)

4.3. Analysis of the clustering results

In this section the clustering results are characterized in terms of (i) *cluster cardinality*, given by the number of tweets per cluster, and (ii) *spatio-temporal cluster distribution*, given by the geographical area and the time span covered by the clusters. As a reference case for the analysis, we selected the collection of tweets posted in the UK partition during time window #1 (i.e., dataset $\mathcal{D}_{(TW1,UK)}$ in Table 4). This time window corresponds to the first stage in the 2014 FIFA World Cup, when there was a larger number of football matches involving many different teams. The tweets are thus potentially characterized by a higher variability of text messages as well as spatial and temporal feature values.

Figure 4 shows the distribution of clusters cardinality in the cluster set computed on dataset $\mathcal{D}_{(TW1,UK)}$. Clusters are sorted along the x axis by increasing value of cardinality. The cluster set includes one cluster with about 800 tweets, while 16.5% of clusters contain from 200 to 400 tweets, 41.5% of clusters from 100 to 200 tweets, and the remaining 41.5% less than 100 tweets. The mean value of cluster size is 132 tweets, while the median value is 111 tweets.

The spatial and temporal distributions of the cluster set are plotted in Figures 5 and 6, respectively. To facilitate understanding of the results, each cluster is concisely represented with the spatial and temporal features of its *centroid*. Moreover, for both features a coarse-grained representation

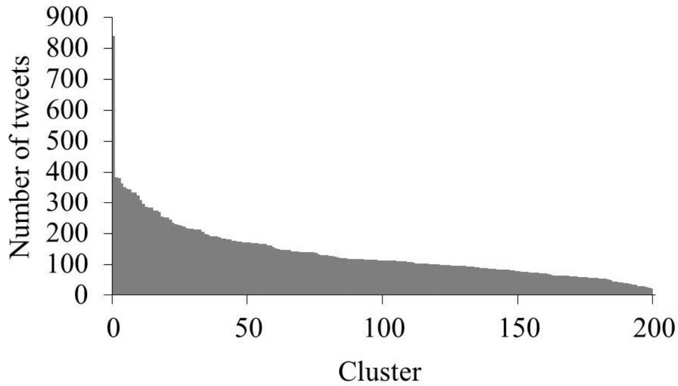


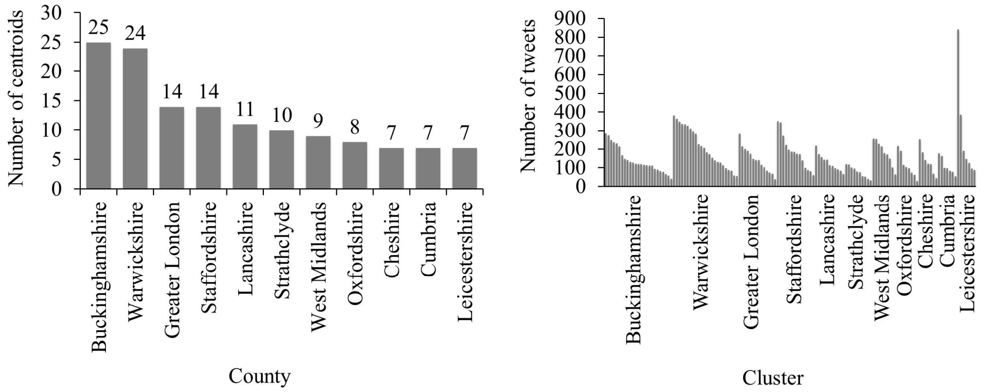
Figure 4: Distribution of number of tweets in the cluster set for dataset $\mathcal{D}_{(TW1,UK)}$

is adopted in place of the original one. Specifically, the *spatial feature* is represented as the *geographical area* where a centroid is located, instead of its GPS coordinates. Since the considered dataset contains tweets posted in UK, the *county* is used here as reference geographical area. County membership of a centroid is calculated based on the boundary GPS coordinates of each county[1] and on the GPS coordinates of the centroid. The *temporal feature* of a centroid is represented in terms of the corresponding *hourly time slot*, instead of the centroid timestamp.

The evaluation of the *spatial distribution* of centroids in the cluster set points out the *locations* in UK where people were more committed to tweeting about the 2014 FIFA World Cup 2014. Figure 5a shows the number of centroids located in each county, while Figure 5b reports the cardinality of the corresponding clusters. For each county, clusters are sorted along the x axis by decreasing value of cardinality. For readability, both figures focus on counties including at least seven centroids.

The results show that a limited subset of counties contain at least seven centroids (11 counties over 89), and about half of the centroids (98 over 200) are located in six counties (i.e., Buckinghamshire, Warwickshire, Greater London, Staffordshire, Lancashire, and Strathclyde). Clusters centered in these six counties overall include about 56% of tweets in dataset $\mathcal{D}_{(TW1,UK)}$. Moreover, thirteen of these clusters are among the fifteen largest clusters

[1] http://www.nearby.org.uk/downloads.html



(a) Number of cluster centroids per county

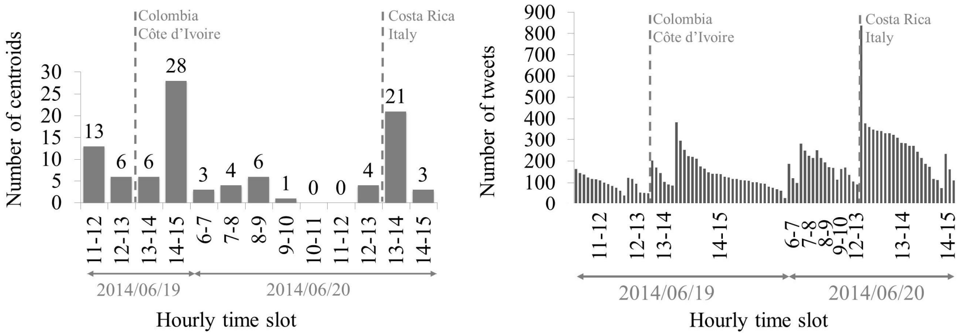
(b) Number of tweets per cluster within counties

Figure 5: Spatial characterization of the cluster set for dataset $\mathcal{D}_{(TW1,UK)}$

in the cluster set (the two largest clusters in the cluster set are centered in Leicestershire county instead). Hence, we can consider the above six counties as the locations where most tweet activity was focused in UK during time window #1.

The evaluation of the *temporal distribution* of centroids in the cluster set reveals the *time periods* when people in UK were more involved in the 2014 FIFA World Cup. As an example, we report the results for a two-day time frame (from June 19th to June 20th) within time window #1. Figure 6a shows the number of centroids located in each hourly time slot, while Figure 6b reports the cardinality of the corresponding clusters. For each hourly time slot, clusters are sorted along the x axis by decreasing value of cardinality.

