Analysis of Twitter Data Using a Multiple-level Clustering Strategy

Analysis of Twitter Data Using a Multiple-Level Clustering Strategy

Elena Baralis, Tania Cerquitelli, Silvia Chiusano, Luigi Grimaudo, Xin Xiao

Dipartimento di Automatica e Informatica Politecnico di Torino - Torino, Italy
Email:{elena.baralis,tania.cerquitelli,silvia.chiusano,luigi.grimaudo,xin.xiao}@polito.it

Keywords Clustering algorithms, association rules, social networks, tweets

Abstract *Twitter, currently the leading microblogging social network, has attracted a great body of research works. This paper proposes a data analysis framework to discover groups of similar twitter messages posted on a given event. By analyzing these groups, user emotions or thoughts that seem to be associated with specific events can be extracted, as well as aspects characterizing events according to user perception. To deal with the inherent sparseness of micro-messages, the proposed approach relies on a multiple-level strategy that allows clustering text data with a variable distribution. Clusters are then characterized through the most representative words appearing in their messages, and association rules are used to highlight correlations among these words. To measure the relevance of specific words for a given event, text data has been represented in the Vector Space Model using the TF-IDF weighting score. As a case study, two real Twitter datasets have been analysed.*

Recently, social networks and online communities, such as Twitter and Facebook, have become a powerful source of knowledge being daily accessed by millions of people. A particular attention has been paid to the analysis of the User Generated Content (UGC) coming from Twitter, which is one of the most popular microblogging websites. Different approaches addressed the discovery of the most relevant user behaviors Bender et al (2008) or topic trends Cagliero and Fiori (2013); Cheong and Lee (2009); Lopes et al (2007).

Twitter textual data (i.e., tweets) can be analysed to discover user thoughts associated with specific events, as well as aspects characterizing events according to user perception. Clustering techniques can provide a coherent summary of tweets, which can be used to provide summary insight into the overall content of the underlying corpus. Nevetheless clustering is a widely studied data mining problem in the text domain, clustering twitter messages imposes new challenges due to their inherent sparsness.

This paper proposes a data analysis framework to discover, in a data collection with a variable distribution, cohesive and well-separated groups of tweets. Our framework exploits a multiple-level clustering strategy that iteratively focuses on disjoint dataset portions and locally identifies clusters. The density-based DBSCAN algorithm Ester et al (1996) has been adopted because it allows the identification of arbitrarily shaped clusters, is less susceptible to noise and outliers, and does not require the specification of the number of expected clusters in the data. To highlight the relevance of specific words for a given tweet or set of tweets, tweets have been represented in the Vector Space Model (VSM) Steinbach et al (2000) using the TF-IDF weighting score Steinbach et al (2000). The cluster content has been compactly represented with the most representative words appearing in their tweets based on the TF-IDF weight. Association rules representing word correlations are also discovered to point out in a compact form the information characterizing each cluster. To our knowledge, this work is the first study addressing a jointly exploitation of a multiple-level clustering strategy with association rules for tweet analysis.

As a reference case study, the proposed framework has been applied to two real datasets retrieved from Twitter. The results showed that, starting from a tweet collection, the framework allows the identification of clusters containing similar messages posted on an event. The multiple-level strategy iterated for three levels compute clusters that progressively contain longer tweets describing the event through a more varied vocabulary, talking about some specific aspects of the event, or reporting user emotions associated with the event.

The paper is organized as follows. Section 1 presents a motivating example. Section 2 describes the proposed framework and describes its building blocks, while the results obtained for the two real datasets are discussed in Section 3. Finally, Section 4 analyses previous related work and Section 5 draws conclusions and future work.

1 Motivating Example

Tweets are short, user-generated, textual messages of at most 140 characters long and publicly visible by default. For each tweet a list of additional features (e.g., GPS coordinates, timestamp) on the context in which tweets have been posted is also available.

This paper focuses on the analysis of the textual part of Twitter data (i.e., on tweets) to provide summary insight into some specific aspects of an event or discover user thoughts associated with specific events. Clustering techniques are used to identify groups of similar tweets. Cluster analysis partitions objects into groups (clusters) so that objects within the same group are more similar to each other than those objects assigned to different groups Pang-Ning T. and Steinbach M. and Kumar V. (2006). Each cluster is then compactly described through the most representative words occurring in their tweets and the association rules modeling correlations among these words. Association rules Han et al (2000) identify collections of itemsets (i.e., sets of words in the tweet analysis) that are statistically related in the underlying dataset. Association rules are usually represented in the form $X \rightarrow Y$, where X and Y are disjoint itemsets (i.e., disjoint conjunctions of words).

