

Twitter data laid almost bare: An insightful exploratory analyser

Original

Twitter data laid almost bare: An insightful exploratory analyser / Xiao, Xin; Attanasio, Antonio; Chiusano, SILVIA ANNA; Cerquitelli, Tania. - In: EXPERT SYSTEMS WITH APPLICATIONS. - ISSN 0957-4174. - STAMPA. - 90:(2017), pp. 501-517. [10.1016/j.eswa.2017.08.017]

Availability:

This version is available at: 11583/2679661 since: 2021-04-07T18:41:05Z

Publisher:

Elsevier

Published

DOI:10.1016/j.eswa.2017.08.017

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

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

219 **1. Introduction**

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

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

246 Although a large body of research focused on Twitter data analysis has
247 already been proposed (e.g., (Phelan et al., 2009; Steiger et al., 2016)),
248 the potential impact of mining social data is still largely unexplored because
249 various critical issues are yet to be addressed when analyzing tons of tweets to
250 identify insightful nuggets. (i) Since a large number of tweets are continuously
251 being posted worldwide, the size of tweet collections to be explored grows
252 at an ever increasing rate. (ii) The collection of tweets generally tends to
253 be scattered in spatio-temporal dimensions, and the conciseness of the tweet
254 messages increases the brevity of their textual content (iii) Furthermore, the
255 distribution of tweets can be characterized by different spatial and temporal

256 granularities. (iv) Mined knowledge should be represented using concise and
257 understandable patterns to enable its exploitation by domain experts. Thus,
258 innovative data analytics solutions are needed to effectively and efficiently
259 mine large Twitter data collections.

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

266 TCHARM is based on two exploratory data mining techniques: (a) *Clus-*
267 *ter analysis*, to identify cohesive groups of tweets with similar text con-
268 tent posted from nearby geographical areas and at close time instances, and
269 (b) *Association rule analysis*, to find significant patterns that concisely de-
270 scribe each computed cluster. To make the proposed methodology scale up
271 to larger datasets, TCHARM exploits the computational advantages of dis-
272 tributed computing frameworks since the current implementation runs on
273 Apache Spark (Zaharia et al., 2010).

274 Unlike previous works (e.g.,(Kim et al., 2011; Lee, 2012; Cunha et al.,
275 2014; Arcaini et al., 2016)), TCharM drives the clustering process by making
276 joint use of the tweet spatio-temporal features and text content. A novel Text
277 And Spatio-TEmporal distance measure, denoted by TASTE, is proposed in
278 this study in order to combine the contributions of all three tweet features in
279 one step. Through TASTE, spatial and temporal distances between tweets
280 are used to modulate the text content distance. By taking into account
281 both spatio-temporal features and text content in the clustering of tweets,
282 TCHARM findings can provide useful insights to identify the users topics
283 of interest in different areas and time periods. For instance, events such as
284 sports, culture and politics, which have widespread visibility, can be useful
285 to understand topics that are popular in different geographical areas. The
286 information provided by the spatio-temporal distribution of such clusters may
287 help characterize peoples involvement in different time frames. TCHARM has
288 been currently integrated into the *K-means* clustering algorithm (Pang-Ning
289 T. and Steinbach M. and Kumar V., 2006), to generate clusters of tweets
290 that can be concisely represented by their centroids.

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

294 This task has been carried out using association rule analysis (Pang-Ning
295 T. and Steinbach M. and Kumar V., 2006), an exploratory data mining
296 technique to extract correlations among data items. Quality indices (e.g.,
297 confidence, support, and lift) are used to distinguish the most significant
298 correlations. Association rule analysis allows the extraction of the most re-
299 current spatio-temporal-text patterns in a systematic and structured way.
300 These patterns describe the cluster content using a concise and clear knowl-
301 edge representation. To further support the exploration of discovered pat-
302 terns, four different classes of association rules have been defined. By “class”
303 we mean a subset of patterns which determines significant relationships be-
304 tween tweet dimensions which can be used to perform a similar in-depth
305 analysis. The identified patterns can provide domain experts with valuable
306 support to identify which topics are most appealing to users in different areas
307 and time periods.

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

324 The rest of the paper is organized as follows: Section 2 summarizes the
325 related work regarding cluster analysis of Twitter data. Section 3 provides
326 an in-depth description of the TCHARM characteristics, while Section 4 dis-
327 cusses the experimental study conducted on the 2014 FIFA World Cup Cham-
328 pionship dataset. Section 5 provides a theoretical and analytical comparison
329 between TCHARM and some previous works on tweet clustering. Section
330 6 discusses the significance of TCharM findings and their possible exploita-
331 tion. Section 7 draws conclusions and future developments of the proposed

332 approach.

333 **2. Related Work**

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

339 The first category refers to methods addressing the analysis of tweet tex-
340 tual content with the aim of (i) characterizing online communities (Rabiger
341 & Spiliopoulou, 2015), (ii) performing spam detection (Thomas et al., 2011),
342 (iii) detecting topics to analyse trends (Baralis et al., 2013; Vicient & Moreno,
343 2015; Yang & Rim, 2014), and (iv) addressing recommendation tasks (Phelan
344 et al., 2009).

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

355 Various approaches have been proposed to cluster tweets collections tak-
356 ing into account textual content and spatio-temporal information (Kim et al.,
357 2011; Steiger et al., 2016), though such works do not jointly exploit all these
358 features in the clustering process. Instead, they typically use a subset of
359 features for clustering, while remaining features are considered either in the
360 post-processing phase, for instance to refine or characterize discovered clus-
361 ters, or in the preprocessing phase, for example to specify spatial or temporal
362 segments in which tweets are locally clustered based on textual content. Kim
363 et al. (2011) cluster tweets based on their GPS coordinates using the K-means
364 algorithm, while Steiger et al. (2016) use a spatio-temporal clustering based
365 on Self Organizing Maps (SOM). In both approaches, discovered clusters are
366 then analysed to identify the main targeted topic. Density based clustering,
367 mainly based on the DBSCAN algorithm, has been also adopted to detect

368 high spatial concentrations or temporal bursts of tweets about specific top-
369 ics (Arcaini et al., 2016; Lee, 2012; Lee et al., 2015; Sakai et al., 2015). For
370 instance, Lee et al. (2015) group user trajectories derived from geo-tagged
371 tweets and explore massive crowd movements, while Sakai et al. (2015) ex-
372 tract local bursty keywords and identify their dense areas to enhance local
373 situation awareness.

374 **Differently from all the works above, the TCHARM framework**
375 ***jointly exploits the spatio-temporal features and tweet textual con-***
376 ***tent to drive the clustering process. Our main purpose is to dis-***
377 ***cover cohesive clusters focused on single topics and, at the same***
378 ***time, with precise spatio-temporal references. Through the TASTE***
379 ***distance measure, TCHARM explores the three dimensions charac-***
380 ***terizing tweets, to discover, in one step, groups of messages with***
381 ***similar content but posted in nearby time and space.***

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

391 In this study, the TCHARM framework has been deployed on Apache
392 Spark. Several open source data mining platforms, like Scikit-learn, Rapid-
393 Miner, Apache Mahout and Apache Spark have proposed their own scala-
394 bility strategies to analyse the huge and rapidly growing amount of data.
395 Such platforms include libraries implementing common machine learning al-
396 gorithms which can be extended or modified by researchers. The adoption of
397 Apache Spark in many research works (including but not limited to tweets) is
398 mainly motivated by both the support for stream analysis (Dasgupta et al.,
399 2015) and the scalable computing framework that makes it possible to speed
400 up existing algorithms for different applications (Capdevila et al., 2016).

401 Tweets about the 2014 FIFA World Cup has been considered as a refer-
402 ence case study for the validation of the proposed framework. Various studies
403 have addressed the analysis of tweets related to this event, with different tar-
404 geted analyses devoted to (i) performing sentiment analysis to characterize
405 U.S. soccer fans' emotional responses (Yu & Wang, 2015); (ii) addressing

406 topic detection through a combined approach based on the DBSCAN algo-
407 rithm and Non-Negative matrix (Godfrey et al., 2014); (iii) tracking user
408 behavior through Latent Dirichlet Allocation (LDA) (Kim et al., 2015). All
409 these approaches analyse the textual content only, while TCHARM clusters
410 the tweet collection besides characterizing the cluster content based on tex-
411 tual and spatio-temporal dimensions.

412 3. TCHARM architecture

413 The main components of the Tweets CHARACTERIZATION Methodology
414 (TCHARM) architecture are shown in Figure 1. The components are briefly
415 introduced below while a more thorough description of each of them is given
416 in the following subsections.

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

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

432 Finally, TCHARM analyses each discovered cluster to mine a set of pat-
433 terns describing the cluster content. Specifically, through *association rule*
434 *analysis*, patterns of relevant correlations among tweets text contents, post-
435 ing times and geographical areas are extracted for each cluster. Extracted
436 rules are then categorized into four classes defined according to the types of
437 correlation among the tweets attributes while, to ease their semantic inter-
438 pretation, the same rules are associated with one of the few reference topic
439 families according to the word set they contain.

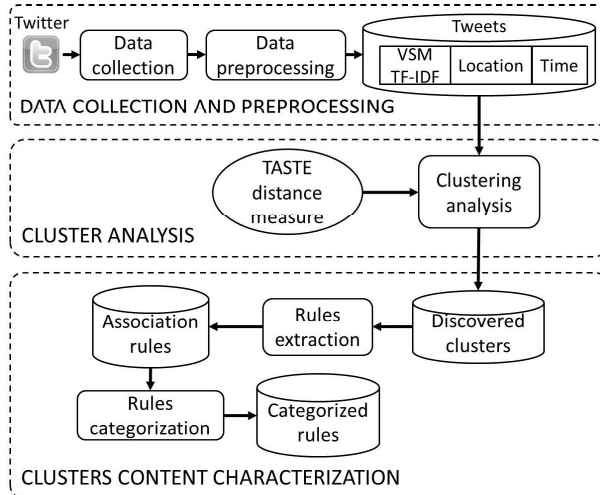


Figure 1: The TCHARM architecture

440 3.1. Twitter data representation

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

447 **Tweet text content.** Tweets are posts published by Twitter users that
448 include also *text messages* 140 characters long at most. Such messages rep-
449 resent the text content used in our analysis. Due to the limited size of the
450 single message and to the high dimensionality of many text content represen-
451 tations, the represented samples are inherently sparse. This property leads
452 to higher levels of noise in the tweet collection, thus adding complexity to
453 the clustering process, which requires an adequate treatment (Jing et al.,
454 2007). Moreover, Twitter messages are usually extremely impure because
455 they include a wide variety of Unicode data, symbols, numbers and links.
456 They therefore need to be properly cleaned and prepared before the analysis.

