

DISCOVERING POPULAR EVENTS FROM TWEETS

Călin Railean, Alexandra Moraru

Artificial Intelligence Department

Jožef Stefan Institute

Jamova cesta 39, 1000 Ljubljana, Slovenia

e-mail: calinraileanfirst@yahoo.com, alexandra.moraru@ijs.si

ABSTRACT

Social events happening in a city can influence and affect a large number of the citizens, directly or indirectly involved. Having a metric to measure the popularity of such events can help in estimating the resources needed for handling them, improving facilities such as public transportation or traffic estimations. This paper reports on the problem of association of Tweet messages to social events in order to discover the popularity of an event. We propose and evaluate a method that computes an association coefficient for an event-tweet pair and discuss the results obtained.

1. INTRODUCTION

The open data movement¹ has gained popularity in the last years from more and more platforms providing their data through open application programming interfaces (APIs) or from other initiatives such of open-data government or linked open data. These sources of data give data mining researchers an increasing number of problems to discover and solve. The challenges now come not from getting the data for testing hypotheses, but from finding the appropriate technologies that can handle such large volumes of data.

Social media channels are recognized as highly interactive platforms that report on all kinds of events happening around the world at any moment in time. As a micro-blogging service, Twitter generates constantly a large number of short messages that give the pulse of the communities involved in using it. Research performed on Twitter messages (i.e. tweets) has been a very popular topic in the last years, with applications ranging from sentiment analysis, to opinion mining and from topic model summarization to event extraction [1][2]. The application we are proposing is that of discovering social events popularity from tweets, where social events refer to concerts, festivals, sport events, conferences, etc. Knowing the popularity of an event can help in improving the organization of the infrastructure in the area of the event's location (i.e. scheduling of public transportation, detour paths for traffic decongesting, etc.) in case of high popularity, as well as in alerting the event's organizers for the need of better promotion of the event, in case of low popularity. We start our experiments from the simple hypothesis that the larger the number of tweets associated

to an event, the more popular the event is. Therefore the problem reduces to identifying the tweets associated with an event. We propose and implement a method that determines wheatear a tweet is associated to an event and we analyze the performance of the method proposed.

Given the streaming nature of twitter data we considered appropriate to perform our experiments following the Complex Event Processing (CEP) principles. CEP is emerging as a new paradigm for continuous processing of streaming data in order to detect relevant information and provide support for timely reactions. The main role of a CEP engine is to detect the occurrence of event patterns on the incoming streaming data [3]. We implemented our application using a CEP platform that provides classic operators for real time processing of streaming data.

The rest of the paper is organized as follows. Section 2 describes in more detail the problem addressed and introduces the event processing concepts and methods used. Section 3 presents the results of our experiments and the evaluation of the method proposed, while section 4 briefly discusses some of the related work. Finally we conclude the paper.

2. CORRELATION OF SOCIAL EVENTS AND TWEETS

The problem that we address in this paper is to determine the popularity of social events (i.e. music concerts) based on their presence in social media (i.e. tweets). We propose an algorithm that computes the degree of association between events and tweets. Furthermore, we consider that the popularity of an event directly depends on the number of tweets associated.

2.1. Dataset Description

The dataset used for experimentation consists of a set of social events and a set of micro-blogs. The social event set has been collected from an online platform² reporting on entertainment events happening in various locations, from sports and concerts to family fun and nightlife. The micro-blogs set refers to tweets collected from the public stream of Twitter³. The dataset refers to London city during the time interval from March 6th to April 11th 2013 and comprises 10033 social events and over 4 million tweets.

¹ http://en.wikipedia.org/wiki/Open_data

² <http://eventful.com/>

³ <http://twitter.com/>

The social events are separated in categories: art, music, conference, attractions, food etc. Each event has 59 fields, from which we were interested in: event title, start and stop time of event, type of event, location, performers name and their “short bio” description. The missing data for some of the events is the stop time, which we determine as the median event duration for each category.