A simplified example of the textual part of two Twitter messages is shown in Figure 1. Both tweets regard the Paralympic Games that took place in London in year 2012. As described in Section 2.1, to suit the textual data to the subsequent data mining steps, tweets are preprocessed in the framework by removing links, stopwords, no-ascii chars, mentions, and replies.

Our proposed framework assigns the two example tweets to two different clusters, due to their quite unlike textual data. Both example tweets contain words as $\{paralympics, olympic, stadium\}$, overall describing the paralympics event. In addition, $\{fireworks, closingceremony\}$ and $\{amazing, athletics\}$ are the representative word sets for Tweets 1 and 2, respectively, reporting the specific subject of each message. The association rules $\{closingceremony \rightarrow fireworks\}$ and $\{amazing \rightarrow athletics\}$ model correlations among representative words in the two tweets. They allow us to point out in a compact form the representative information characterizing the two messages. While the first tweet talks about a specific event in the closing ceremony (i.e., the fireworks), the second one reports a positive opinion of people attending the event.

2 The Proposed Multiple-Level Clustering Framework

The proposed framework to analyse Twitter data is shown in Figure 2 and detailed in the following subsections.

```
TWEET 1 - text: {Fireworks on! paralympics closingceremony at Olympic Stadium}
TWEET 2 - text: {go to Olympic Stadium for amazing athletics at Paralympics}
```

Figure 1: Two simplified example tweets

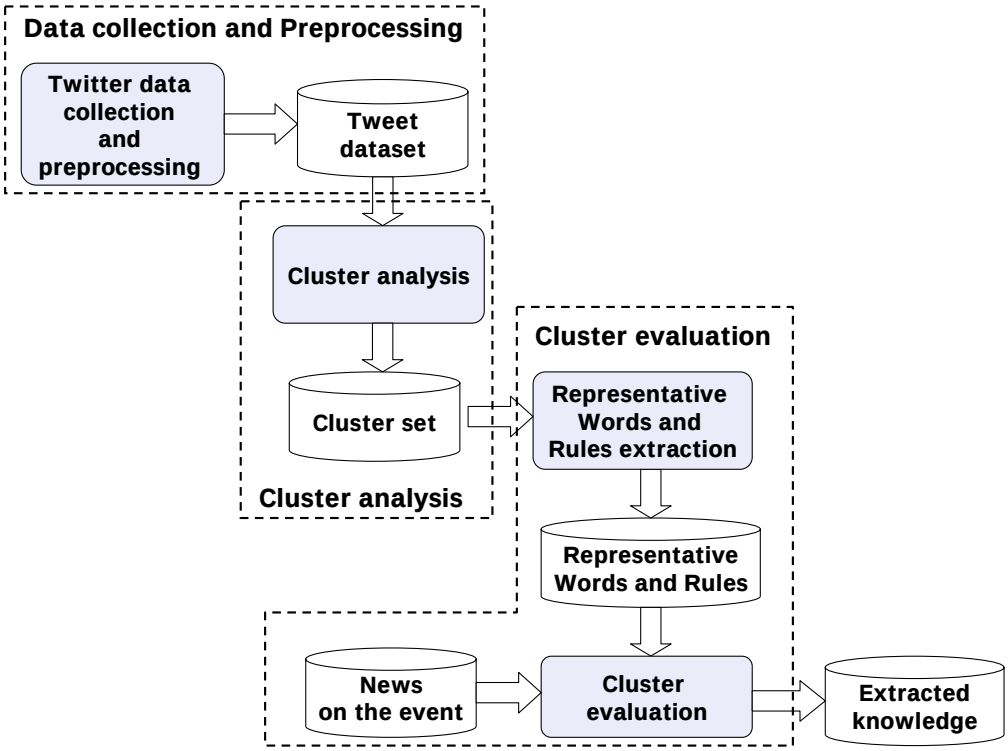

Figure 2: The proposed multiple-level clustering framework for tweet analysis

The textual content of Twitter posts (i.e., the tweets) is retrieved through the Twitter Stream APIs (Application Programming Interfaces) and preprocessed to make it suitable for the subsequent mining steps. The multiple-level clustering approach is then applied to discover, in a dataset with a variable distribution, groups of tweets with a similar informative content. The DBSCAN algorithm has been exploited for the cluster analysis.