457 **Tweet temporal feature** corresponds to the *timestamp* including date and
458 time instant when the tweet was posted. In this study, we omit the temporal
459 information possibly appearing in the tweet message, since it is considered
460 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

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

474 Table 1 reports an example tweet including the three features. The tweet
475 refers to the 2014 FIFA World Cup, considered as a reference case study in
476 this paper. The tweet was posted on Friday morning, June 20th 2014, at
477 9:26 a.m. from Banbury City (UK), according to geo-coordinate values.

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

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

491 *is the tweet text content.*

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

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

511 **Definition 3.2 (Tweet text content representation).** Let $\tau_i =$
512 (t_i, s_i, W_i) be an arbitrary tweet in collection \mathcal{D} . The tweet text con-
513 tent W_i is a vector of k elements corresponding to words in Σ (i.e., $k = |\Sigma|$).
514 Each vector element $W_i[j]$ contains the TF-IDF weight of word w_j for
515 tweet τ_i . $W_i[j]$ is computed as $W_i[j] = TF(\tau_i, w_j) \cdot IDF(w_j)$, where terms
516 $TF(\tau_i, w_j)$ and $IDF(w_j)$ are defined as follows:

- 517 1. $TF(\tau_i, w_j)$ is the relative frequency of word w_j for tweet τ_i .
518 $TF(\tau_i, w_j) = f(\tau_i, w_j) / \sum_{l=1}^k f(\tau_i, w_l)$, where $f(\tau_i, w_j)$ is the number
519 of times word w_j appeared in tweet τ_i and $\sum_{l=1}^k f(\tau_i, w_l)$ is the total
520 number of words contained in τ_i .
- 521 2. $IDF(w_j)$ is the relative frequency of word w_j in \mathcal{D} . $IDF(w_j) =$
522 $\log(|\mathcal{D}|/|\mathcal{D}_j|)$ where $|\mathcal{D}|$ is the number of tweets in \mathcal{D} and $|\mathcal{D}_j|$, $\mathcal{D}_j =$
523 $\{\tau_i \in \mathcal{D} : f(\tau_i, w_j) > 0\} \subseteq \mathcal{D}$, is the number of tweets in \mathcal{D} which
524 contain (at least once) word w_j .

525 Mathematically, the base of the *log* function for IDF computation in
526 Definition 3.2 does not influence the overall results as it constitutes a constant
527 multiplicative factor (Robertson, 2004). The TF-IDF weight $W_i[j]$ for word
528 w_j in tweet τ_i is high when w_j appears with high frequency in tweet τ_i but
529 low frequency in tweets in the collection \mathcal{D} . When word w_j appears in more
530 tweets, the ratio inside the IDF *log* function approaches 1, and both the
531 $IDF(w_j)$ value and the TF-IDF weight $W_i[j]$ become close to 0. Hence, the
532 approach tends to filter out common words. In short messages like tweets, the
533 TF-IDF weighting score could actually be reduced to a pure IDF scheme due
534 to the limited word frequency within each tweet. Nevertheless, we preserved
535 the TF-IDF approach to consider also possible word repetitions.

536 3.2. Twitter data collection and preprocessing

537 Twitter data for the TCHARM framework are retrieved through the Twit-
538 ter Stream Application Programming Interfaces (APIs) by specifying a set of
539 filter parameters (e.g., keywords, hashtags). Collected data include all infor-
540 mation characterizing tweets useful for the subsequent data analysis phase,
541 i.e., tweet message, publication time and geographical location of the user.
542 Of the tweets collected, only those in English are considered.

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

550 3.3. Cluster analysis of tweets

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

557 In TCHARM, the K-means algorithm is used for clustering tweet data
558 collections. K-means has been widely used in different applications domains,
559 including tweets analysis, providing good quality solutions. The K-means

560 algorithm discovers K clusters modeled by their representatives, named *cen-*
561 *troids*, given by the mean value of the objects in the clusters. Initially, K
562 tweets of the tweet collection \mathcal{D} are randomly chosen as centroids. Then each
563 tweet $\tau_i \in \mathcal{D}$ is assigned to the cluster of the nearest centroid. Finally, the
564 centroids are relocated by computing the mean of the tweets features within
565 each cluster. The process iterates until a convergence criterion is met, i.e.,
566 the centroids do not change or some objective functions are achieved.

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

572 3.3.1. The TASTE distance measure

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

575 **Definition 3.3 (TASTE distance measure).** Let $\tau_i = (t_i, s_i, W_i)$ and
576 $\tau_j = (t_j, s_j, W_j)$ be two arbitrary tweets in collection \mathcal{D} . The TASTE dis-
577 tance measure between tweets τ_i and τ_j is defined as

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

578 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
579 $d_W(W_i, W_j)$, $d_s(s_i, s_j)$, and $d_t(t_i, t_j)$ measure the distance on tweet text con-
580 tent, spatial feature, and temporal feature, respectively. These distances
581 have been normalized in the range $[0, 1]$ using the *min-max* normalization
582 method (Pang-Ning T. and Steinbach M. and Kumar V., 2006).

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

586 In the TASTE measure, spatial and temporal distances ($d_s(s_i, s_j)$ and
587 $d_t(t_i, t_j)$) modulate the text content distance ($d_W(W_i, W_j)$) to determine the
588 overall value of $d_{TASTE}(\tau_i, \tau_j)$. The exponential form is used for $d_s(s_i, s_j)$
589 and $d_t(t_i, t_j)$ to significantly penalize tweets with a large space and/or time
590 distance.

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

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

600 In TASTE, three different measures are used to compute $d_W(W_i, W_j)$,
 601 $d_s(s_i, s_j)$, and $d_t(t_i, t_j)$ based on the data type describing tweet text content,
 602 spatial feature and temporal feature.

603 **Text content distance measure ($d_W(W_i, W_j)$).** The distance between
 604 the weighted word frequency vectors W_i and W_j of tweets τ_i and τ_j is evalu-
 605 ated using the *cosine distance measure* (Pang-Ning T. and Steinbach M. and
 606 Kumar V., 2006), which has often been used to compare documents in text
 607 mining (Steinbach et al., 2000). We define the text content distance measure
 608 $d_W(W_i, W_j)$ as

$$d_W(W_i, W_j) = \arccos(\cos(W_i, W_j)). \quad (2)$$

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

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

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

612 The value range is $[0, 1]$ for the cosine similarity $\cos(W_i, W_j)$, while the
 613 value range for the content distance measure $d_W(W_i, W_j)$ is $[0, \pi/2]$. When
 614 $\cos(W_i, W_j) = 1$, then $d_W(W_i, W_j) = 0$ which describes the exact similarity of
 615 text content for tweets τ_i and τ_j . When $\cos(W_i, W_j) = 0$, then $d_W(W_i, W_j) =$
 616 $\pi/2$ which points out that tweets τ_i and τ_j have completely different texts.

617 **Temporal distance measure ($d_t(t_i, t_j)$).** The tweet temporal feature is
 618 an integer number representing the time instant when the tweet was posted.
 619 The *Euclidean distance* (Pang-Ning T. and Steinbach M. and Kumar V.,
 620 2006) is adopted here as the distance on temporal features t_i and t_j of tweets

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

$$d_t(t_i, t_j) = |t_i - t_j|. \quad (4)$$

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

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

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

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

635 The content, spatial and temporal distance measures defined above satisfy
 636 the positivity, symmetry, and triangle inequality properties that characterize
 637 a metric (Pang-Ning T. and Steinbach M. and Kumar V., 2006). It easily
 638 follows that the TASTE measure also verifies these properties. Specifically,
 639 the following properties hold. (i) *Positivity*: $d_{TASTE}(\tau_i, \tau_j) \geq 0$ for all $\tau_j, \tau_i \in$
 640 \mathcal{D} , while $d_{TASTE}(\tau_i, \tau_j) = 0$ only if $\tau_i = \tau_j$. (ii) *Symmetry*: $d_{TASTE}(\tau_i, \tau_j) =$
 641 $d_{TASTE}(\tau_j, \tau_i)$ for all $\tau_j, \tau_i \in \mathcal{D}$. (iii) *Triangle inequality*: $d_{TASTE}(\tau_i, \tau_j) \leq$
 642 $d_{TASTE}(\tau_i, \tau_k) + d_{TASTE}(\tau_k, \tau_j)$ for all $\tau_i, \tau_k, \tau_j \in \mathcal{D}$.

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

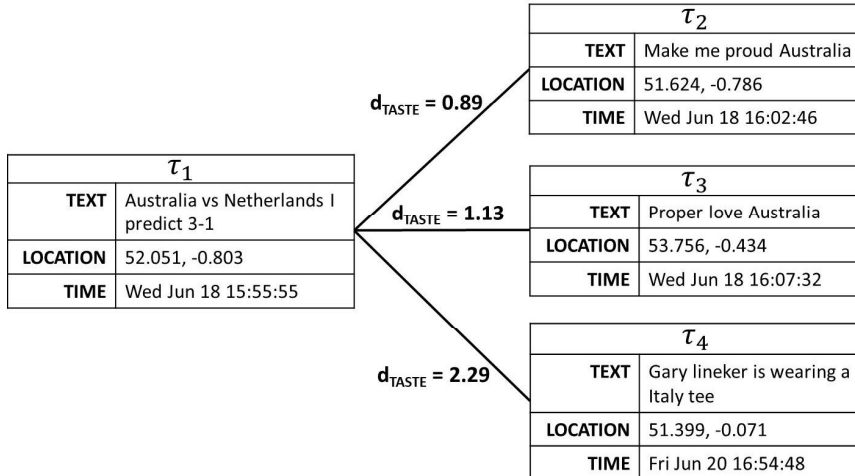


Figure 2: Sample tweets about 2014 FIFA World Cup with TASTE distance values

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

658 3.3.2. Clustering Evaluation

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

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

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

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

670 3.4. Clusters content characterization

671 After the cluster set is generated, in TCHARM each cluster is then locally
672 explored to characterize its content. Specifically, each cluster is analysed to
673 discover underlying correlations in the text content, and between text content
674 and the spatial and temporal features characterizing tweets. Cluster charac-
675 terization makes use of *association rules* as reference pattern type (Agrawal
676 et al., 1993).