Results point out that the number of clusters, as well as the number of tweets per cluster, increases in correspondence of two events, i.e., the football matches *Colombia - Cote D'Ivoire* and *Italy - Costa Rica* (the starting hour for both matches is highlighted with a dashed line in Figures 6a and 6b). More specifically, in Figure 6a a peak occurs in the hourly time slot when goals were scored in each of the two matches. For match *Colombia - Cote D'Ivoire*, the peak of 28 centroids occurs in time slot 2014/06/19 [14:00-15:00) which corresponds to the second half of the match when three goals were scored. Instead, for the match *Italy - Costa Rica*, the peak of 21 centroids occurs in time slot 2014/06/20 [13:00-14:00) which corresponds to the first



(a) Number of cluster centroids per hourly time slot

(b) Number of tweets per cluster within hourly time slot

Figure 6: Temporal characterization of the cluster set for dataset $\mathcal{D}_{(TW1,UK)}$ during a two-days time frame

half of the match when the only goal of the match was scored.

To deepen the analysis of the spatio-temporal span for the discovered clusters, we focus on four example clusters selected among those with the centroid located in the Greater London county. The characteristics of these clusters are summarized in Table 5 in terms of (i) spatial and temporal features of the cluster centroid, (ii) cluster cardinality, (iii) cluster spatial cohesion as average geographical distance between tweets in the cluster and the cluster centroid, and (iv) cluster temporal cohesion as average time distance between tweets in the cluster and the cluster centroid. Since all the centroids are located in the Greater London county, to describe their spatial features Table 5 also reports the town where each centroid is placed.

Clusters manifest a good temporal cohesion since the average time distance is always about 20 minutes. This temporal span is suitable to associate clusters to some specific events. For example, clusters A and C span on time intervals including the *Colombia - Cote D'Ivoire* and *Italy - Costa Rica* football matches, respectively. Tweets in clusters B and D mainly discuss the elimination of the England football team that occurred the day before. These tweets may have been posted in response to news reporting this event on sports channels (also mentioned in tweet messages and taking place near the centroid time).

Clusters also demonstrate a reasonable spatial cohesion around their centroid, since tweets within each cluster are mainly (or even exclusively) posted

	Cluster centroid		Cluster content		
Cluster ID	**Spatial location of centroid (County:City)**	**Temporal slot of centroid (Date:hourly time slot)**	**# of tweets**	**Avg GPS distance (km)**	**Avg time distance (min)**
A	Greater London: Harrow	2014/06/19 [14-15)	113	59.25	26
B	Greater London: Stratford	2014/06/25 [08-09)	188	42.35	20
C	Greater London: Uxbridge	2014/06/20 [13-14)	283	68.73	23
D	Greater London: London	2014/06/25 [07-08)	197	42.24	19

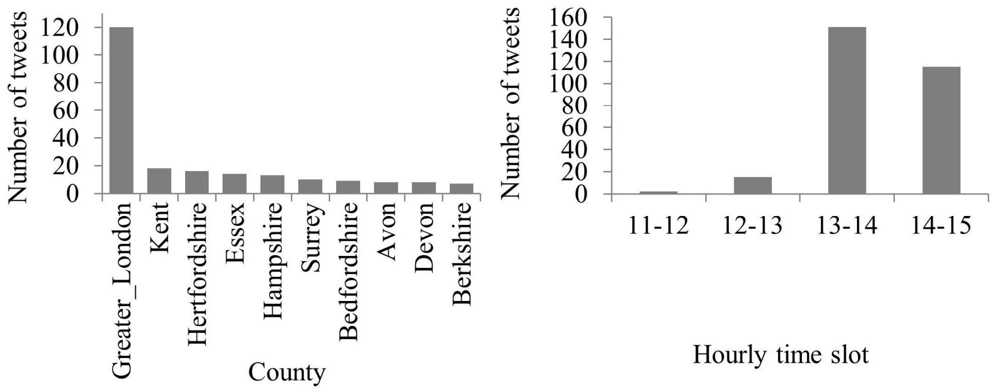
Table 5: Characterization of four example clusters centered in Greater London county

in the same county where the centroid is located. The larger geographical area covered by each cluster is due to the fact that events related to the FIFA World Cup are of widespread interest.

As an example, Figure 7 reports the distribution of the number of tweets in the top ten counties and over time for the cluster with the highest cardinality in Table 5, i.e., cluster C. Most tweets were posted in the Greater London county where the cluster centroid is located, while the other tweets are mainly spread out in four of the neighboring counties. Furthermore, the tweets were mainly posted during the hourly time slots adjacent to the slot of the centroid.

4.4. Clusters characterization using association rules

The cluster content is concisely described here using association rules to model correlations among tweet features (text content, location, and time). The rules are extracted according to the rule templates defined in Section 3.4.2 and the topic families reported in the same section. To discuss the type of information that can be mined using these patterns, some example rules are reported in the next subsections. These rules have been extracted from (i) one sample cluster, (ii) clusters mined in time window #1



(a) Number of tweets per county

(b) Number of tweets per hourly time slot

Figure 7: Spatial and temporal characterization of cluster C (from Table 5)

and from different geographical partitions, and (iii) clusters computed for different time windows from the UK partition. For the rule extraction, we enforced *support* $\geq$ 1%, and $lift > 1$ to prune both negatively correlated and uncorrelated item combinations.

4.4.1. Analysis of rules on a sample cluster

Cluster C (see Table 5) from dataset $\mathcal{D}_{(TW1,UK)}$ was selected as the reference case for the analysis. To reduce data fragmentation in the extracted patterns, caused by the spatio-temporal sparsity of the data collection, the tweet geo-coordinates have been mapped to the corresponding counties and the tweet posting timestamp to the corresponding 2-hours time slot.

Experimental results showed that the association rules generated from cluster C concern a variety of topics such as events, emotional states and celebrities, mainly related to the *Italy - Costa Rica* football match scheduled on June 20th, 2014. A selection of significant rules is reported in Table 6 and they are briefly described below.

Analysis of correlations in tweet text content (class TC). The rules in the class TC model correlations in the tweet text content. The information about when and where tweets were posted is concisely described as spatial and temporal details of the cluster centroid. Rules like R_1 and R_2 represent strong pairwise correlations (according to the lift value) among words in tweet messages. Rule R_1 captures a positive emotional state in people for the Costa

Rica football team. Instead, in rule R_2 people talked about the celebrity Gary Lineker, a retired English footballer and current sports broadcaster, who wore an Italy shirt. The reason is that the victory of Italy over Costa Rica would have allowed the England football team to keep their World Cup hopes alive.

Analysis of correlations between the location where tweets were posted and tweets text content (class L-TC). Rules in the L-TC class, like rules R_3 and R_4, point out the geographical areas where certain topics are discussed. Rule R_3 reveals that a negative emotional state about the England football team arises from people located in the Greater London county. This opinion may be due to the fact that the England football team did not win any match in the first stage of the World Cup. Instead, rule R_4 reports that people in the Greater London county are watching how the Costa Rica football team performs in matches.