Clustering results are evaluated through the Silhouette Rousseeuw (1987) quality index, balancing both intra-cluster homogeneity and inter-cluster separation. To analyse tweets contained in the cluster set, each cluster has been characterized with the most representative words appearing in its tweets and the association rules modeling correlations among these words. We validated both the meaning and the importance of the information extracted from the tweet datasets with the support of news available on the web. This allows us to properly frame the context in which tweets were posted.

2.1 Twitter Data Collection and Preprocessing

Tweet content and their relative contextual data are retrieved through the Stream Application Programming Interfaces (APIs). Data is gathered by establishing and maintaining a continuous connection with the stream endpoint.

To suit the raw tweet textual to the following mining process, some preliminary data cleaning and processing steps have been applied. The textual message content is first preprocessed by eliminating stopwords, numbers, links, non-ascii characters, mentions, and replies. Then, it is represented by means of the Bag-of-Word (BOW) representation Steinbach et al (2000).

Tweets are transformed using the Vector Space Model (VSM) Steinbach et al (2000). Each tweet is a vector in the word space. Each vector element corresponds to a different word and is associated with a weight describing the word relevance for the tweet. The Term Frequency (TF) - Inverse Document Frequency (IDF) scheme Steinbach et al (2000) has been adopted to weight word frequency. This data representation allows highlighting the relevance of specific words for each tweet. It reduces the importance of common terms in the collection, ensuring that the matching of tweets is more influenced by discriminative words with relatively low frequency in the collection. In short-messages as tweets, the TF-IDF weighting score could actually boild down to a pure IDF due to the limited word frequency within each tweet. Nevertheless, we preserved the TF-IDF approach to consider also possible word repetitions.

The tweet collection is then partitioned based on trending topics, identified by analysing the most frequent hashtags. A dataset partition is analyzed as described in the following sections.

2.2 Cluster Analysis

Differently from other clustering methods, density-based algorithms can effectively discover clusters of arbitary shape and filter out outliers, thus increasing cluster homogeneity. Additionally, the number of expected clusters in the data is not required. Tweet datasets can include outliers as messages posted on some specific topics and clusters can be non-spherical shaped. Besides, the expected number of clusters can be hardly guessed a priori, because our aim is discovering groups of similar tweets through an explorative data analysis. For these reasons, the DBSCAN density-based method has been selected for tweet cluster analysis.

In the DBSCAN algorithm Ester et al (1996), clusters are identified as dense areas of data objects surrounded by an area of low density. Density is evaluated based on the user-specified parameters Eps and $MinPts$. A dense region in the data space is a n-dimensional sphere with radius Eps and containing at least $MinPts$ objects. DBSCAN iterates over the data objects in the collection by analyzing their neighborhood. It classifies objects as being (i) in the interior of a dense region (a core point), (ii) on the edge of a dense region (a border point), or (iii) in a sparsly occupied region (a noise or outlier point). Any two core points that are close enough (within a distance Eps of one another) are put in the same cluster. Any border point close enough to a core point is put in the same cluster as the core point. Outlier points (i.e., points far from any core point) are isolated.

One single execution of DBSCAN discovers dense groups of tweets according to one specific setting of the Eps and $MinPts$ parameters. Tweets in lower density areas are labeled as outliers and not assigned to any cluster. Hence, different parameter settings are needed to discover clusters in datasets with a variable data distribution as the one considered in this study.

In application domains where data collections have a variable distribution, clustering algorithms can be applied in a multiple-level fashion Antonelli et al (2013). In this study we coupled a *multiple-level clustering* approach with association rule mining to discover representative clusters and the information characterizing them. Our approach iteratively applies the DBSCAN algorithm on different (disjoint) dataset portions. The whole original dataset is clustered at the first level. Then, at each subsequent level, tweets labeled as outliers in the previous level are re-clustered. The DBSCAN parameters Eps and $MinPts$ are properly set at each level by addressing the following issues. To discover representative clusters for the dataset, we aim at avoiding clusters including few tweets. In addition, to consider all different posted information, we aim at limiting the number of tweets labeled as outliers and thus unclustered.