677 3.4.1. Association rules extraction

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

682 Consider an arbitrary cluster C included in the cluster set computed on
683 tweet collection \mathcal{D} . The *transactional tweet dataset* $\mathcal{D}_{\mathcal{T}}(C)$ for cluster C is a
684 set of transactions. Each *transaction* \mathcal{T}_i corresponds to a tweet $\tau_i \in C$ and
685 it consists of a set of tweet features called *items*, represented in the form
686 $\{\textit{attribute} : \textit{value}\}$. The items of the generic transaction \mathcal{T}_i are (i) each
687 single *word* $w \in W_i$ appearing in the text content of tweet τ_i , (ii) the value
688 of the *spatial feature* s_i of τ_i , and (iii) the value of the *temporal feature* t_i of
689 τ_i .

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

701 Consider, for example, association rule $r : \{\textit{start}, \textit{world}, \textit{cup}\} \Rightarrow \{\textit{love}\}$
702 (*supp* = 1.1%, *conf* = 60%) mined from cluster C . Rule r talks about people’s
703 feelings on the World Cup game. The rule represents relationships that

704 emerge from tweets messages contained in C , i.e., the correlation between
705 subset of words included in these messages. According to the rule support
706 and confidence values, 1.1% of tweets in cluster C contain all the words
707 appearing in the rule (i.e., *start*, *world*, *cup* and *love*), but the word *love*
708 appears in 60% of tweets including the words *start*, *world* and *cup*.

709 In some cases, measuring the strength of a rule in terms of support and
710 confidence values may be misleading. When the rule consequent has a high
711 support value, the rule may be characterized by a high confidence value
712 even if its actual strength is relatively low. To overcome this issue, the
713 *lift* (or correlation) index (Pang-Ning T. and Steinbach M. and Kumar V.,
714 2006) may be used, beyond the confidence index, to measure the (symmetric)
715 correlation between sets X and Y . Lift values below 1 show a negative
716 correlation between sets X and Y , while values above 1 indicate a positive
717 correlation. In this study, to mine patterns representing strong correlations
718 among features characterizing tweets, the selection of association rules is
719 based on confidence and lift values.

720 3.4.2. Association rule categorization

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

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

- 740 2. *Location-TextContent class (L-TC)*. This class analyses the correlations
741 between the words in tweet messages and the locations where tweets
742 have been posted. It makes it possible to identify the topics attract-
743 ing/involving users in a given location.
- 744 3. *Time-TextContent class (T-TC)*. This class analyses the correlation
745 between words in tweet messages and the time when tweets have been
746 posted so as to discover the topics attracting/involving users in a given
747 time frame.
- 748 4. *Location-Time-TextContent class (L-T-TC)*. This class considers all the
749 properties characterizing tweets in order to analyse the correlation be-
750 tween the words in tweet messages together with the time when, and
751 the location where, the tweets were posted. It makes it possible to dis-
752 cover the topics attracting/involving users in a given time frame and
753 location.

Class	Example question	Example pattern	
		Association Rule	Meaning
TC	What are the topics attracting/involving users?	$\{\text{world,final}\} \Rightarrow \{\text{cup}\}$ $\text{centroid}(T = y, L = x)$	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?	$\{L = x\} \Rightarrow$ $\{\text{TC} = (\text{german, win, argentina})\}$	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?	$\{T = y\} \Rightarrow$ $\{\text{TC} = (\text{best, player, playerName})\}$	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?	$\{T = y, L = x\} \Rightarrow$ $\{\text{TC} = (\text{good, performance, playerName})\}$	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

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

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

774 3.5. TCHARM implementation

775 The entire data analysis process (preprocessing, clustering, and associ-
776 ation rules extraction) in TCHARM has been implemented as a Scala ap-
777 plication in the open source computing framework *Apache Spark* (version
778 1.5) (Zaharia et al., 2010). This framework was selected because it is cur-
779 rently one of the leading platforms for data analytics and provides a Machine
780 Learning library (MLlib) which has been exploited and extended in this study
781 to support all the functionalities of TCHARM.

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

793 The preliminary data collection step relies on Twitter’s Streaming Appli-
794 cation Programming Interfaces (APIs) to retrieve tweets data. The Stream-
795 ing APIs provide low latency access to Twitter’s global stream of tweets data
796 by establishing and maintaining a continuous connection with the stream
797 endpoint. A Java crawler is used to collect and parse tweets in real time
798 based on a predefined set of keywords (e.g., “worldcup2014”, “fifaworldcup”
799 in our case study), with a case-insensitive search.

800 4. Experimental Results

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

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

814 4.1. Datasets

815 The public stream endpoint offered by the Twitter APIs was monitored
816 over a time period of 27 days from June 18th to July 14th 2014, by tracking

817 a selection of keywords related to the 2014 FIFA World Cup (e.g., “world-
818 cup2014”, “fifaworldcup”). Tweets in English and with the exact GPS coor-
819 dinates of the user location were extracted. The resulting collection includes
820 302,052 tweets. To ease the computation of temporal distances between
821 tweets in the clustering phase, all timestamps have been converted according
822 to the reference time zone of *America/Sao Paulo*, in Brazil, where the 2014
823 FIFA World Cup was held.

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

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

835 The tweet spatial distribution was then locally analysed within each of
836 the three time windows based on tweet geo-coordinates. In each time window
837 tweets appeared to be widely dispersed and geographically partitioned into
838 different areas. English speaking countries like the United Kingdom (UK),
839 USA, and Central America show higher tweets concentrations than other
840 areas. Following this evaluation of tweet spatial distribution, we selected
841 two *spatial partitions*, corresponding to *UK* and *USA*, for each time window.
842 Table 4 summarizes the main characteristics of the six resulting datasets
843 which are used as reference case studies for the experimental evaluation.
844 Each dataset was named using the corresponding spatio-temporal segment.
845 For example, dataset $\mathcal{D}_{(TW1,UK)}$ contains tweets posted during time window
846 #1 in UK.

847 4.2. Parameters configuration for cluster analysis

848 We set the parameters for the clustering analysis to best fit the use case
849 considered, the 2014 FIFA World Cup, which involves people worldwide.
850 Aimed at discovering clusters including tweets about the same topics but
851 posted in nearby locations and time periods, we assigned the same relevance
852 to spatial and temporal terms in modulating the text distance, i.e., we set
853 $k_s = k_t = 0.5$. 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

854 most reactions to a given event (e.g., a football match) to be published as soon
 855 as the same event occurs (or within a short delay), even from quite distant
 856 locations. Indeed, while users interested in the same event can be also located
 857 in different areas, it is unlikely that they tweet at completely different times.
 858 Therefore, to group tweets with very close temporal distances, we set the
 859 weight of the temporal exponent p_t to a higher value than the spatial one p_s .
 860 We empirically found that $p_s = 3$ and $p_t = 6$ provide the lowest variability of
 861 SSE among clusters for different values of K (number of clusters) on datasets
 862 in Table 4. For each dataset, we evaluated the average SSE among the
 863 resulting clusters for a range of values of K . K was then set to 200 as a
 864 good trade-off to minimize SSE and to limit the number of clusters as well.
 865 As an example, Figure 3 plots the decrease of the average SSE for dataset
 866 $\mathcal{D}_{(TW1,UK)}$ when increasing the value of K . SSE abruptly decreases until
 867 $K = 150$, after which it goes down at a lower rate. Since we needed to
 868 limit both the desired number of clusters and the expected value of SSE, we
 869 assumed that $K = 200$ was a good trade off between these two objectives.

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

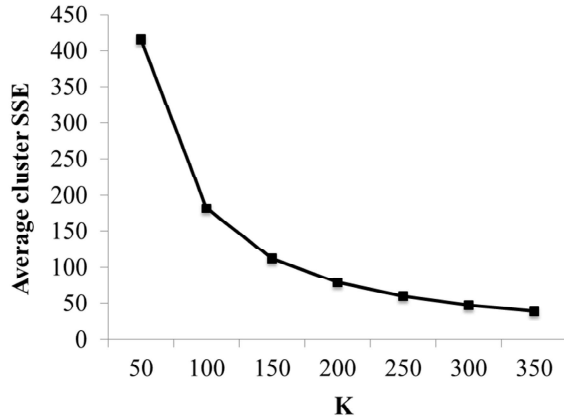


Figure 3: Variation of average cluster SSE with respect to the number of clusters (K) for dataset $\mathcal{D}_{(TW1,UK)}$ ($p_s = 3$, $p_t = 6$, $k_s = k_t = 0.5$)

874 *4.3. Analysis of the clustering results*

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

885 Figure 4 shows the distribution of clusters cardinality in the cluster set
886 computed on dataset $\mathcal{D}_{(TW1,UK)}$. Clusters are sorted along the x axis by
887 increasing value of cardinality. The cluster set includes one cluster with
888 about 800 tweets, while 16.5% of clusters contain from 200 to 400 tweets,
889 41.5% of clusters from 100 to 200 tweets, and the remaining 41.5% less than
890 100 tweets. The mean value of cluster size is 132 tweets, while the median
891 value is 111 tweets.

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

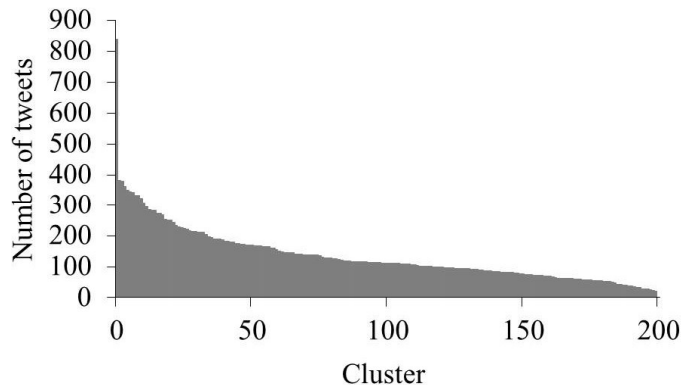


Figure 4: Distribution of number of tweets in the cluster set for dataset $\mathcal{D}_{(TW1,UK)}$

896 is adopted in place of the original one. Specifically, the *spatial feature*
 897 is represented as the *geographical area* where a centroid is located, instead of its
 898 GPS coordinates. Since the considered dataset contains tweets posted in UK,
 899 the *county* is used here as reference geographical area. County membership
 900 of a centroid is calculated based on the boundary GPS coordinates of each
 901 county¹ and on the GPS coordinates of the centroid. The *temporal feature*
 902 of a centroid is represented in terms of the corresponding *hourly time slot*,
 903 instead of the centroid timestamp.