Analysis of correlations between the time when tweets were posted and tweets text content (class T-TC). Rules in the T-TC class, such as R_5 and R_6, point out the time slot when certain topics are discussed. Rule R_5 describes the association between people's disappointment about the behavior of the Italian football team and the time slot including the football match *Italy - Costa Rica*. In fact, after the goal scored by Costa Rica in the first half of the match, the Italian team did not respond with any winning actions in the second half of the match. Rule R_6 highlights the people's interest in the comments on the *Italy - Costa Rica* match by a former English player (Robbie Savage) hired as pundit by the British Broadcasting Corporation (BBC) for the 2014 FIFA World Cup.

Analysis of correlations between location where, and time when, tweets were posted and tweet text content (class L-T-TC). R_7 and R_8 are example rules belonging to this class and they both show that the goal scored by Costa Rica and the consequent defeat of the Italian team in the time slot including the first half of the match was a hot topic in the Greater London county.

It is worth noting that the rules of classes *L-TC*, *T-TC* and *L-T-TC*, characterized by positive correlation and high confidence values, always include the same county and hourly time slot of the centroid. This provides further evidence in support of the high spatio-temporal cohesion of cluster C around its centroid.

Rule class	**Rule ID**	**Topic family**	**Rule**	**supp [%]**	**conf [%]**	**lift**
TC	R_1	Emotional state	{fancy,costa,rica} ⇒ {chances}	1.1	75	53.25
	R_2	Celebrity	{shirt,italy} ⇒ {lineker}	1.1	100	56.80
L-TC	R_3	Emotional state	{TC = (bad,england)} ⇒ {L = Greater London}	1.1	100	2.37
	R_4	Event	{TC = (watching,costa,rica)} ⇒ {L = Greater London}	1.4	66	1.58
T-TC	R_5	Emotional state	{TC = (bad,italy)} ⇒ {T = 2014-06-20 [12:00-14:00)}	1.1	50	1.23
	R_6	Celebrity	{TC = (robbiesavage,playing,italy,costa)} ⇒ {T = 2014-06-20 [12:00-14:00)}	1.1	100	1.71
L-T-TC	R_7	Event	{T = 2014-06-20 [12:00-14:00), TC = (lose,italy) } ⇒ {L = Greater London}	1.1	60	1.42
	R_8	Event	{T = 2014-06-20 [12:00-14:00), L = Greater London, TC=(costa,rica)} ⇒ {TC=(goal)}	1.1	15	10.65

Table 6: Example rules from cluster C (*centroid*(T = *2014-06-20 [12:00-14:00)*, L = *Greater London*)) from dataset $\mathcal{D}_{(TW1,UK)}$ (see Table 5)

4.4.2. Analysis of rules across geographical partitions

In this section we analyse how people's interest in events occurring within a given time window vary across different geographical areas. We compared the association rules mined from clusters computed in UK and USA areas when considering time window #1 (datasets $\mathcal{D}_{(TW1,UK)}$ and $\mathcal{D}_{(TW1,USA)}$). To reduce data fragmentation in the mined patterns, we adopted a coarse spatio-temporal data representation suitable for both cases considered. Specifically, tweet geo-coordinates have been mapped to the nearest city and the tweet posting timestamp to the corresponding day. Some sample rules modeling correlations in the tweet text content (class TC) are shown in Table 7, but the following discussion is based on the overall results.

People in the UK area commented mostly on matches involving the England football team (e.g., rule R_1), or other teams included in the same group as England. Moreover, an odd episode involving a single player was the main topic of various clusters (R_2). Instead, clusters from many locations of the USA reveal that people were interested in matches involving various football teams, also those not included in the same group as their national team. For instance, rule R_3 refers to the match between Italy and Costa Rica and rule R_4 to the match involving Nigeria and Argentina.

The behaviour observed may be related to the people's different interests in the two geographical areas. Overall, football is more popular in England than in USA, where people are mostly interested in other sports. While in England people particularly focus on events related to their national team, in USA they show a more general interest in the FIFA World Cup, also for events involving teams other than their national team.

4.4.3. Analysis of rules across time windows

In this section we analyse how the interests of people tweeting from the same geographical area vary for events that occurred in different time windows. We compared rules mined from clusters computed in the UK area in the three time windows (datasets $\mathcal{D}_{(TW1,UK)}$, $\mathcal{D}_{(TW2,UK)}$, and $\mathcal{D}_{(TW3,UK)}$). We adopted the same spatio-temporal data representation used for the analysis discussed in Section 4.4.2. Table 8 shows some example rules from the TC class, but the discussion is based on the overall results.

It is worth noting how interests varied after the elimination of England team which happened at the end of time window #1. The extracted rules show that people in UK shifted their attention to matches involving other teams. Various clusters in time window #2 are focused on the

Rule id	Partition	Topic family	Rule	supp [%]	conf [%]	lift
R_1	UK	Event	{uruguay} $\Rightarrow$ {england} *centroid*(T = *2014-06-19*, L = *Perth*)	5.0	100	2.38
R_2	UK	Celebrity	{suarez,someone} $\Rightarrow$ {bite} *centroid*(T = *2014-06-25*, L = *Rugeley*)	3.0	80	26.90
R_3	USA	Event	{costa,rica} $\Rightarrow$ {italy} *centroid*(T = *2014-06-20*, L = *Whittier,CA*)	8.3	64	1.67
R_4	USA	Event	{nigeria} $\Rightarrow$ {argentina} *centroid*(T = *2014-06-25*, L = *Banning,CA*)	2.1	53	7.16

Table 7: Example rules (class TC) characterizing clusters in UK and USA areas in time window #1 (datasets $\mathcal{D}_{(TW1,UK)}$ and $\mathcal{D}_{(TW1,USA)}$)

Germany – Algeria football match (played on June 30^{th}, 2014), and are mostly about the tactics (R_5) and performance (R_6) of the German team.

During time window #3, the final match became one of the most popular topics (R_7). Nevertheless, the attention of people in UK also moved towards other topics loosely related to the competition. For instance, the latest transfer of player Luis Suarez away from an English club was mainly discussed on July 11^{th} 2014, on the same day as the official announcement (R_8), while the next match of the England team, scheduled for November against Scotland (R_9), became popular just after the final World Cup match, on July 14^{th} 2014.

4.5. Execution time and scalability

The execution time for the cluster set computation on the six datasets in Table 4 spans from 12m 13s for the smallest dataset ($\mathcal{D}_{(TW2,UK)}$, 15,175 tweets) up to 33m 34s for the largest one ($\mathcal{D}_{(TW3,USA)}$, 50,028 tweets). The execution time for association rules extraction is less variable and has an overall mean value of 53s. Increasing the number of executors does not yield better performance in terms of clustering execution time due to the limited size of these datasets. Thus, experiments for these datasets were performed using one execution node.