The cosine similarity measure has been adopted to evaluate the similarity between tweets represented in the VSM model using the TF-IDF method. This measure has been often used to compare documents in text mining Steinbach et al (2000).

2.3 Cluster Evaluation

The discovered cluster set is evaluated using the Silhouette index Kaufman, L. and Rousseeuw, P. J. (1990). Silhouette allows evaluating the appropriateness of the assignment of a data object to a cluster rather than to another by measuring both intra-cluster cohesion and inter-cluster separation.

The silhouette value for a cluster C is the average silhouette value on all its tweets. Negative silhouette values represent wrong tweet placements, while positive silhouette values a better tweet assignments. Clusters with silhouette values in the range [0.51,0.70] and [0.71,1] respectively show that a reasonable and a strong structure have been found Kaufman, L. and Rousseeuw, P. J. (1990). The cosine similarity metric has been used for silhouette evaluation, since this measure was used to evaluate tweet similarity in the cluster analysis (see Section 2.2).

Each cluster has been characterized in terms of the words appearing in its tweets and the association rules modeling strong correlations among these words. News available on the web are used to properly frame the context in which tweets were posted and validate the extracted information. Specifically, the most representative words for each cluster are highlighted. These words are the relevant words for the cluster based on the TF-IDF weight. They occur with higher frequency in tweets in the cluster than in tweets contained in other clusters.

The quality of an association rule $X \rightarrow Y$, with X and Y disjoint itemsets (i.e., sets of words in this study), is usually measured by rule support and confidence. Rule support is the percentage of tweets containing both X and Y. Rule confidence is the percentage of tweets with X that also contain Y, and describes the strength of the implication. To rank the most interesting rules, we also used the lift index Pang-Ning T. and Steinbach M. and Kumar V. (2006), which measures the (symmetric) correlation between sets X and Y. Lift values below 1 show a negative correlation between sets X and Y, while values above 1 indicate a positive correlation. The interest of rules having a lift value close to 1 may be marginal. In this work, to mine association rules representing strong word correlations, rules with high confidence value and lift grater than one have been selected.

3 First Experimental Validation

This section presents and discusses the preliminary results obtained when analysing two real collections of twitter messages with the proposed framework.

3.1 Datasets

We evaluated the usefulness and applicability of the proposed approach on two real datasets retrieved from Twitter (http://twitter.com). Our framework exploits a crawler to access the Twitter global stream efficiently. To generate the real Twitter datasets we monitored the public stream endpoint offered by the Twitter APIs over a 1-month time period and tracked a selection of keywords ranging over two different topics, i.e., Sport and Music. The crawler establishes and maintains a continuous connection with the stream endpoint to collect and store Twitter data.

For both Twitter data collections, we analyzed the most frequent hashtags to discover trending topics. Among them, we selected the following two reference datasets for our experimental evaluation: the *paralympics* and the *concert* datasets. The *paralympics* dataset contains tweets on the Paralympic Games that took place in London in year 2012. The *concert* dataset contains tweets on the Madonna's concert held in September 6, 2012, at the Yankee Stadium located at The Bronx in New York City. Madonna is an American singer-songwriter and this concert was part of the "Mdna 2012 World Tour". Tweets in each dataset are preprocessed as described in Section 2.1. Hashtags used for tweets selection have been removed from the corresponding dataset, because appearing in all its tweets.

The main characteristics of the two datasets are as follows. The paralympics dataset contains 1,696 tweets with average length 6.89. The concert dataset contains 2,960 tweets with average length 6.38. Datasets used in the experiments are available at DBDMG (2013).

3.2 Framework Configuration

In the proposed framework, the procedures for data transformation and cluster evaluation have been developed in the Java programming language. These procedures transform the tweet collection into the VSM representation using the TF-IDF scheme and compute the silhouette values for the cluster set provided by the cluster analysis. The DBSCAN Ester et al (1996) and FPGrowth Han et al (2000) algorithms available in the RapidMiner toolkit Rapid Miner Project (2013) have been used for the cluster analysis and association rule extraction, respectively.