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

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

¹<http://www.nearby.org.uk/downloads.html>

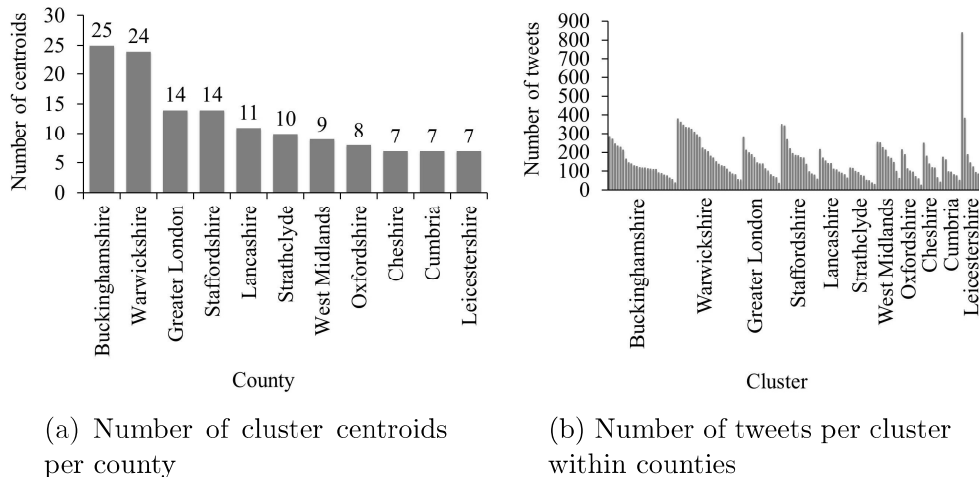


Figure 5: Spatial characterization of the cluster set for dataset $\mathcal{D}_{(TW1,UK)}$

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

921 The evaluation of the *temporal distribution* of centroids in the cluster
 922 set reveals the *time periods* when people in UK were more involved in the
 923 2014 FIFA World Cup. As an example, we report the results for a two-
 924 day time frame (from June 19th to June 20th) within time window #1.
 925 Figure 6a shows the number of centroids located in each hourly time slot,
 926 while Figure 6b reports the cardinality of the corresponding clusters. For
 927 each hourly time slot, clusters are sorted along the x axis by decreasing
 928 value of cardinality.

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

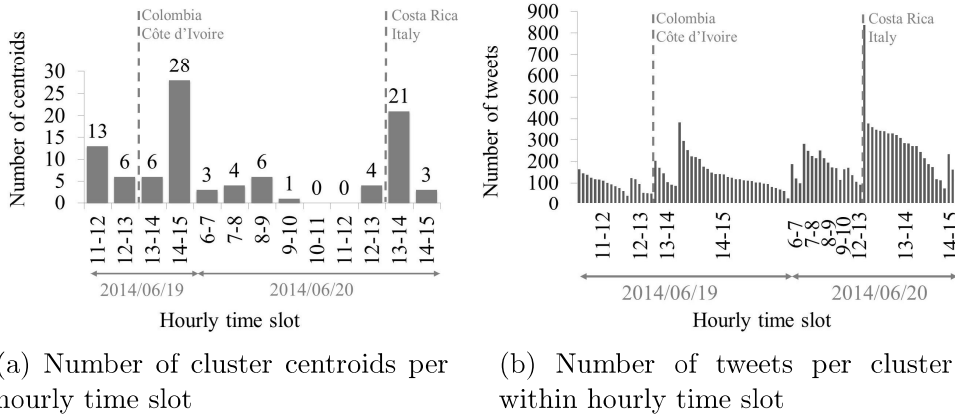


Figure 6: Temporal characterization of the cluster set for dataset $\mathcal{D}_{(TW1,UK)}$ during a two-days time frame

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

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

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

959 Clusters also demonstrate a reasonable spatial cohesion around their cen-
 960 troid, since tweets within each cluster are mainly (or even exclusively) posted

Cluster ID	Cluster centroid		Cluster content		
	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

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

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

971 4.4. Clusters characterization using association rules

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

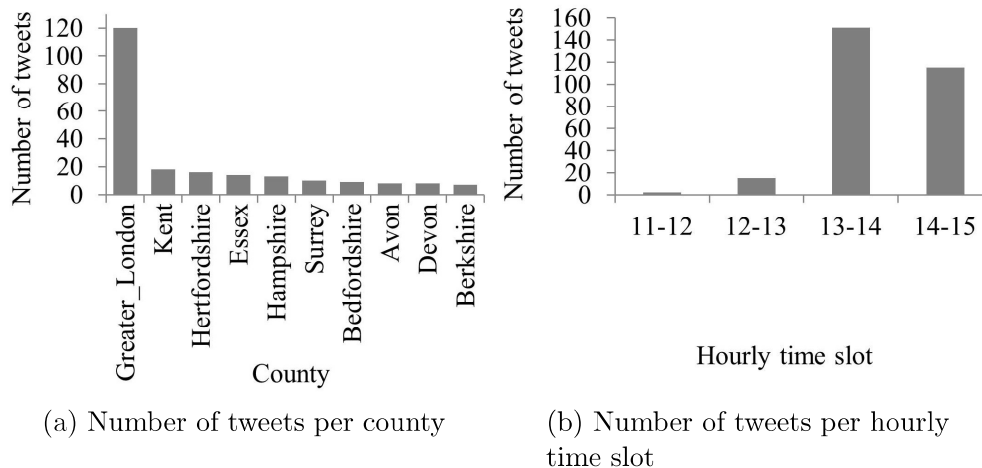


Figure 7: Spatial and temporal characterization of cluster C (from Table 5)

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

983 4.4.1. Analysis of rules on a sample cluster

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

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

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

1000 Rica football team. Instead, in rule R_2 people talked about the celebrity
1001 Gary Lineker, a retired English footballer and current sports broadcaster,
1002 who wore an Italy shirt. The reason is that the victory of Italy over Costa
1003 Rica would have allowed the England football team to keep their World Cup
1004 hopes alive.

1005 *Analysis of correlations between the location where tweets were posted and*
1006 *tweets text content (class L-TC).* Rules in the L-TC class, like rules R_3 and
1007 R_4 , point out the geographical areas where certain topics are discussed. Rule
1008 R_3 reveals that a negative emotional state about the England football team
1009 arises from people located in the Greater London county. This opinion may
1010 be due to the fact that the England football team did not win any match
1011 in the first stage of the World Cup. Instead, rule R_4 reports that people in
1012 the Greater London county are watching how the Costa Rica football team
1013 performs in matches.

1014 *Analysis of correlations between the time when tweets were posted and tweets*
1015 *text content (class T-TC).* Rules in the T-TC class, such as R_5 and R_6 ,
1016 point out the time slot when certain topics are discussed. Rule R_5 describes
1017 the association between people’s disappointment about the behavior of the
1018 Italian football team and the time slot including the football match *Italy -*
1019 *Costa Rica*. In fact, after the goal scored by Costa Rica in the first half of
1020 the match, the Italian team did not respond with any winning actions in
1021 the second half of the match. Rule R_6 highlights the people’s interest in
1022 the comments on the *Italy - Costa Rica* match by a former English player
1023 (Robbie Savage) hired as pundit by the British Broadcasting Corporation
1024 (BBC) for the 2014 FIFA World Cup.

1025 *Analysis of correlations between location where, and time when, tweets were*
1026 *posted and tweet text content (class L-T-TC).* R_7 and R_8 are example rules
1027 belonging to this class and they both show that the goal scored by Costa
1028 Rica and the consequent defeat of the Italian team in the time slot including
1029 the first half of the match was a hot topic in the Greater London county.

1030 It is worth noting that the rules of classes *L-TC*, *T-TC* and *L-T-TC*,
1031 characterized by positive correlation and high confidence values, always in-
1032 clude the same county and hourly time slot of the centroid. This provides
1033 further evidence in support of the high spatio-temporal cohesion of cluster C
1034 around its centroid.

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

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

1035 4.4.2. Analysis of rules across geographical partitions

1036 In this section we analyse how people’s interest in events occurring within
1037 a given time window vary across different geographical areas. We compared
1038 the association rules mined from clusters computed in UK and USA areas
1039 when considering time window #1 (datasets $\mathcal{D}_{(TW1,UK)}$ and $\mathcal{D}_{(TW1,USA)}$). To
1040 reduce data fragmentation in the mined patterns, we adopted a coarse spatio-
1041 temporal data representation suitable for both cases considered. Specifically,
1042 tweet geo-coordinates have been mapped to the nearest city and the tweet
1043 posting timestamp to the corresponding day. Some sample rules modeling
1044 correlations in the tweet text content (class TC) are shown in Table 7, but
1045 the following discussion is based on the overall results.

1046 People in the UK area commented mostly on matches involving the Eng-
1047 land football team (e.g., rule R_1), or other teams included in the same group
1048 as England. Moreover, an odd episode involving a single player was the main
1049 topic of various clusters (R_2). Instead, clusters from many locations of the
1050 USA reveal that people were interested in matches involving various football
1051 teams, also those not included in the same group as their national team. For
1052 instance, rule R_3 refers to the match between Italy and Costa Rica and rule
1053 R_4 to the match involving Nigeria and Argentina.

1054 The behaviour observed may be related to the people’s different interests
1055 in the two geographical areas. Overall, football is more popular in England
1056 than in USA, where people are mostly interested in other sports. While in
1057 England people particularly focus on events related to their national team,
1058 in USA they show a more general interest in the FIFA World Cup, also for
1059 events involving teams other than their national team.

1060 4.4.3. Analysis of rules across time windows

1061 In this section we analyse how the interests of people tweeting from the
1062 same geographical area vary for events that occurred in different time win-
1063 dows. We compared rules mined from clusters computed in the UK area
1064 in the three time windows (datasets $\mathcal{D}_{(TW1,UK)}$, $\mathcal{D}_{(TW2,UK)}$, and $\mathcal{D}_{(TW3,UK)}$).
1065 We adopted the same spatio-temporal data representation used for the anal-
1066 ysis discussed in Section 4.4.2. Table 8 shows some example rules from the
1067 TC class, but the discussion is based on the overall results.