Tweets have 90 fields that include, along with the tweet text, information such as: hash tags, time of posting tweet and geographical coordinates.

2.2. Event Processing with NEsper

NEesper is the .NET version of Esper [4] that shares the same syntax; therefore, throughout the rest of the paper we will refer to it as Esper. The Esper system provides the functionalities of an Event Stream Processing (ESP) system, as well as those of a Complex Event Processing (CEP) system. The interaction with Esper is supported by the Event Processing Language (EPL) that defines the main operators for expressing queries that are run by the engine. It is designed for a high-volume of data where one cannot store all information in database and process it in real time by using classical database queries. It is used in several areas such as finance, fraud detection, medicine where decisions need to be made as fast as possible.

The principle of the Esper system is that it allows registering queries in the engine and creates a listener class that will be called if the incoming event matches one of the inserted queries. The query can contain timeline windows, filtering, aggregation and sorting operators. Another functionality of the Esper system is to generate a new

stream as combinations between two or more input streams. The EPL statement used in our applications is the following:

```
select * from pattern [every Event ->
every Tweet (Event.Stop_Time-
Tweet.Time>0)]
```

which is similar to an inner-join statement from classical database management systems, where the join condition is represented by the time constraint. In order to obtain only the combinations that overlap over time, we use a pattern-based event stream structure, on which we specify the time constraints we want to impose. Therefore, the event stream generated will contain only the event-tweet pairs for which the timestamp for the tweet is between that start and stop time of the event or two hours before the event (the start time of the event is altered with two hours before the initial start time). For each event-tweet pair we call a method that calculates the degree of association between a tweet and a social event. To determine the degree of association, we use an *association coefficient* (AC) defined by the next formula:

$$AC = 0.5 * P + 0.25 * W + 0.125 * L + 0.125 * B,$$

where P = 1 if tweet text contains the event’s performer name, W is calculated by number of words matching between the tweet’s text and the event’s title divided by the total number of words in the event’s title, L = 1 if location name is found in tweet’s text and B = 1 if the tweet’s text contains short-bio description of the performer. The weights of these parameters are set based on a common sense understanding of their meaning.