To select the number of iterations for the multiple-level clustering strategy and the DBSCAN parameters for each level, we addressed the following issues. We aim at avoiding clusters including few tweets, to discover representative clusters, and at limiting the number of unclustered tweets, to consider all posted information. For both datasets we adopted a three-level clustering approach, with each level focusing on a different dataset part. The Eps and $MinPts$ values at each iteration level for the two datasets are reported in Section 3.3.

To extract association rules representing strong correlations among words appearing in tweets contained in each cluster, we considered a minimum confidence threshould greater than or equal to 80%, lift greater than 1, and a minimum support threshold greater than or equal to 10%.

3.3 Analysis of the Clustering Results

Starting from a collection of Twitter data related to an event, the proposed framework allows the discovery of a set of clusters containing similar tweets. The multiple-level DBSCAN approach, iterated for three levels, computed clusters progressively containing longer tweets, that (i) describe the event through a more varied vocabulary, (ii) focus on some specific aspects of the event, or (ii) report user emotions and thoughts associated with the event.

First-level clusters contain tweets mainly describing general aspects of the event. Second-level clusters collect more diversified tweets that describe some specific aspects of the event or express user opinions about the event. Tweets become progressively longer and more focused in third-level clusters, indicating that some additionally specific aspects have been addressed. Since at each level clusters contain more specific messages, a lower number of tweets are contained in each cluster and the cluster size tends to reduce progressively. By further applying the DBSCAN algorithm on the subsequent levels, fragmented groups of tweets can be identified. Clusters show good cohesion and separation as they are characterized by high silhouette values. Both the meaning and the importance of the information extracted from the two datasets has been validated with the support of news on the event available on the web.

Cluster properties are discussed in detail in the following subsections. Tables 1 and 2 report, for each first- and second-level cluster in the two datasets, the number of tweets, the average tweet length, the silhouette value, and the most representative words. Representative association rules are also reported, pointing out in a compact form the discriminative information characterizing each cluster. Clusters are named as C_{i_j} in the

tables, where j denotes the level of the multiple-level DBSCAN approach providing the cluster and i locally identifies the cluster at each level j.

3.3.1 Tweet Analysis in the Paralympics Dataset.

First-level clusters can be partitioned into the following groups: clusters containing tweets that (i) post general information about the event (clusters C_{1_1} and C_{2_1}), (ii) regard a specific discipline (C_{3_1}) or team (C_{4_1} and C_{5_1}) among those involved in the event, (iii) report user emotions (C_{6_1}), and (iv) talk about the closing ceremony (C_{7_1}).

Specifically, clusters C_{1_1} and C_{2_1} mainly contains information about the event location (rule $\{london\} \rightarrow \{stadium, olympics\}$). Clusters C_{4_1} is about the Great Britain team taking part in the Paralympics event (rule $\{teamgb\} \rightarrow \{olympic\}$). Clusters C_{3_1} and C_{6_1} focus on the athletics discipline. While cluster C_{3_1} simply associates athletics with the Olympic event, users in cluster C_{6_1} express their appreciation on the athletics competitions they are attending (rule $\{athletics\} \rightarrow \{amazing, day\}$). Finally, tweets in cluster C_{7_1} talk about the seats of people attending the final ceremony (rule $\{closingceremony, stadium\} \rightarrow \{seats\}$).

Second-level clusters contain more diversified tweets. The following categories of clusters can be identified: clusters with tweets posting information on (i) specific events in the closing ceremony (clusters C_{1_2} and C_{2_2}), (ii) specific teams (cluster C_{3_2}) or competitions (cluster C_{4_2}) in Paralympics, and (iii) thoughts of people attending Paralympics (cluster C_{5_2}).

More in detail, cluster C_{1_2} focuses on the flame that was put out on the day of the closing celebration (rule $\{stadium, london\} \rightarrow \{flame, closingceremony\}$), while cluster C_{2_2} is on the fireworks that lit up London's Olympic stadium in the closing ceremony (rule $\{stadium, closingceremony\} \rightarrow \{fireworks\}$). Cluster C_{3_2} is about the Great Britain team taking part to athletics discipline (rule $\{teamgb, park\} \rightarrow \{athletics\}$). Tweets in cluster C_{4_2} address the final basketball competition in the North Greenwich Arena. They contain the information about the event location and the German women's team involved in the competition (rules $\{final\} \rightarrow \{north, germany\}$ and $\{final\} \rightarrow \{basketball, germany\}$). Tweets in cluster C_{5_2} show an enthusiastic feeling on Paralympics (rule $\{stadium, olympic\} \rightarrow \{london, fantasticfriday\}$) and the desire to share pictures on them (rule $\{pic, dreams\} \rightarrow \{stadium, time\}$).