1068 It is worth noting how interests varied after the elimination of Eng-
1069 land team which happened at the end of time window #1. The extracted
1070 rules show that people in UK shifted their attention to matches involv-
1071 ing 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	$\{\text{uruguay}\} \Rightarrow \{\text{england}\}$ $\text{centroid}(T = 2014-06-19,$ $L = \text{Perth})$	5.0	100	2.38
R_2	UK	Celebrity	$\{\text{suarez, someone}\} \Rightarrow \{\text{bite}\}$ $\text{centroid}(T = 2014-06-25,$ $L = \text{Rugeley})$	3.0	80	26.90
R_3	USA	Event	$\{\text{costa, rica}\} \Rightarrow \{\text{italy}\}$ $\text{centroid}(T = 2014-06-20,$ $L = \text{Whittier, CA})$	8.3	64	1.67
R_4	USA	Event	$\{\text{nigeria}\} \Rightarrow \{\text{argentina}\}$ $\text{centroid}(T = 2014-06-25,$ $L = \text{Banning, CA})$	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 $\mathcal{D}_{(TW1,USA)}$)

1072 *Germany – Algeria* football match (played on June 30th, 2014), and are
1073 mostly about the tactics (R_5) and performance (R_6) of the German team.

1074 During time window #3, the final match became one of the most popular
1075 topics (R_7). Nevertheless, the attention of people in UK also moved towards
1076 other topics loosely related to the competition. For instance, the latest trans-
1077 fer of player Luis Suarez away from an English club was mainly discussed on
1078 July 11th 2014, on the same day as the official announcement (R_8), while the
1079 next match of the England team, scheduled for November against Scotland
1080 (R_9), became popular just after the final World Cup match, on July 14th
1081 2014.

1082 4.5. Execution time and scalability

1083 The execution time for the cluster set computation on the six datasets
1084 in Table 4 spans from 12m 13s for the smallest dataset ($\mathcal{D}_{(TW2,UK)}$, 15,175
1085 tweets) up to 33m 34s for the largest one ($\mathcal{D}_{(TW3,USA)}$, 50,028 tweets). The
1086 execution time for association rules extraction is less variable and has an
1087 overall mean value of 53s. Increasing the number of executors does not yield
1088 better performance in terms of clustering execution time due to the limited
1089 size of these datasets. Thus, experiments for these datasets were performed
1090 using one execution node.

Rule id	Time window	Topic description	Rule	supp [%]	conf [%]	lift
R_1	1	Event	$\{\text{uruguay}\} \Rightarrow \{\text{england}\}$ <i>centroid</i> (T = 2014-06-19, L = Perth)	5.0	100	2.38
R_2	1	Celebrity	$\{\text{suarez, someone}\} \Rightarrow \{\text{bite}\}$ <i>centroid</i> (T = 2014-06-25, L = Rugeley)	3.0	80	26.90
R_5	2	Event	$\{\text{line,high}\} \Rightarrow \{\text{germany}\}$ <i>centroid</i> (T = 2014-06-30, L = London)	2.0	100	1.02
R_6	2	Emotional state	$\{\text{good}\} \Rightarrow \{\text{germany}\}$ <i>centroid</i> (T = 2014-06-30, L = Stirling)	2.0	58	1.22
R_7	3	Event	$\{\text{world, cup}\} \Rightarrow \{\text{final}\}$ <i>centroid</i> (T = 2014-07-13, L = Newcastle)	10.2	99	2.91
R_8	3	Celebrity	$\{\text{suarez}\} \Rightarrow \{\text{good,luck}\}$ <i>centroid</i> (T = 2014-07-11, L = London)	2.3	77	24.40
R_9	3	Event	$\{\text{november}\} \Rightarrow$ $\{\text{england,scotland}\}$ <i>centroid</i> (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 $\mathcal{D}_{(TW1,UK)}$, $\mathcal{D}_{(TW2,UK)}$, $\mathcal{D}_{(TW3,UK)}$)

1091 The capacity of the clustering algorithm integrating the TASTE measure
1092 to scale up to bigger data collections was assessed by measuring the execution
1093 time when varying (i) the number of tweets under analysis and (ii) the number
1094 of parallel executors. For scalability analysis, to get a larger number of tweets
1095 including all (text, temporal, and spatial) features, we have considered the
1096 location specified in the user profile as reference location information. Indeed
1097 the amount of tweets with geo-coordinates is much less than the number of
1098 tweets with location information in the user profile due to the limitation
1099 of GPS enabled devices. Geo-coordinates for the location extracted from
1100 the user profile have been calculated using Bing Maps Locations API. The

1101 resulting dataset, named \mathcal{D}'' , includes about 23.5 million tweets.

1102 To study scalability by varying the number of tweets, we considered dif-
1103 ferent sample rates of dataset \mathcal{D}'' and one executor for process running. In-
1104 creasing the number of tweets from 50,000 to about 2.35 million (10% of
1105 whole \mathcal{D}''), we notice an increment of the execution time (from 33m 34s to
1106 14h 31m). However, the growth rate of the execution time (about 25) is
1107 almost half the growth rate of the dataset size (about 47).

1108 To study scalability by varying the number of executors, we considered
1109 the whole dataset \mathcal{D}'' . The results show that, when increasing the number
1110 of executors from 4 to 8, the K-means algorithm integrating the TASTE
1111 measure scales almost linearly. The execution time is about 35h 43m with 4
1112 nodes; it decreases to about 19h 24m with 6 nodes, and to 10h 45m with 8
1113 nodes. Thus, with a suitable number of parallel executors, the clustering task
1114 is capable of handling also bigger data, evenly distributing the load across
1115 the nodes. When fewer than 4 executors are used, the process exceeded 48
1116 hours of execution and it was interrupted due to the very large dataset size.

1117 5. Comparison with previous studies

1118 This section discusses both a theoretical and analytical comparison be-
1119 tween our work and four previous studies on clustering Twitter data: (Kim
1120 et al., 2011),(Arcaini et al., 2016), (Lee, 2012), and (Cunha et al., 2014).
1121 These studies have proposed distance measures which combine the same
1122 tweet features considered in TASTE, or a subset of them. Specifically, the
1123 work in (Kim et al., 2011) takes into account the tweet spatial feature, while
1124 the spatio-temporal features are considered in (Arcaini et al., 2016), and both
1125 the text content and the spatial feature are evaluated in (Lee, 2012). A first
1126 attempt in considering all the three tweet features was proposed in (Cunha
1127 et al., 2014). Like in TCHARM, in these studies the geographic and tem-
1128 poral distances between tweets are computed using the Haversine and the
1129 Euclidean distance, respectively. The text content is represented with the
1130 BOW model, and the word relevance is weighted with the TF-IDF (Cunha
1131 et al., 2014) or the BursT (Lee, 2012) score; the cosine similarity is used to
1132 compare messages.

1133 For each study we present the objective of the work and the methodology
1134 for clustering tweets, including the clustering algorithm, the distance func-
1135 tions used and the strategy adopted for combining tweet features. Then, we
1136 discuss the analytical comparison between these works and our approach.

1137 In the following, we adopt the same notations as in Sections 3.1 and 3.3.
1138 An arbitrary tweet τ_i is a triplet $\tau_i = (t_i, s_i, W_i)$ where t_i and s_i are respec-
1139 tively the temporal and spatial features of τ_i , while $W_i \subseteq \Sigma$ is the tweet
1140 text content. Given two tweets $\tau_i = (t_i, s_i, W_i)$ and $\tau_j = (t_j, s_j, W_j)$ their
1141 temporal, spatial and content distances are denoted by $d_t(t_i, t_j)$, $d_s(s_i, s_j)$,
1142 and $d_W(W_i, W_j)$, respectively.

1143 The work in Kim et al. (2011) aims at providing (near-)real time infor-
1144 mation to users about events happening close to their location. Tweets are
1145 clustered through the K-means algorithm by considering their geographic dis-
1146 tance. The discovered cluster set is then analysed to detect clusters that can
1147 reveal the occurrence of an event. The values of the tweet temporal feature
1148 are used to filter computed clusters by comparing their temporal aspects. If
1149 the number of tweets from a given cluster exceed far from those from clusters
1150 found in vicinity in the past, the cluster is considered unusual and an event
1151 may happen there. For tweets included in unusual clusters, the text content
1152 is explored to extract representative keywords, which are sent to nearby users
1153 to inform them about the possible events.

1154 The study in Arcaini et al. (2016) focuses on discovering spatio-temporal
1155 periodic and aperiodic characteristics of events to support situation aware-
1156 ness. Tweets collections are analysed off-line with a DBSCAN based algo-
1157 rithm (GT-DBSCAN) to extract dense clusters of arbitrary shapes. The
1158 tweet text content is explored in a preprocessing phase to filter the subset
1159 of tweets relevant for the subsequent cluster analysis. Messages about spe-
1160 cific events are selected by properly setting keywords for tweets search. To
1161 drive the clustering process, three distance measures, considering the tweet
1162 temporal and spatial features, are evaluated: (i) a temporal distance, (ii) a
1163 geographic distance, and (iii) a geographic-temporal distance, basically a
1164 combination of the two above. In this study we focus on the latter distance
1165 measure for performance comparison. The geographic-temporal distance is
1166 defined as the maximum value between the (normalized) geographic and
1167 temporal distances.

1168 The work in Lee (2012) proposes a (near-)real time temporal-text cluster-
1169 ing approach to detect bursts of tweets representing unexpectedly frequent
1170 occurrences of a certain topic in a short period of time. A sliding window
1171 of fixed time length is used to filter only the most recent tweets, which are
1172 then considered in the analysis. Selected tweets are clustered using the Incre-
1173 mentalDBSCAN algorithm (Ester et al., 1998), to detect dense clusters with
1174 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)	$d_{Arc}(\tau_i, \tau_j) = [\text{Max}(d_s, d_t)]^\beta, \beta \in (0,1]$ d_s and d_t values expressed as the number of elementary units ϵ_s and ϵ_t , respectively
Lee (2012)	$d_{Lee}(\tau_i, \tau_j) = d_W \cdot e^{\zeta d_t/M}$ 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	$d_{TASTE}(\tau_i, \tau_j) = d_W \cdot (k_s \cdot e^{p_s \cdot d_s} + k_t \cdot e^{p_t \cdot d_t})$ $k_s, k_t, p_s, p_t \in \mathbb{R}; k_s, k_t \in [0, 1]$ 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} .