Rule id	Time window	Topic description	Rule	supp [%]	conf [%]	lift
R_1	1	Event	{uruguay} ⇒ {england} *centroid*(T = *2014-06-19*, L = *Perth*)	5.0	100	2.38
R_2	1	Celebrity	{suarez, someone} ⇒ {bite} *centroid*(T = *2014-06-25*, L = *Rugeley*)	3.0	80	26.90
R_5	2	Event	{line,high} ⇒ {germany} *centroid*(T = *2014-06-30*, L = *London*)	2.0	100	1.02
R_6	2	Emotional state	{good} ⇒ {germany} *centroid*(T = *2014-06-30*, L = *Stirling*)	2.0	58	1.22
R_7	3	Event	{world, cup} ⇒ {final} *centroid*(T = *2014-07-13*, L = *Newcastle*)	10.2	99	2.91
R_8	3	Celebrity	{suarez} ⇒ {good,luck} *centroid*(T = *2014-07-11*, L = *London*)	2.3	77	24.40
R_9	3	Event	{november} ⇒ {england,scotland} *centroid*(T = *2014-07-14*, L = *Broxbourne*)	1.8	100	36.71

Table 8: Example rules (class TC) characterizing clusters across the three time windows in UK area (datasets $\mathcal{D}_{(TW1,UK)}$, $\mathcal{D}_{(TW2,UK)}$, $\mathcal{D}_{(TW3,UK)}$)

The capacity of the clustering algorithm integrating the TASTE measure to scale up to bigger data collections was assessed by measuring the execution time when varying (i) the number of tweets under analysis and (ii) the number of parallel executors. For scalability analysis, to get a larger number of tweets including all (text, temporal, and spatial) features, we have considered the location specified in the user profile as reference location information. Indeed the amount of tweets with geo-coordinates is much less than the number of tweets with location information in the user profile due to the limitation of GPS enabled devices. Geo-coordinates for the location extracted from the user profile have been calculated using Bing Maps Locations API. The resulting dataset, named $\mathcal{D}''$, includes about 23.5 million tweets.

To study scalability by varying the number of tweets, we considered different sample rates of dataset $\mathcal{D}''$ and one executor for process running. Increasing the number of tweets from 50,000 to about 2.35 million (10% of whole $\mathcal{D}''$), we notice an increment of the execution time (from 33m 34s to 14h 31m). However, the growth rate of the execution time (about 25) is almost half the growth rate of the dataset size (about 47).

To study scalability by varying the number of executors, we considered the whole dataset $\mathcal{D}''$. The results show that, when increasing the number of executors from 4 to 8, the K-means algorithm integrating the TASTE measure scales almost linearly. The execution time is about 35h 43m with 4 nodes; it decreases to about 19h 24m with 6 nodes, and to 10h 45m with 8 nodes. Thus, with a suitable number of parallel executors, the clustering task is capable of handling also bigger data, evenly distributing the load across the nodes. When fewer than 4 executors are used, the process exceeded 48 hours of execution and it was interrupted due to the very large dataset size.

5. Comparison with previous studies

This section discusses both a theoretical and analytical comparison between our work and four previous studies on clustering Twitter data: (Kim et al., 2011),(Arcaini et al., 2016), (Lee, 2012), and (Cunha et al., 2014). These studies have proposed distance measures which combine the same tweet features considered in TASTE, or a subset of them. Specifically, the work in (Kim et al., 2011) takes into account the tweet spatial feature, while the spatio-temporal features are considered in (Arcaini et al., 2016), and both the text content and the spatial feature are evaluated in (Lee, 2012). A first attempt in considering all the three tweet features was proposed in (Cunha et al., 2014). Like in TCharM, in these studies the geographic and temporal distances between tweets are computed using the Haversine and the Euclidean distance, respectively. The text content is represented with the BOW model, and the word relevance is weighted with the TF-IDF (Cunha et al., 2014) or the BursT (Lee, 2012) score; the cosine similarity is used to compare messages.

For each study we present the objective of the work and the methodology for clustering tweets, including the clustering algorithm, the distance functions used and the strategy adopted for combining tweet features. Then, we discuss the analytical comparison between these works and our approach.

In the following, we adopt the same notations as in Sections 3.1 and 3.3. An arbitrary tweet τ_i is a triplet $\tau_i = (t_i, s_i, W_i)$ where t_i and s_i are respectively the temporal and spatial features of τ_i, while $W_i \subseteq \Sigma$ is the tweet text content. Given two tweets $\tau_i = (t_i, s_i, W_i)$ and $\tau_j = (t_j, s_j, W_j)$ their temporal, spatial and content distances are denoted by $d_t(t_i, t_j)$, $d_s(s_i, s_j)$, and $d_W(W_i, W_j)$, respectively.

The work in Kim et al. (2011) aims at providing (near-)real time information to users about events happening close to their location. Tweets are clustered through the K-means algorithm by considering their geographic distance. The discovered cluster set is then analysed to detect clusters that can reveal the occurrence of an event. The values of the tweet temporal feature are used to filter computed clusters by comparing their temporal aspects. If the number of tweets from a given cluster exceed far from those from clusters found in vicinity in the past, the cluster is considered unusual and an event may happen there. For tweets included in unusual clusters, the text content is explored to extract representative keywords, which are sent to nearby users to inform them about the possible events.

The study in Arcaini et al. (2016) focuses on discovering spatio-temporal periodic and aperiodic characteristics of events to support situation awareness. Tweets collections are analysed off-line with a DBSCAN based algorithm (GT-DBSCAN) to extract dense clusters of arbitrary shapes. The tweet text content is explored in a preprocessing phase to filter the subset of tweets relevant for the subsequent cluster analysis. Messages about specific events are selected by properly setting keywords for tweets search. To drive the clustering process, three distance measures, considering the tweet temporal and spatial features, are evaluated: (i) a temporal distance, (ii) a geographic distance, and (iii) a geographic-temporal distance, basically a combination of the two above. In this study we focus on the latter distance measure for performance comparison. The geographic-temporal distance is defined as the maximum value between the (normalized) geographic and temporal distances.

The work in Lee (2012) proposes a (near-)real time temporal-text clustering approach to detect bursts of tweets representing unexpectedly frequent occurrences of a certain topic in a short period of time. A sliding window of fixed time length is used to filter only the most recent tweets, which are then considered in the analysis. Selected tweets are clustered using the IncrementalDBSCAN algorithm (Ester et al., 1998), to detect dense clusters with shapes changing over time and to remove uninformative tweets (outliers).