Third-level clusters (with DBSCAN parameters $MinPts = 15$, $Eps = 0.65$) show a similar trend to second-level clusters. For example, clusters contain tweets on some specific aspects of the closing ceremony, as the participation of the ColdPlay band (rule $\{london\} \rightarrow \{coldplay, watching\}$), or tweets about a positive feeling on the Paralympics event (rules $\{love\} \rightarrow \{summer, olympics\}$ and $\{gorgeous\} \rightarrow \{day\}$). By stopping the multiple-level DBSCAN approach at this level, 808 tweets labeled as outliers remain unclustered, with respect to the initial collection of 1,696 tweets.

3.3.2 Tweet Analysis in the Concert Dataset.

Among first-level clusters, we can identify groups of tweets mainly posting information on the concert location (clusters C_{1_1}, C_{2_1}, and C_{3_1} with rule $\{concert, mdna\} \rightarrow \{yankee\}$). The remaining clusters talk about some aspects of the concert. For example, cluster C_{4_1} regards the opening act (rule $\{yankee, stadium\} \rightarrow \{opening, act\}$). Cluster C_{5_1} is on the participation of the Avicii singer (rule $\{wait\} \rightarrow \{yankee, avicii\}$), cluster C_{6_1} on the "forgive" writing on Madonna's back (rule $\{forgive\} \rightarrow \{stadium, nyc\}$), and cluster C_{7_1} is about the raining weather (rule $\{rain\} \rightarrow \{yankee, stadium\}$). Finally, cluster C_{8_1} regards people sharing concert pictures (rule $\{queen\} \rightarrow \{instagram\}$).

In second-level clusters, tweets focus on more specific aspects related to the concert. For example tweets in cluster C_{2_2} refer to Madonna with the "madge" nickname typically used by her fans (rule $\{singing\} \rightarrow \{stadium, madge\}$).

Similar to the paralympics dataset, also in the concert dataset third-level clusters (with DBSCAN parameter Eps=0.77 and $MinPts$=23) show a similar trend to second-level clusters. For example, clusters contain tweets regarding some particular songs. At this stage, 1660 tweets labeled as outliers remain unclustered, with respect to the initial collection of 2,960 tweets considered at the first level.

3.4 Performance Evaluation

Experiments were performed on a 2.66 GHz Intel(R) Core(TM)2 Quad PC with 8 GB main memory running linux (kernel 3.2.0). The run time of DBScan at the first, second, and third level is respectively 2 min 9 sec, 1 min 9 sec, and 48 sec for the paralympics dataset, and 4 min 4 sec, 1 min 53 sec, and 47 sec for the concert dataset. The run time progressively reduces because less tweets are considered at each subsequent level. The time for association rule extraction is about 24 sec for the cluster set at each level.

Table 1: First- and second-level clusters in the paralympics dataset (DBSCAN parameters $MinPts$=30, Eps=0.39 and $MinPts$=25, Eps=0.49 for first- and second-level iterations, respectively)

First-level clusters

Cluster	Tweets	Avg Length	Avg Sil	Words	Association Rules
C_{1_1}	70	3	1	olympic, stadium	olympic→ stadium
C_{2_1}	30	7.33	0.773	olympics, london, stadium	london→ stadium, olympics
C_{3_1}	124	4.47	0.603	london, park, athletics, day	london, day→ athletics olympic→ park, athletics
C_{4_1}	30	6.67	0.710	heats, teamgb, olympic	teamgb→ olympic heats→ teamgb
C_{5_1}	30	5.67	0.806	mens, olympic, stadium	mens→ olympic
C_{6_1}	40	6	0.620	day, pic, amazing, athletics	athletics→ amazing, day day, pic→ stadium
C_{7_1}	36	5.72	0.804	closingceremony, seats, park, stadium	closingceremony, stadium→ seats olympic, park→ closingceremony