1175 Clusters are calculated by evaluating the temporal-text distance between
1176 tweets. In d_{Lee} , the temporal distance is used to modulate the text content
1177 distance. The exponential form has been adopted for the time distance to
1178 significantly penalize tweets far distant in time. Finally, geo-spatial keywords
1179 are extracted from message in each computed cluster to estimate location of
1180 detected events.

1181 The authors of Cunha et al. (2014) address the problem of identifying
1182 and displaying tweets profiles considering four different facets characterizing
1183 tweets: temporal, spatial, and context features and user social connections.
1184 Tweets are clustered with the DBSCAN algorithm Ester et al. (1996) to
1185 detect arbitrarily shaped clusters and to remove outliers from the results. The
1186 adopted distance measure is a linear combination of the four considered tweet
1187 features, i.e., the distance on time, space, text content, and social relations
1188 (d_{So}). The social distance term d_{So} evaluates the connections between users
1189 represented as nodes of a graph connected through edges. It is computed as
1190 the geodesic distance (i.e., the number of edges of the shortest path) between
1191 two nodes in the graph Bouttier et al. (2003).

1192 Based on the purposes of this paper, we want to evaluate the ability of
1193 each distance measure above in discovering cohesive clusters of tweets to be
1194 represented through their centroids. Hence, keeping the K-means algorithm

1195 used in TCHARM as a reference clustering method, we applied in turn each
1196 distance measure. Since the TCHARM methodology aims at discovering
1197 cohesive clusters considering temporal and spatial tweet features and text
1198 content, we omitted the social distance for the measure proposed in (Cunha
1199 et al., 2014). For the sake of brevity, the resulting clustering methods are de-
1200 noted by Cunha-14 (Cunha et al., 2014), Lee-12 (Lee, 2012), Arcaini-16 (Ar-
1201 caini et al., 2016), and Kim-11 (Kim et al., 2011). The approach proposed
1202 in this study adopting the TASTE measure is denoted by TCHARM.

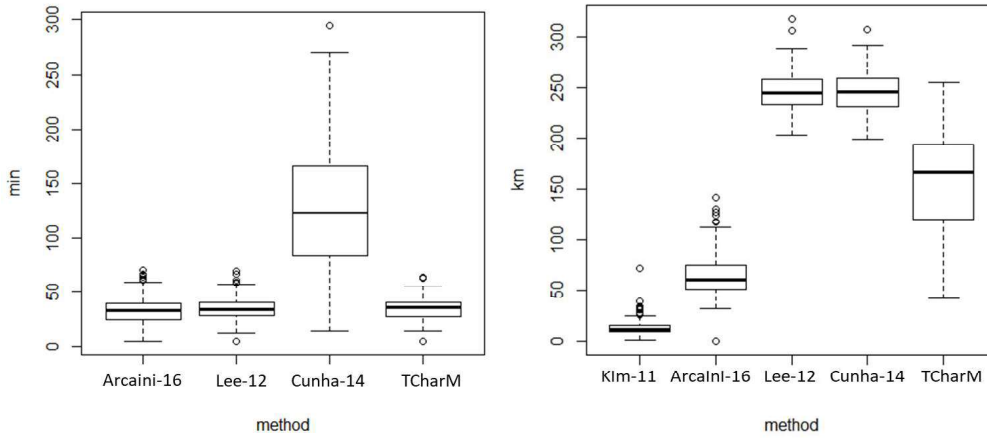
1203 We evaluated the cluster cohesion as the average geo-
1204 graphic/temporal/text content distance between tweets in the cluster
1205 and the cluster centroid. Lower values of these average distances point out
1206 a higher degree of cohesion on the corresponding tweet dimension.

1207 The comparison was performed with the $\mathcal{D}_{(TW1,UK)}$ dataset. To produce
1208 comparable cluster sets, we forced $K=200$ as expected number of clusters
1209 for all the distance measures (i.e., the same value selected for TCHARM in
1210 Section 4.2). We suitably tuned the parameters to use each distance measure
1211 at its best with the $\mathcal{D}_{(TW1,UK)}$ datasets and with the K-means algorithm.
1212 Starting from the configuration proposed in each study (considered as default
1213 configuration), we performed several runs to tune the parameters of each
1214 distance measure, with the aim of reducing the average cluster SSE as well
1215 as the distance values for all the tweet dimensions they consider. Selected
1216 parameter values are reported in Figure 8.

1217 For each method, box plots in Figure 8 illustrate the distributions of the
1218 average geographic/temporal/text content distance between tweets in each
1219 cluster and cluster centroid, while Table 10 reports the average values. Note
1220 that the temporal box plot for the Kim-11’s measure is not represented in
1221 Figure 8 as its values are too high compared to the other methods.

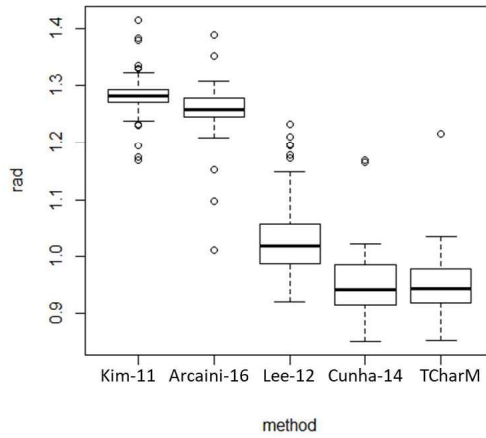
1222 Clusters manifest the highest text content cohesion with TCHARM,
1223 Cunha-14 and Lee-12 distance measures, which provide comparable results.
1224 The highest temporal cohesion is provided by Arcaini-16, TCHARM and Lee-
1225 12, which achieve similar performance. The highest spatial cohesion is given
1226 by Kim-11, followed by Arcaini-16, and then TCHARM.

1227 These results point out that TCHARM provides clusters with an over-
1228 all good cohesion on all the three facets characterizing tweets. Specifically,
1229 computed clusters show the highest cohesion on the text content and on the
1230 temporal feature, and the third best spatial cohesion. Yet it should be noted
1231 that, when setting parameters in TASTE, we gave more importance to the
1232 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.

1233 Clusters provided by Arcaini-16, Lee-12, and Kim-11 methods show a
1234 good cohesion on the tweet features considered in their proposed distance
1235 measures, but the cohesion on the remaining features is far lower than
1236 TCHARM. Clusters tend to be spread over a larger geographic area (Lee-12)
1237 or a longer time period (Kim-11) than TCHARM, or to discuss more different
1238 topics (Kim-11, Arcaini-16). These results demonstrate that, to obtain clus-
1239 ters suitable for a subsequent characterization of their spatial, temporal and
1240 text features, it is convenient to consider all the three dimensions directly
1241 in the clustering phase. Otherwise, further post-processing steps would be
1242 required to characterize the clusters with the features previously left out.

1243 Results also highlight that, when all three features are considered to clus-
1244 ter tweets, their contributions should be properly weighted in the distance
1245 measure. A liner combination of the content, spatial, and temporal distances
1246 as the one proposed in Cunha-14 turns out to be less suitable than our ap-
1247 proach since discovered clusters manifest a temporal and spatial cohesion
1248 lower than TCHARM.

1249 To deepen into the comparison of the methods above, we used the Ad-
1250 justed Rand Index (ARI) (Hubert & Arabie, 1985) to evaluate the agreement
1251 between the cluster sets generated using the TASTE measure and those ob-
1252 tained with the other distance measures. The ARI computes the rate of
1253 pairwise agreements between two partitions of a set. It allows a more accu-
1254 rate estimation of the agreement between two partitions than the standard

1255 Rand Index (Rand, 1971). Basically, ARI rescales the Rand Index value
1256 **with respect to its expected value for two independent clustering**
1257 **algorithms.** ARI has a maximum value of 1 for two identical partitions,
1258 and an expected value of 0 for two independent random partitions. Higher
1259 ARI values imply higher levels of agreement between two partitions.

1260 The computed values of ARI report a moderate agreement between the
1261 cluster set provided by TCHARM and the one computed by Cunha-14 (ARI
1262 = 0.45). The agreement decreases with Lee-11 (ARI = 0.13), Arcaini-16
1263 (ARI = 0.03), and Kim-11 (ARI = 0.005) methods which consider a subset
1264 of tweet features.

1265 The results from the analytical comparison suggest that clusters discov-
1266 ered using other distance measures have quite different properties than those
1267 provided by TCHARM.

1268 From a temporal perspective, clusters can have a higher temporal span.
1269 Indeed, while our clusters are centered around events of interest (see Sec-
1270 tion 4.3), we noticed that clusters computed with other methods (Kim-11 and
1271 Cunha-14) can include more than one event (e.g., more football matches).
1272 Similarly, the distance measures that do not provide a good cluster spatial
1273 cohesion lead to clusters of tweets spread across more counties (Lee, 2012).
1274 The two aspects above prevent from performing qualitative analyses based
1275 on fine-grained temporal and spatial resolutions. Finally, the lower text simi-
1276 larity among tweets in the clusters (Arcaini-16) makes it difficult to associate
1277 a single prevailing topic with each cluster and to generate significant asso-
1278 ciation rules (i.e., with high values of quality indices as support, confidence
1279 and lift).

1280 Thus, with the adoption of other distance measures than TASTE, a
1281 further level of segmentation would be required to identify the main topics
1282 in each cluster, or to partition the cluster content into subsets which refer to
1283 shorter time windows or more limited geographic areas.

1284 6. Discussion

1285 In this section we discuss the results discovered through TCHARM. The
1286 discussion addresses the data analysis phases in TCHARM, the computa-
1287 tional cost of TCHARM, and the possible exploitation of the TCHARM find-
1288 ings.

1289 (i) *Discovering in one step cohesive spatio-temporal clusters focused on spe-*
1290 *cific topics.* The TCHARM findings demonstrate the ability of the proposed
1291 methodology to properly analyse large tweet collections distributed over time
1292 and space as well as addressing various topics for automatically computing
1293 cohesive clusters. TCHARM allows data miners to discover clusters useful
1294 for identifying *when* and *where* people were more involved and about *which*
1295 topics. The 2014 FIFA World Cup use case considered in this study enables
1296 a thorough validation of computed clusters due to the availability of a time
1297 schedule for the main events (e.g., football matches) and web news about the
1298 other events or celebrities somehow involved. The experimental evaluation
1299 conducted on six different datasets showed that mined clusters are centered
1300 in time in correspondence with an event related to the 2014 FIFA World Cup
1301 and they mainly include messages about the event. Moreover, the clusters
1302 present a good spatio-temporal cohesion around their centroid.