Study	Distance measure
Kim et al. (2011)	$d_{Kim}(\tau_i, \tau_j) = d_s$
Arcaini et al. (2016)	$d_{Arc}(\tau_i, \tau_j) = [\mathrm{Max}(d_s,d_t)]^{\beta}$, $\beta \in (0,1]$ d_s and d_t values expressed as the number of elementary units ϵ_s and ϵ_t, respectively
Lee (2012)	$d_{Lee}(\tau_i, \tau_j) = d_W \cdot \mathrm{e}^{\zeta d_t / M}$ M: time unit; ζ: exponential decay rate factor.
Cunha et al. (2014)	$d_{Cun}(\tau_i, \tau_j) = w_W \cdot d_W + w_t \cdot d_t + w_s \cdot d_s + w_{So} \cdot d_{So}$ $w_W, w_t, w_s, w_{So} \in [0,1]$ and $w_W + w_t + w_s + w_{So} = 1$
TCharM	$d_{TASTE}(\tau_i, \tau_j) = d_W \cdot (k_s \cdot \mathrm{e}^{p_s \cdot d_s} + k_t \cdot \mathrm{e}^{p_t \cdot d_t})$ $k_s, k_t, p_s, p_t \in \mathbb{R}$; $k_s, k_t \in [0,1]$ and $k_s + k_t = 1$.

Table 9: Distance measures for tweet comparison proposed in four reference previous studies and in TCharM. For a pair of tweets (τ_i, τ_j), their spatial distance $d_s(s_i, s_j)$ is shortly denoted by d_s, the temporal distance $d_t(t_i, t_j)$ by d_t, the content distance $d_W(W_i, W_j)$ by d_W, and the social distance $d_{So}(user_i, user_j)$ by d_{So}.

Clusters are calculated by evaluating the temporal-text distance between tweets. In d_{Lee}, the temporal distance is used to module the text content distance. The exponential form has been adopted for the time distance to significantly penalize tweets far distant in time. Finally, geo-spatial keywords are extracted from message in each computed cluster to estimate location of detected events.

The authors of Cunha et al. (2014) address the problem of identifying and displaying tweets profiles considering four different facets characterizing tweets: temporal, spatial, and context features and user social connections. Tweets are clustered with the DBSCAN algorithm Ester et al. (1996) to detect arbitrarily shaped clusters and to remove outliers from the results. The adopted distance measure is a linear combination of the four considered tweet features, i.e., the distance on time, space, text content, and social relations (d_{So}). The social distance term d_{So} evaluates the connections between users represented as nodes of a graph connected through edges. It is computed as the geodesic distance (i.e., the number of edges of the shortest path) between two nodes in the graph Bouttier et al. (2003).

Based on the purposes of this paper, we want to evaluate the ability of each distance measure above in discovering cohesive clusters of tweets to be represented through their centroids. Hence, keeping the K-means algorithm used in TCharM as a reference clustering method, we applied in turn each distance measure. Since the TCharM methodology aims at discovering cohesive clusters considering temporal and spatial tweet features and text content, we omitted the social distance for the measure proposed in (Cunha et al., 2014). For the sake of brevity, the resulting clustering methods are denoted by Cunha-14 (Cunha et al., 2014), Lee-12 (Lee, 2012), Arcaini-16 (Arcaini et al., 2016), and Kim-11 (Kim et al., 2011). The approach proposed in this study adopting the TASTE measure is denoted by TCharM.

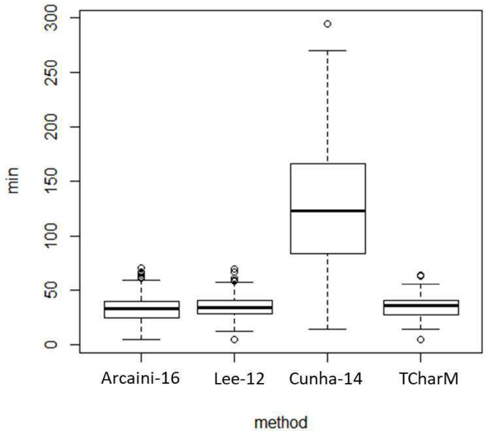
We evaluated the cluster cohesion as the average geographic/temporal/text content distance between tweets in the cluster and the cluster centroid. Lower values of these average distances point out a higher degree of cohesion on the corresponding tweet dimension.

The comparison was performed with the $\mathcal{D}_{(TW1,UK)}$ dataset. To produce comparable cluster sets, we forced K=200 as expected number of clusters for all the distance measures (i.e., the same value selected for TCharM in Section 4.2). We suitably tuned the parameters to use each distance measure at its best with the $\mathcal{D}_{(TW1,UK)}$ datasets and with the K-means algorithm. Starting from the configuration proposed in each study (considered as default configuration), we performed several runs to tune the parameters of each distance measure, with the aim of reducing the average cluster SSE as well as the distance values for all the tweet dimensions they consider. Selected parameter values are reported in Figure 8.

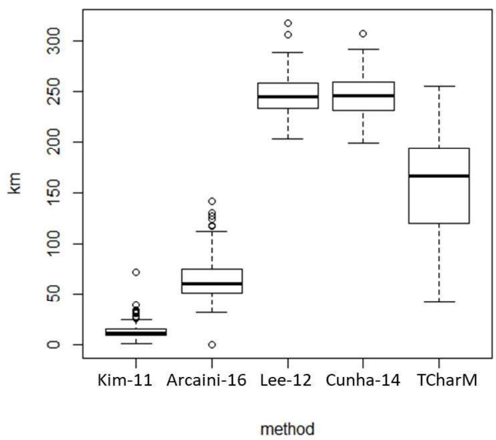
For each method, box plots in Figure 8 illustrate the distributions of the average geographic/temporal/text content distance between tweets in each cluster and cluster centroid, while Table 10 reports the average values. Note that the temporal box plot for the Kim-11's measure is not represented in Figure 8 as its values are too high compared to the other methods.

Clusters manifest the highest text content cohesion with TCharM, Cunha-14 and Lee-12 distance measures, which provide comparable results. The highest temporal cohesion is provided by Arcaini-16, TCharM and Lee-12, which achieve similar performance. The highest spatial cohesion is given by Kim-11, followed by Arcaini-16, and then TCharM.

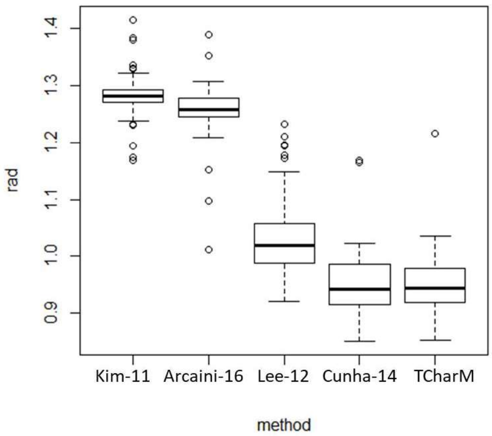
These results point out that TCharM provides clusters with an overall good cohesion on all the three facets characterizing tweets. Specifically, computed clusters show the highest cohesion on the text content and on the temporal feature, and the third best spatial cohesion. Yet it should be noted that, when setting parameters in TASTE, we gave more importance to the temporal cohesion than to the spatial one.