Second-level clusters

Cluster	Tweets	Avg Length	Avg Sil	Words	Association Rules
C_{1_2}	90	5.67	0.398	flame, closingceremony, london, stadium	stadium,london→ flame,closingceremony
C_{2_2}	36	6.67	0.616	fireworks, closingceremony, hart, stadium	stadium,closingceremony→ fireworks fireworks, hart→ stadium
C_{3_2}	26	6.08	0.722	teamgb, athletics, park, olympic, london	teamgb, park→ olympic teamgb, park→ athletics olympic, park→ teamgb, london
C_{4_2}	34	9.65	0.502	greenwich, north, arena, basketball germany, final, womens	final→ north, germany final→ basketball, germany final→ womens, germany
C_{5_2}	40	6.5	0.670	fantasticfriday, dreams, time, pic olympic, london, stadium	pic, dreams→ stadium,time stadium,olympic→ london, fantasticfriday

4 Related Work

The application of data mining techniques to discover relevant knowledge from the User Generated Content (UGC) of online communities and social networks has become an appealing research topic. Many research efforts have been devoted to improving the understanding of online resources Li et al (2008a); Yin et al (2009), designing and building query engines that fruitfully exploit semantics in social networks Bender et al (2008); Heymann et al (2008), and identifying the emergent topics Alvanaki et al (2012); Mathioudakis and Koudas (2010). Research activity has been carried out to on Twitter data to discover hidden co-occurrences Li et al (2008b) and associations among Twitter UGC Cagliero and Fiori (2013); Cheong and Lee (2009); Lopes et al (2007), and analyse Twitter UGC using clustering algorithms Qing Chen (2010); Kim et al (2012); Subramani et al (2011).

Specifically, in Li et al (2008b) frequently co-occurring user-generated tags are extracted to discover social interests for users, while in Lopes et al (2007) association rules are exploited to visualize relevant topics within a textual document collection. Cheong and Lee (2009) discovers trend patterns in Twitter data to identify users who contribute towards the discussions on specific trends. The approach proposed in Cagliero and Fiori (2013), instead, exploits generalized association rules for topic trend analysis. A parallel effort has been devoted to studying the emergent topics from Twitter UGC Alvanaki et al (2012); Mathioudakis and Koudas (2010). For example, in Mathioudakis and Koudas (2010) bursty keywords (i.e., keywords that unexpectedly increase the appearance rate) are firstly identified. Then, they are clustered based on their co-occurrences.

Research works also addressed the Twitter data analysis using clustering techniques. Qing Chen (2010) proposed to overcome the short-length tweet messages with an extended feature vector along with a semi-supervised clustering technique. The wikipedia search has been exploited to expand the feature set, while the bisecting k-Means has been used to analyze the training set. In Kim et al (2012), the Core-Topic-based Clustering (CTC) method has been proposed to extract topics and cluster tweets. Community detection in social networks using density-based clustering has been addressed in Subramani et al (2011) using the density-based OPTICS clustering algorithm.

Unlike the above cited papers, our work jointly exploits a multiple-level clustering technique and association rules mining to compactly point out, in tweet collections with a variable distribution, the information posted on an event.

5 Conclusions and Future Work

This paper presents a framework for the analysis of Twitter data aimed at discovering, in a compact form, the information posted by users about an event as well as the user perception of the event. Our preliminar experimental evaluation performed on two real datasets shows the effectiveness of the approach in discovering interesting knowledge.

Table 2: First- and second- level clusters in the concert dataset (DBSCAN parameters $MinPts$=40, Eps=0.41 and $MinPts$=21, Eps=0.62 for the first- and second-level iterations, respectively)