1303 Differently from previous work (see Section 2), TCHARM clusters Twit-
1304 ter data taking into account in one step both spatio-temporal features and
1305 text content. TCHARM relies on the TASTE measure which combines the
1306 contributions of all three features above. TASTE modulates the distance
1307 between tweet messages through their distance in time and space, and it is
1308 aimed at discovering groups of tweets about the same topic but posted in
1309 nearby time periods and locations. Parameters of the TASTE measure can
1310 be conveniently tuned to fit scenarios with different spatial and temporal
1311 granularities.

1312 **The analytical comparison in Section 5 shows that TCHARM is**
1313 **competitive in terms of cluster cohesion, in almost all dimensions.**
1314 **In particular, it overperforms all the other measures in the text**
1315 **average distance. Indeed, the multiplicative (exponential) factors**
1316 **for time and space distances are suitably applied to the text dis-**
1317 **tance, based on the hypothesis that a tight temporal and spatial**
1318 **proximity can contribute in detecting clusters of tweets about the**
1319 **same topic. As already demonstrated in Section 4.3, such clusters**
1320 **are temporally centered within the time interval of the event they**
1321 **refer to (e.g., a football match).**

1322 **None of the measures considered for comparison performs far bet-**
1323 **ter than TASTE in more than one dimension. Moreover, the lower**
1324 **spatial cohesion obtained with TASTE is mainly due to our choice**
1325 **to assign a lower weight to spatial distance ($p_s = 3$), preferring the**
1326 **temporal cohesion ($p_t = 6$).**

1327 *(ii) Cluster characterization through rules analysis.* TCHARM deeply ex-
1328 plores the resulting clusters through association rule analysis to discover cor-
1329 relations among topics (such as events, celebrities, emotional states) and
1330 spatio-temporal features. While rule class TC makes possible the identifica-
1331 tion of the main topics discussed in each cluster, the other rule classes enable
1332 a deeper characterization by correlating topics with time periods (class T-
1333 TC), geographical areas (L-TC), or both of them (class L-T-TC). This cluster
1334 characterization allows data miners to better understand popular topics in
1335 different geographical areas and through different time windows. Moreover,
1336 association rules represent the mined knowledge in a concise and easily un-
1337 derstandable form.

1338 The 2014 FIFA World Cup use case allows us to qualitatively validate
1339 various mined rules. Rule analysis pointed out some of the interests and
1340 reactions of sports fans and supporters that were in some cases predictable
1341 (e.g., the disappointment of people from England over the English team’s
1342 defeats). However, it also highlighted some aspects not so evident a priori,
1343 like those about celebrities statements or the major interest in USA for the
1344 team of Argentina. We believe that TCHARM can be applied also in other
1345 scenarios, for understanding people’s reactions and interests.

1346 *(iii) TCHARM performance.* From a computational point of view, TCHARM
1347 has a major advantage with respect to related works, since it is implemented
1348 on Apache Spark and can distribute computational load across parallel ex-
1349 ecutors. Tests performed on big collections of tweets (Section 4.5) prove the
1350 good scalability of our implementation of TCHARM and, in particular, of
1351 the clustering algorithm integrating the TASTE measure. Thus, TCHARM
1352 can be applied also to use cases with a higher cardinality of data and it is
1353 still capable to provide results in a reasonable time.

1354 *(iv) Exploitation of the mined knowledge.* TCHARM findings provide a
1355 spatio-temporal overview of people involvement in occurred events. This
1356 knowledge, hidden in Twitter data collections, can have a variety of practi-
1357 cal applications in different domains.

1358 In case of events with a wide and spread out audience (as FIFA World
1359 Cup), TCHARM findings can provide useful insights to understand how peo-
1360 ple located in different geographical areas perceive an event and to char-
1361 acterize the different facets of people involvement in different time frames.
1362 From a business perspective, this knowledge can be very useful to improve

1363 service/product provision and support targeted advertising of certain ser-
1364 vices/products. For instance, the information about favourite teams or play-
1365 ers in specific areas and moments can be used to provide targeted adver-
1366 tising that leverages on such features. Also during 2014 FIFA World Cup,
1367 advertising companies demonstrated great interest in social trends to plan
1368 marketing strategies. This was particularly evident with some viral topics
1369 as some brands gained visibility by proposing advertisements based on viral
1370 marketing strategies, mostly on social networks (Jenkins, 2014; Bud, 2014).
1371 TCHARM can thus be an effective methodology to enable a deeper analy-
1372 sis of spatio-temporal trends on social networks, showing when and where
1373 certain topics spread among users.

1374 We believe that TCHARM can be profitably applied also in different
1375 domains. In a smart urban environment, for example, social networks are
1376 currently recognized as powerful instruments to enable citizen interaction
1377 and participation. Citizens may use Twitter to report information related
1378 to a variety of aspects such as urban safety, traffic and services (e.g., bike
1379 sharing, public transport offer, etc.). City administration is interested in
1380 better understanding where and when citizens report issues about the above
1381 aspects, to eventually undertake appropriate and targeted responses to cit-
1382 izens' concerns. The application of TCHARM to such collections of tweets
1383 would help to find out in which areas of the city and in which periods of time
1384 citizens discuss and complain about some issues. Clustering analysis would
1385 extract spatio-temporally defined clusters of topics reported by citizens. Rule
1386 analysis would then better highlight the degrees of correlation among topics,
1387 times and places of discussion and describe how the same topics evolve across
1388 different periods and through nearby urban areas.

1389 7. Conclusion

1390 In this paper we introduce TCHARM, a novel exploratory data mining
1391 methodology to analyse Twitter datasets. Its aim is to discover significant
1392 and cohesive groups of tweets by considering three facets of Twitter data:
1393 spatial, temporal, and text content information. The TASTE measure is one
1394 of the main added values of TCHARM as it allows the K-means algorithm
1395 to discover clusters with suitable levels of spatial and temporal cohesion,
1396 centered on specific events and including tweets which can be concisely rep-
1397 resented by their centroids with an acceptable approximation. Moreover,
1398 through association rules mining, TCHARM provides us with a set of pat-

1399 terns that concisely describe the most significant characteristics of tweets in
1400 clusters. The TCHARM system has been deployed on Apache Spark to dis-
1401 tribute computational load across parallel executors and reduce the overall
1402 execution time also with huge amounts of data.

1403 The experimental validation conducted on tweets collected for the 2014
1404 FIFA World Cup demonstrated the ability of TCHARM in efficiently charac-
1405 terizing collections of tweets in terms of distribution of people involvement,
1406 topic identification, and correlations among tweet features. As a matter of
1407 fact, we managed to isolate groups of tweets focused on a few topics, tem-
1408 porarily associated to actual events (e.g., football matches), and posted from
1409 a limited geographical area. Compared with other approaches for tweet clus-
1410 tering, clusters computed using the TASTE measure confirmed an overall
1411 better cohesion balanced between the three tweet features.

1412 TCHARM can be an effective methodology to enable a deeper analysis of
1413 spatio-temporal trends on social networks, showing the different patterns of
1414 user involvement in certain topics or events. TCHARM can be used to anal-
1415 yse global events like the FIFA World Cup at a local scale and, for instance,
1416 to assess the popularity of soccer matches and football players in different
1417 areas and time periods. This information could be very useful for compa-
1418 nies to improve their services and products and to optimize their marketing
1419 strategies. For example, information about favourite teams and players in
1420 specific areas and moments can be used to provide targeted advertising that
1421 leverages on the characteristics of the computed clusters.

1422 There is still room for improvement of the TCHARM methodology in
1423 order to mitigate some of its weaknesses. Five promising future research
1424 directions have been identified.

1425 In the current implementation of TCHARM, the number of expected clus-
1426 ters for the k-means algorithm and the parameters in the TASTE measure
1427 should be experimentally tuned by trading-off the cardinality of the cluster
1428 set and the expected quality of clusters. However, the selection of the proper
1429 TCHARM configuration can be a very time-consuming activity. The design
1430 of innovative *self-tuning configuration strategies* Di Corso et al. (2017) to
1431 automatically identify the suitable TCHARM set up for each targeted data
1432 collection can permit the use of TCHARM in various application domains.
1433 These strategies would simplify the analysts role by relieving the end-user of
1434 the burden of configuring the overall cluster analysis process.

1435 The ability of TCHARM to discover cohesive and significant clusters may
1436 decrease when data sparseness further increases. In this case, a larger number

1437 of clusters should be generated to discover groups with good quality, but these
1438 groups may be limited in size. To deal with this issue, *data taxonomies* on
1439 the three facets characterizing tweets can be climbed during the clustering
1440 process. The use of data taxonomies can result into coarse-grained data
1441 representations with a lower degree of sparsness and allows the evaluation of
1442 data correlations at different abstraction levels.

1443 The use of K-means clustering, rather than other clustering algorithms
1444 as density-based methods, was motivated in this study by the purpose of
1445 generating clusters of tweets that can be concisely represented by their cen-
1446 troids. However, TCHARM inherits one of the main weaknesses of K-means,
1447 which is more sensitive to outliers in the dataset. A future task is to conduct
1448 a detailed study on evaluating the *integration of other candidate clustering*
1449 *methods* in TCHARM and their ability to identify more cohesive and signif-
1450 icant clusters of tweets.

1451 Currently, the proposed TASTE measure weights various tweet facets,
1452 but omits other aspects such as the characteristics of users who posted tweets
1453 and their social relationships. Considering also *user information* in the clus-
1454 ter analysis would be very helpful to discover spatio-temporal patterns of
1455 communities of users and to better profile how the user interests evolve over
1456 time. As a future work, we will study an improvement of the TASTE mea-
1457 sure with the aim of evaluating also data about users.

1458 Finally, in this study we have applied the TCHARM engine for the off-
1459 line analysis of spatio-temporal-text information from tweets posted within a
1460 (relatively large) time window. As a future study, TCHARM can be applied
1461 for the (near-)real time analysis, for instance of tweets collected every hour,
1462 to investigate the spatial evolution of clusters and related topics with a low
1463 time granularity. This approach would provide a deeper overview of the
1464 spatio-temporal dynamics of people’s interests. Thanks to the deployment
1465 on a cloud-based platform as Apache Spark, TCHARM can analyse huge
1466 amounts of data thus providing results in a reasonable time consistent with
1467 a near-real time analysis.