(a) Average temporal distance from centroid



(b) Average spatial distance from centroid



(c) Average text content distance from centroid

Figure 8: Distributions of the average temporal, spatial, text content distances from cluster centroids, for each method. The temporal box plot for Kim-11 is not represented as its values are too high. Parameter configurations are as follows. Arcaini-16: ($\epsilon_s = 2km$, $\epsilon_t = 1200s$, $\beta = 1$), Cunha-14: ($w_s = w_W = 0.25, w_t = 0.5, w_{So} = 0$), Lee-12: ($\zeta/M = 12h^{-1}$).

Method	Avg time distance (min)	Avg GPS distance (km)	Avg text content distance (rad)
Kim-11	3905	**14**	1.28
Arcaini-16	**33**	66	1.26
Lee-12	**35**	246	**1.03**
Cunha-14	126	245	**0.95**
TCharM	**35**	158	**0.95**

Table 10: Average value of mean temporal, spatial, and text content distances between tweets and their centroids for each distance measure.

Clusters provided by Arcaini-16, Lee-12, and Kim-11 methods show a good cohesion on the tweet features considered in their proposed distance measures, but the cohesion on the remaining features is far lower than TCharM. Clusters tend to be spread over a larger geographic area (Lee-12) or a longer time period (Kim-11) than TCharM, or to discuss more different topics (Kim-11, Arcaini-16). These results demonstrate that, to obtain clusters suitable for a subsequent characterization of their spatial, temporal and text features, it is convenient to consider all the three dimensions directly in the clustering phase. Otherwise, further post-processing steps would be required to characterize the clusters with the features previously left out.

Results also highlight that, when all three features are considered to cluster tweets, their contributions should be properly weighted in the distance measure. A liner combination of the content, spatial, and temporal distances as the one proposed in Cunha-14 turns out to be less suitable than our approach since discovered clusters manifest a temporal and spatial cohesion lower than TCharM.

To deepen into the comparison of the methods above, we used the Adjusted Rand Index (ARI) (Hubert & Arabie, 1985) to evaluate the agreement between the cluster sets generated using the TASTE measure and those obtained with the other distance measures. The ARI computes the rate of pairwise agreements between two partitions of a set. It allows a more accurate estimation of the agreement between two partitions than the standard

Rand Index (Rand, 1971). Basically, ARI rescales the Rand Index value **with respect to its expected value for two independent clustering algorithms.** ARI has a maximum value of 1 for two identical partitions, and an expected value of 0 for two independent random partitions. Higher ARI values imply higher levels of agreement between two partitions.

The computed values of ARI report a moderate agreement between the cluster set provided by TCharMand the one computed by Cunha-14 (ARI = 0.45). The agreement decreases with Lee-11 (ARI = 0.13), Arcaini-16 (ARI = 0.03), and Kim-11 (ARI = 0.005) methods which consider a subset of tweet features.

The results from the analytical comparison suggest that clusters discovered using other distance measures have quite different properties than those provided by TCharM.

From a temporal perspective, clusters can have a higher temporal span. Indeed, while our clusters are centered around events of interest (see Section 4.3), we noticed that clusters computed with other methods (Kim-11 and Cunha-14) can include more than one event (e.g., more football matches). Similarly, the distance measures that do not provide a good cluster spatial cohesion lead to clusters of tweets spread across more counties (Lee, 2012). The two aspects above prevent from performing qualitative analyses based on fine-grained temporal and spatial resolutions. Finally, the lower text similarity among tweets in the clusters (Arcaini-16) makes it difficult to associate a single prevailing topic with each cluster and to generate significant association rules (i.e., with high values of quality indices as support, confidence and lift).

Thus, with the adoption of other distance measures than TASTE, a further level of segmentation would be required to identify the main topics in each cluster, or to partition the cluster content into subsets which refer to shorter time windows or more limited geographic areas.

6. Discussion

In this section we discuss the results discovered through TCharM. The discussion addresses the data analysis phases in TCharM, the computational cost of TCharM, and the possible exploitation of the TCharM findings.

(i) Discovering in one step cohesive spatio-temporal clusters focused on specific topics. The TCharM findings demonstrate the ability of the proposed methodology to properly analyse large tweet collections distributed over time and space as well as addressing various topics for automatically computing cohesive clusters. TCharM allows data miners to discover clusters useful for identifying *when* and *where* people were more involved and about *which* topics. The 2014 FIFA World Cup use case considered in this study enables a thorough validation of computed clusters due to the availability of a time schedule for the main events (e.g., football matches) and web news about the other events or celebrities somehow involved. The experimental evaluation conducted on six different datasets showed that mined clusters are centered in time in correspondence with an event related to the 2014 FIFA World Cup and they mainly include messages about the event. Moreover, the clusters present a good spatio-temporal cohesion around their centroid.

Differently from previous work (see Section 2), TCharM clusters Twitter data taking into account in one step both spatio-temporal features and text content. TCharM relies on the TASTE measure which combines the contributions of all three features above. TASTE modulates the distance between tweet messages through their distance in time and space, and it is aimed at discovering groups of tweets about the same topic but posted in nearby time periods and locations. Parameters of the TASTE measure can be conveniently tuned to fit scenarios with different spatial and temporal granularities.

The analytical comparison in Section 5 shows that TCharM **is competitive in terms of cluster cohesion, in almost all dimensions. In particular, it overperforms all the other measures in the text average distance. Indeed, the multiplicative (exponential) factors for time and space distances are suitably applied to the text distance, based on the hypothesis that a tight temporal and spatial proximity can contribute in detecting clusters of tweets about the same topic. As already demonstrated in Section 4.3, such clusters are temporally centered within the time interval of the event they refer to (e.g., a football match).**

None of the measures considered for comparison performs far better than TASTE **in more than one dimension. Moreover, the lower spatial cohesion obtained with** TASTE **is mainly due to our choice to assign a lower weight to spatial distance ($p_s = 3$), preferring the temporal cohesion ($p_t = 6$).**

(ii) Cluster characterization through rules analysis. TCharM deeply explores the resulting clusters through association rule analysis to discover correlations among topics (such as events, celebrities, emotional states) and spatio-temporal features. While rule class TC makes possible the identification of the main topics discussed in each cluster, the other rule classes enable a deeper characterization by correlating topics with time periods (class T-TC), geographical areas (L-TC), or both of them (class L-T-TC). This cluster characterization allows data miners to better understand popular topics in different geographical areas and through different time windows. Moreover, association rules represent the mined knowledge in a concise and easily understandable form.