First-level clusters					
Cluster	**Tweets**	**Avg Length**	**Avg Sil**	**Words**	**Association Rules**
C_{1_1}	148	5.05	0.817	concert, mdna, yankee, stadium	concert, yankee→ stadium concert, mdna→ yankee
C_{2_1}	340	4	1	bronx, yankee, stadium	yankee, stadium→ bronx
C_{3_1}	160	3	1	yankee, stadium	stadium→ yankee
C_{4_1}	40	6	0.950	opening, act, mdna, yankee, stadium	act→ opening yankee, stadium→ opening, act
C_{5_1}	60	6	0.779	avicii, wait, concert	wait→ yankee, avicii
C_{6_1}	84	6.19	0.794	forgive, nyc, mdna, stadium	forgive→ stadium, nyc
C_{7_1}	40	7	0.986	rain, yankee, stadium	rain→ yankee, stadium
C_{8_1}	40	6	0.751	queen, instagram, nyc	queen→ instagram
Second-level clusters					
Cluster	**Tweets**	**Avg Length**	**Avg Sil**	**Words**	**Association Rules**
C_{1_2}	60	6.67	0.523	raining, mdna, stop	raining→ mdna, stop
C_{2_2}	40	7	0.667	madge, dame, named, singing	singing→ stadium, madge madge, singing, named→ stadium, dame
C_{3_2}	44	7.64	0.535	surprise, brother, birthday, avicii, minute	yankee, stadium, surprise→ birthday
C_{4_2}	22	8.55	0.893	style, way, vip, row livingthedream	style→ vip, livingthedream

Other interesting future research directions to further improve the performance of our framework will be considering also the additional features (e.g., GPS coordinates) available in Twitter data. Furthermore, a real-time and distributed analysis of Twitter data can be addressed to support the analysis of huge data collection, also regarding parallel events.

References

Alvanaki F, Michel S, Ramamritham K, Weikum G (2012) See what's enblogue - real-time emergent topic identification in social media. In: 15th Int. Conf. on Extending Database Technology, pp 336–347

Antonelli D, Baralis E, Bruno G, Cerquitelli T, Chiusano S, Mahoto N (2013) Analysis of diabetic patients through their examination history. Expert Systems with Applications 40(11)

Bender M, Crecelius T, Kacimi M, Michel S, Neumann T, Parreira J, Schenkel R, Weikum G (2008) Exploiting social relations for query expansion and result ranking. In: IEEE 24th Int. Conf. on Data Engineering Workshop, pp 501 –506

Cagliero L, Fiori A (2013) Generalized association rule mining from Twitter. Intelligent Data Analysis 17(4)

Cheong M, Lee V (2009) Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. In: 2nd ACM Workshop on Social web search and mining, pp 1–8

DBDMG (2013) Available at http://dbdmg.polito.it/wordpress/research/analysis-of-twitter-data-using-a-multiple-level-clustering-strategy/

Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Knowledge Discovery and Data Mining (KDD), pp 226–231

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

Heymann P, Ramage D, Garcia-Molina H (2008) Social tag prediction. In: 31st Int. ACM SIGIR Conf. on Research and development in information retrieval, pp 531–538

Kaufman, L and Rousseeuw, P J (1990) Finding groups in data: An introduction to cluster analysis. Wiley

Kim S, Jeon S, Kim J, Park Y, Yu H (2012) Finding core topics: Topic extraction with clustering on tweet. In: 2012 Second International Conference on Cloud and Green Computing, CGC 2012, Xiangtan, Hunan, China, November 1-3, 2012, pp 777–782

Li X, Guo L, Zhao Y (2008a) Tag-based social interest discovery. In: 17th Int. Conf. on World Wide Web, pp 675–684

Li X, Guo L, Zhao YE (2008b) Tag-based social interest discovery. In: 17th Int. Conf. on World Wide Web, pp 675–684

Lopes AA, Pinho R, Paulovich FV, Minghim R (2007) Visual text mining using association rules. Comput Graph 31(3):316–326

Mathioudakis M, Koudas N (2010) Twittermonitor: trend detection over the twitter stream. In: ACM Int. Conf. on Management of data, pp 1155–1158

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

Qing Chen KL Shipper T (2010) Tweets mining using wikipedia and impurity cluster measurement. In: Int. Conf. Intelligence and Security Informatics, pp 141–143

Rapid Miner Project RM (2013) The Rapid Miner Project for Machine Learning. Available: http://rapid-i.com/ Last access on January 2013

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Computational and Applied Mathematics pp 53–65

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

Subramani K, Velkov A, Ntoutsi I, Kroger P (2011) Density-based community detection in social networks. In: IEEE Int. Conf. on Internet Multimedia Systems Architecture and Application, pp 1–8

Yin Z, Li R, Mei Q, Han J (2009) Exploring social tagging graph for web object classification. In: 15th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining, pp 957–966