1468 **References**

- 1469 Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules
1470 between sets of items in large databases. In *SIGMOD Conference* (pp.
1471 207–216).

- 1472 Arcaini, P., Bordogna, G., Ienco, D., & Sterlacchini, S. (2016). User-driven
1473 geo-temporal density-based exploration of periodic and not periodic events
1474 reported in social networks. *Inf. Sci.*, *340*, 122–143.
- 1475 Baralis, E., Cerquitelli, T., Chiusano, S., Grimaudo, L., & Xiao, X. (2013).
1476 Analysis of twitter data using a multiple-level clustering strategy. In *Model
1477 and Data Engineering* (pp. 13–24). Springer.
- 1478 Bernabe-Moreno, J., Tejada-Lorente, A., Porcel, C., & Herrera-Viedma, E.
1479 (2015). A new model to quantify the impact of a topic in a location over
1480 time with social media. *Expert Systems with Applications*, *42*, 3381 – 3395.
- 1481 Bouttier, J., Francesco, P. D., & Guitter, E. (2003). Geodesic distance in
1482 planar graphs. *Nuclear Physics B*, *663*, 535 – 567.
- 1483 Bud (2014). Bud light on twitter. <http://t.co/Kj69E17MRE>.
- 1484 Cagliero, L., Cerquitelli, T., Garza, P., & Grimaudo, L. (2014). Twitter
1485 data analysis by means of strong flipping generalized itemsets. *Journal of
1486 Systems and Software*, *94*, 16–29.
- 1487 Capdevila, J., Pericacho, G., Torres, J., & Cerquides, J. (2016). Scaling
1488 dbscan-like algorithms for event detection systems in twitter. *Lecture Notes
1489 in Computer Science (including subseries Lecture Notes in Artificial Intel-
1490 ligence and Lecture Notes in Bioinformatics)*, *10048 LNCS*, 356–373.
- 1491 Cunha, T., Soares, C., & Mendes Rodrigues, E. (2014). Tweepfiles: Detec-
1492 tion of spatio-temporal patterns on twitter. In X. Luo, J. X. Yu, & Z. Li
1493 (Eds.), *Advanced Data Mining and Applications: 10th International Con-
1494 ference, ADMA 2014, Guilin, China, December 19-21, 2014. Proceedings*
1495 (pp. 123–136). Cham: Springer International Publishing.
- 1496 Dasgupta, S. S., Natarajan, S., Kaipa, K. K., Bhattacharjee, S. K., &
1497 Viswanathan, A. (2015). Sentiment analysis of facebook data using hadoop
1498 based open source technologies. In *2015 IEEE International Conference
1499 on Data Science and Advanced Analytics (DSAA)* (pp. 1–3).
- 1500 Di Corso, E., Cerquitelli, T., & Ventura, F. (2017). Self-tuning techniques
1501 for large scale cluster analysis on textual data collections. In *Proceedings
1502 of the Symposium on Applied Computing SAC '17* (pp. 771–776). New
1503 York, NY, USA: ACM.

- 1504 Ester, M., Kriegel, H.-P., Sander, J., Wimmer, M., & Xu, X. (1998). Incre-
1505 mental clustering for mining in a data warehousing environment. In *Pro-*
1506 *ceedings of the 24rd International Conference on Very Large Data Bases*
1507 *VLDB '98* (pp. 323–333). San Francisco, CA, USA: Morgan Kaufmann
1508 Publishers Inc.
- 1509 Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based al-
1510 gorithm for discovering clusters a density-based algorithm for discovering
1511 clusters in large spatial databases with noise. In *Proceedings of the Sec-*
1512 *ond International Conference on Knowledge Discovery and Data Mining*
1513 *KDD'96* (pp. 226–231). AAAI Press.
- 1514 García-Gavilanes, R., Kaltenbrunner, A., Sáez-Trumper, D., Baeza-Yates,
1515 R., Aragón, P., & Laniado, D. (2014). Who are my audiences? a study
1516 of the evolution of target audiences in microblogs. In L. M. Aiello, &
1517 D. McFarland (Eds.), *Social Informatics: 6th International Conference,*
1518 *SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings* (pp.
1519 561–572). Springer International Publishing.
- 1520 Godfrey, D., Johns, C., Meyer, C., Race, S., & Sadek, C. (2014). A case
1521 study in text mining: Interpreting twitter data from world cup tweets.
1522 *arXiv preprint arXiv:1408.5427*, .
- 1523 Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candi-
1524 date generation. In *SIGMOD'00, Dallas, TX*, .
- 1525 Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classifi-*
1526 *cation*, 2, 193–218.
- 1527 Ikeda, K., Hattori, G., Ono, C., Asoh, H., & Higashino, T. (2013). Twitter
1528 user profiling based on text and community mining for market analysis.
1529 *Knowledge-Based Systems*, 51, 35 – 47.
- 1530 Jenkins, T. (2014). World cup 2014: fans get bitten by the luis suarez bug
1531 - in pictures. [https://www.theguardian.com/football/gallery/2014/
1532 jun/26/world-cup-2014-luis-suarez-advert-bite](https://www.theguardian.com/football/gallery/2014/jun/26/world-cup-2014-luis-suarez-advert-bite).
- 1533 Jing, L., Ng, M. K., & Huang, J. Z. (2007). An entropy weighting k-means
1534 algorithm for subspace clustering of high-dimensional sparse data. *IEEE*
1535 *Transactions on Knowledge and Data Engineering*, 19, 1026–1041.

- 1536 Kim, J. W., Kim, D., Keegan, B., Kim, J. H., Kim, S., & Oh, A. (2015).
1537 Social media dynamics of global co-presence during the 2014 fifa world cup.
1538 In *Proceedings of the 33rd Annual ACM Conference on Human Factors in*
1539 *Computing Systems* (pp. 2623–2632). ACM.
- 1540 Kim, T., Huerta-Canepa, G., Park, J., Hyun, S. J., & Lee, D. (2011). What’s
1541 happening: Finding spontaneous user clusters nearby using twitter. In
1542 *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Iner-*
1543 *national Conference on Social Computing (SocialCom), 2011 IEEE Third*
1544 *International Conference on* (pp. 806–809). IEEE.
- 1545 Lee, C. H. (2012). Mining spatio-temporal information on microblogging
1546 streams using a density-based online clustering method. *Expert Systems*
1547 *with Applications*, 39, 9623–9641.
- 1548 Lee, R., Wakamiya, S., & Sumiya, K. (2015). Exploring geospatial cognition
1549 based on location-based social network sites. *World Wide Web*, 18, 845–
1550 870.
- 1551 Lloret, E., Balahur, A., Gómez, J. M., Montoyo, A., & Palomar, M. (2012).
1552 Towards a unified framework for opinion retrieval, mining and summariza-
1553 tion. *Journal of Intelligent Information Systems*, 39, 711–747.
- 1554 Manning, C. D., Raghavan, P., Schütze, H. et al. (2008). *Introduction to*
1555 *information retrieval* volume 1. Cambridge university press Cambridge.
- 1556 Pang-Ning T. and Steinbach M. and Kumar V. (2006). *Introduction to Data*
1557 *Mining*. Addison-Wesley.
- 1558 Phelan, O., Mccarthy, K., & Smyth, B. (2009). Using twitter to recommend
1559 real-time topical news. In *Proceedings of the third ACM conference on*
1560 *Recommender systems* (pp. 385–388). ACM.
- 1561 Rabiger, S., & Spiliopoulou, M. (2015). A framework for validating the merit
1562 of properties that predict the influence of a twitter user. *Expert Systems*
1563 *with Applications*, 42, 2824 – 2834.
- 1564 Rand, W. M. (1971). Objective criteria for the evaluation of clustering meth-
1565 ods. *Journal of the American Statistical Association*, 66, 846–850.

- 1566 Robertson, S. (2004). Understanding inverse document frequency: on theo-
1567 retical arguments for idf. *Journal of Documentation*, *60*, 503–520.
- 1568 Saito, K., Ohara, K., Kimura, M., & Motoda, H. (2015). Change point de-
1569 tection for burst analysis from an observed information diffusion sequence
1570 of tweets. *Journal of Intelligent Information Systems*, *44*, 243–269.
- 1571 Sakai, T., Tamura, K., Kotozaki, S., Hayashida, T., & Kitakami, H. (2015).
1572 Real-time local topic extraction using density-based adaptive spatiotempo-
1573 ral clustering for enhancing local situation awareness. In *2015 7th Interna-
1574 tional Joint Conference on Knowledge Discovery, Knowledge Engineering
1575 and Knowledge Management (IC3K)* (pp. 203–210). volume 01.
- 1576 Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for
1577 automatic indexing. *Commun. ACM*, *18*, 613–620.
- 1578 Steiger, E., Resch, B., & Zipf, A. (2016). Exploration of spatiotemporal and
1579 semantic clusters of twitter data using unsupervised neural networks. *Int.
1580 J. Geogr. Inf. Sci.*, *30*, 1694–1716.
- 1581 Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document
1582 clustering techniques. In *KDD Workshop on Text Mining*.
- 1583 Thomas, K., Grier, C., Song, D., & Paxson, V. (2011). Suspended accounts
1584 in retrospect: An analysis of twitter spam. In *Proceedings of the 2011 ACM
1585 SIGCOMM conference on Internet measurement conference* (pp. 243–258).
1586 ACM.
- 1587 Vicient, C., & Moreno, A. (2015). Unsupervised topic discovery in micro-
1588 blogging networks. *Expert Systems with Applications*, *42*, 6472 – 6485.
- 1589 Yang, M.-C., & Rim, H.-C. (2014). Identifying interesting twitter contents
1590 using topical analysis. *Expert Systems with Applications*, *41*, 4330 – 4336.
- 1591 Yu, Y., & Wang, X. (2015). World cup 2014 in the twitter world: A big data
1592 analysis of sentiments in u.s. sports fans’ tweets. *Computers in Human
1593 Behavior*, *48*, 392–400.
- 1594 Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010).
1595 Spark: Cluster computing with working sets. In *Proceedings of the 2Nd
1596 USENIX Conference on Hot Topics in Cloud Computing HotCloud’10* (pp.
1597 10–10). Berkeley, CA, USA: USENIX Association.