The 2014 FIFA World Cup use case allows us to qualitatively validate various mined rules. Rule analysis pointed out some of the interests and reactions of sports fans and supporters that were in some cases predictable (e.g., the disappointment of people from England over the English team's defeats). However, it also highlighted some aspects not so evident a priori, like those about celebrities statements or the major interest in USA for the team of Argentina. We believe that TCharM can be applied also in other scenarios, for understanding people's reactions and interests.

(iii) TCharM *performance.* From a computational point of view, TCharM has a major advantage with respect to related works, since it is implemented on Apache Spark and can distribute computational load across parallel executors. Tests performed on big collections of tweets (Section 4.5) prove the good scalability of our implementation of TCharM and, in particular, of the clustering algorithm integrating the TASTE measure. Thus, TCharM can be applied also to use cases with a higher cardinality of data and it is still capable to provide results in a reasonable time.

(iv) Exploitation of the mined knowledge. TCharM findings provide a spatio-temporal overview of people involvement in occurred events. This knowledge, hidden in Twitter data collections, can have a variety of practical applications in different domains.

In case of events with a wide and spread out audience (as FIFA World Cup), TCharM findings can provide useful insights to understand how people located in different geographical areas perceive an event and to characterize the different facets of people involvement in different time frames. From a business perspective, this knowledge can be very useful to improve service/product provision and support targeted advertising of certain services/products. For instance, the information about favourite teams or players in specific areas and moments can be used to provide targeted advertising that leverages on such features. Also during 2014 FIFA World Cup, advertising companies demonstrated great interest in social trends to plan marketing strategies. This was particularly evident with some viral topics as some brands gained visibility by proposing advertisements based on viral marketing strategies, mostly on social networks (Jenkins, 2014; Bud, 2014). TCharM can thus be an effective methodology to enable a deeper analysis of spatio-temporal trends on social networks, showing when and where certain topics spread among users.

We believe that TCharM can be profitably applied also in different domains. In a smart urban environment, for example, social networks are currently recognized as powerful instruments to enable citizen interaction and participation. Citizens may use Twitter to report information related to a variety of aspects such as urban safety, traffic and services (e.g., bike sharing, public transport offer, etc.). City administration is interested in better understanding where and when citizens report issues about the above aspects, to eventually undertake appropriate and targeted responses to citizens' concerns. The application of TCharM to such collections of tweets would help to find out in which areas of the city and in which periods of time citizens discuss and complain about some issues. Clustering analysis would extract spatio-temporally defined clusters of topics reported by citizens. Rule analysis would then better highlight the degrees of correlation among topics, times and places of discussion and describe how the same topics evolve across different periods and through nearby urban areas.

7. Conclusion

In this paper we introduce TCharM, a novel exploratory data mining methodology to analyse Twitter datasets. Its aim is to discover significant and cohesive groups of tweets by considering three facets of Twitter data: spatial, temporal, and text content information. The TASTE measure is one of the main added values of TCharM as it allows the K-means algorithm to discover clusters with suitable levels of spatial and temporal cohesion, centered on specific events and including tweets which can be concisely represented by their centroids with an acceptable approximation. Moreover, through association rules mining, TCharM provides us with a set of patterns that concisely describe the most significant characteristics of tweets in clusters. The TCharM system has been deployed on Apache Spark to distribute computational load across parallel executors and reduce the overall execution time also with huge amounts of data.

The experimental validation conducted on tweets collected for the 2014 FIFA World Cup demonstrated the ability of TCharM in efficiently characterizing collections of tweets in terms of distribution of people involvement, topic identification, and correlations among tweet features. As a matter of fact, we managed to isolate groups of tweets focused on a few topics, temporarily associated to actual events (e.g., football matches), and posted from a limited geographical area. Compared with other approaches for tweet clustering, clusters computed using the TASTE measure confirmed an overall better cohesion balanced between the three tweet features.

TCharM can be an effective methodology to enable a deeper analysis of spatio-temporal trends on social networks, showing the different patterns of user involvement in certain topics or events. TCharM can be used to analyse global events like the FIFA World Cup at a local scale and, for instance, to assess the popularity of soccer matches and football players in different areas and time periods. This information could be very useful for companies to improve their services and products and to optimize their marketing strategies. For example, information about favourite teams and players in specific areas and moments can be used to provide targeted advertising that leverages on the characteristics of the computed clusters.

There is still room for improvement of the TCharM methodology in order to mitigate some of its weaknesses. Five promising future research directions have been identified.

In the current implementation of TCharM, the number of expected clusters for the k-means algorithm and the parameters in the TASTE measure should be experimentally tuned by trading-off the cardinality of the cluster set and the expected quality of clusters. However, the selection of the proper TCharM configuration can be a very time-consuming activity. The design of innovative *self-tuning configuration strategies* Di Corso et al. (2017) to automatically identify the suitable TCharM set up for each targeted data collection can permit the use of TCharM in various application domains. These strategies would simplify the analysts role by relieving the end-user of the burden of configuring the overall cluster analysis process.

The ability of TCharM to discover cohesive and significant clusters may decrease when data sparseness further increases. In this case, a larger number of clusters should be generated to discover groups with good quality, but these groups may be limited in size. To deal with this issue, *data taxonomies* on the three facets characterizing tweets can be climbed during the clustering process. The use of data taxonomies can result into coarse-grained data representations with a lower degree of sparsness and allows the evaluation of data correlations at different abstraction levels.

The use of K-means clustering, rather than other clustering algorithms as density-based methods, was motivated in this study by the purpose of generating clusters of tweets that can be concisely represented by their centroids. However, TCharM inherits one of the main weaknesses of K-means, which is more sensitive to outliers in the dataset. A future task is to conduct a detailed study on evaluating the *integration of other candidate clustering methods* in TCharM and their ability to identify more cohesive and significant clusters of tweets.

Currently, the proposed TASTE measure weights various tweet facets, but omits other aspects such as the characteristics of users who posted tweets and their social relationships. Considering also *user information* in the cluster analysis would be very helpful to discover spatio-temporal patterns of communities of users and to better profile how the user interests evolve over time. As a future work, we will study an improvement of the TASTE measure with the aim of evaluating also data about users.

Finally, in this study we have applied the TCharM engine for the offline analysis of spatio-temporal-text information from tweets posted within a (relatively large) time window. As a future study, TCharM can be applied for the (near-)real time analysis, for instance of tweets collected every hour, to investigate the spatial evolution of clusters and related topics with a low time granularity. This approach would provide a deeper overview of the spatio-temporal dynamics of people's interests. Thanks to the deployment on a cloud-based platform as Apache Spark, TCharM can analyse huge amounts of data thus providing results in a reasonable time consistent with a near-real time analysis.

References

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *SIGMOD Conference* (pp. 207–216).

Arcaini, P., Bordogna, G., Ienco, D., & Sterlacchini, S. (2016). User-driven geo-temporal density-based exploration of periodic and not periodic events reported in social networks. *Inf. Sci.*, *340*, 122–143.

Baralis, E., Cerquitelli, T., Chiusano, S., Grimaudo, L., & Xiao, X. (2013). Analysis of twitter data using a multiple-level clustering strategy. In *Model and Data Engineering* (pp. 13–24). Springer.

Bernabe-Moreno, J., Tejeda-Lorente, A., Porcel, C., & Herrera-Viedma, E. (2015). A new model to quantify the impact of a topic in a location over time with social media. *Expert Systems with Applications*, *42*, 3381 – 3395.

Bouttier, J., Francesco, P. D., & Guitter, E. (2003). Geodesic distance in planar graphs. *Nuclear Physics B*, *663*, 535 – 567.

Bud (2014). Bud light on twitter. `http://t.co/Kj69El7MRE`.

Cagliero, L., Cerquitelli, T., Garza, P., & Grimaudo, L. (2014). Twitter data analysis by means of strong flipping generalized itemsets. *Journal of Systems and Software*, *94*, 16–29.

Capdevila, J., Pericacho, G., Torres, J., & Cerquides, J. (2016). Scaling dbscan-like algorithms for event detection systems in twitter. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10048 LNCS*, 356–373.

Cunha, T., Soares, C., & Mendes Rodrigues, E. (2014). Tweeprofiles: Detection of spatio-temporal patterns on twitter. In X. Luo, J. X. Yu, & Z. Li (Eds.), *Advanced Data Mining and Applications: 10th International Conference, ADMA 2014, Guilin, China, December 19-21, 2014. Proceedings* (pp. 123–136). Cham: Springer International Publishing.

Dasgupta, S. S., Natarajan, S., Kaipa, K. K., Bhattacherjee, S. K., & Viswanathan, A. (2015). Sentiment analysis of facebook data using hadoop based open source technologies. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 1–3).

Di Corso, E., Cerquitelli, T., & Ventura, F. (2017). Self-tuning techniques for large scale cluster analysis on textual data collections. In *Proceedings of the Symposium on Applied Computing* SAC '17 (pp. 771–776). New York, NY, USA: ACM.

Ester, M., Kriegel, H.-P., Sander, J., Wimmer, M., & Xu, X. (1998). Incremental clustering for mining in a data warehousing environment. In *Proceedings of the 24rd International Conference on Very Large Data Bases* VLDB '98 (pp. 323–333). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* KDD'96 (pp. 226–231). AAAI Press.

García-Gavilanes, R., Kaltenbrunner, A., Sáez-Trumper, D., Baeza-Yates, R., Aragón, P., & Laniado, D. (2014). Who are my audiences? a study of the evolution of target audiences in microblogs. In L. M. Aiello, & D. McFarland (Eds.), *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings* (pp. 561–572). Springer International Publishing.

Godfrey, D., Johns, C., Meyer, C., Race, S., & Sadek, C. (2014). A case study in text mining: Interpreting twitter data from world cup tweets. *arXiv preprint arXiv:1408.5427*, .

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *In SIGMOD'00, Dallas, TX*, .

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Ikeda, K., Hattori, G., Ono, C., Asoh, H., & Higashino, T. (2013). Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems*, *51*, 35 – 47.

Jenkins, T. (2014). World cup 2014: fans get bitten by the luis suarez bug - in pictures. `https://www.theguardian.com/football/gallery/2014/jun/26/world-cup-2014-luis-suarez-advert-bite`.

Jing, L., Ng, M. K., & Huang, J. Z. (2007). An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Transactions on Knowledge and Data Engineering*, *19*, 1026–1041.

Kim, J. W., Kim, D., Keegan, B., Kim, J. H., Kim, S., & Oh, A. (2015). Social media dynamics of global co-presence during the 2014 fifa world cup. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 2623–2632). ACM.

Kim, T., Huerta-Canepa, G., Park, J., Hyun, S. J., & Lee, D. (2011). What's happening: Finding spontaneous user clusters nearby using twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (pp. 806–809). IEEE.

Lee, C. H. (2012). Mining spatio-temporal information on microblogging streams using a density-based online clustering method. *Expert Systems with Applications*, *39*, 9623–9641.

Lee, R., Wakamiya, S., & Sumiya, K. (2015). Exploring geospatial cognition based on location-based social network sites. *World Wide Web*, *18*, 845–870.

Lloret, E., Balahur, A., Gómez, J. M., Montoyo, A., & Palomar, M. (2012). Towards a unified framework for opinion retrieval, mining and summarization. *Journal of Intelligent Information Systems*, *39*, 711–747.

Manning, C. D., Raghavan, P., Schütze, H. et al. (2008). *Introduction to information retrieval* volume 1. Cambridge university press Cambridge.

Pang-Ning T. and Steinbach M. and Kumar V. (2006). *Introduction to Data Mining*. Addison-Wesley.

Phelan, O., Mccarthy, K., & Smyth, B. (2009). Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems* (pp. 385–388). ACM.

Rabiger, S., & Spiliopoulou, M. (2015). A framework for validating the merit of properties that predict the influence of a twitter user. *Expert Systems with Applications*, *42*, 2824 – 2834.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*, 846–850.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation*, *60*, 503–520.

Saito, K., Ohara, K., Kimura, M., & Motoda, H. (2015). Change point detection for burst analysis from an observed information diffusion sequence of tweets. *Journal of Intelligent Information Systems*, *44*, 243–269.

Sakai, T., Tamura, K., Kotozaki, S., Hayashida, T., & Kitakami, H. (2015). Real-time local topic extraction using density-based adaptive spatiotemporal clustering for enhancing local situation awareness. In *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)* (pp. 203–210). volume 01.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, *18*, 613–620.

Steiger, E., Resch, B., & Zipf, A. (2016). Exploration of spatiotemporal and semantic clusters of twitter data using unsupervised neural networks. *Int. J. Geogr. Inf. Sci.*, *30*, 1694–1716.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. In *KDD Workshop on Text Mining*.

Thomas, K., Grier, C., Song, D., & Paxson, V. (2011). Suspended accounts in retrospect: An analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference* (pp. 243–258). ACM.

Vicient, C., & Moreno, A. (2015). Unsupervised topic discovery in micro-blogging networks. *Expert Systems with Applications*, *42*, 6472 – 6485.

Yang, M.-C., & Rim, H.-C. (2014). Identifying interesting twitter contents using topical analysis. *Expert Systems with Applications*, *41*, 4330 – 4336.

Yu, Y., & Wang, X. (2015). World cup 2014 in the twitter world: A big data analysis of sentiments in u.s. sports fans' tweets. *Computers in Human Behavior*, *48*, 392–400.

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing* HotCloud'10 (pp. 10–10). Berkeley, CA, USA: USENIX Association.