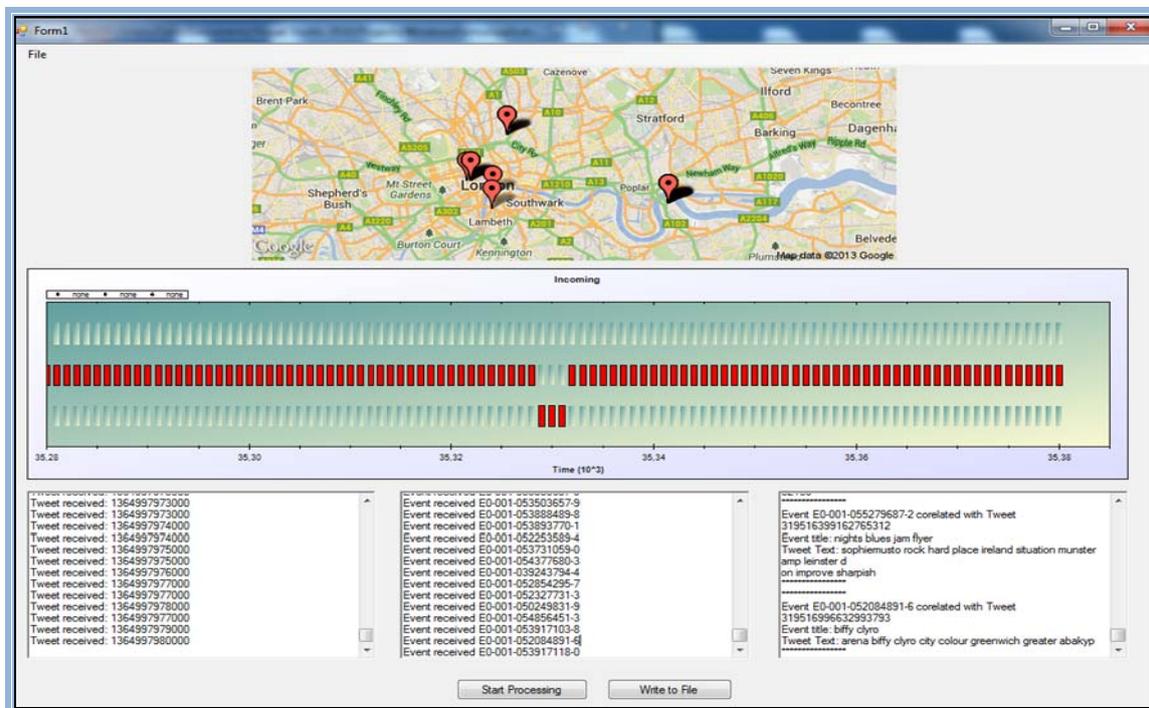


Figure 1: GUI of the application developed for running experiments

In addition, in the preprocessing steps we used a stop-word list for eliminating words such as: “this”, “and”, “that” from the event’s title and tweet’s text.

As an illustrative example, let us consider the event with the title “Halestorm”. The performer’s name is included in the title and the associated tweet is: “I’m at @eballroomcamden for Halestorm, In This Moment and Sacred Mother Tongue”. As the word “Halestorm” from the tweet is found in the event title and in the performer’s name, we assign an AC value of 0.75 (0.25 points for W and 0.5 for P).

For a better visualization of how the application processing flows, we created a graphical user interface (GUI) that shows us in real time all steps of data processing (Figure 1): locations of events, events and tweets received and some information about them. The application also registers all correlated events and tweets and gives the option to save the results at any time.

3. RESULTS

The results of the processing performed with the Esper system comprise 15455 tweets correlated with 572 music events having the AC higher than 0.25 (with an average of 27 tweets per event). From these, 5600 tweets were associated in the 2 hours interval before the event start time and the rest during the events. Figure 2 illustrates the influence of AC over the average number of tweets associated with an event, so if the AC threshold increases, the number of tweets per event decrease. Specific examples of event-tweet associations can be found in Table 1. We can observe that the association with event title “Union” is incorrect because the tweet refers to union council and not Union music band. Similarly, two more examples are incorrect because the event titles have commonly used

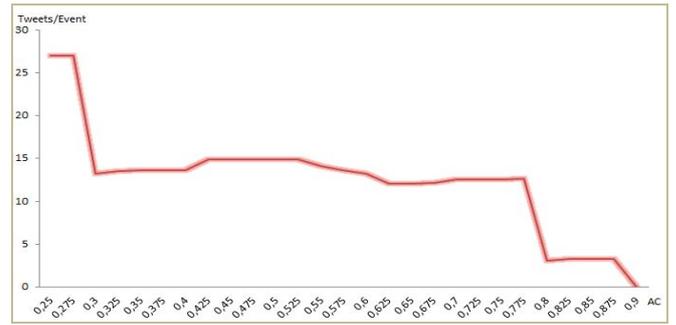


Figure 2 : Tweets-Events ratio for different values of AC

words and they are mistakenly associated with the tweet. However, the other examples show correct associations, indicating that the coefficient defined can yield positive results. Moreover, it can be observed that the highest values of AC are obtained by the events that have the name of the artist in their title. In order to analyze the performance of AC we manually evaluated a sample of event-tweet association, as it is explained in the next section.

3.1. Evaluation and Discussion

A low value of AC is understood as a low degree of association between the tweet and the event analyzed. In order to analyze the performance of AC we have manually evaluated a random set of associations of events and tweets. We first set a threshold value for AC to 0.25 and then we randomly selected 100 associations of events and tweets where AC was higher than 0.25. Two human annotators have analyzed the tweet and the event title and evaluated them as correct or incorrect. The inter-annotator agreement has been calculated in order to illustrate the utility of the annotator’s results. The inter-annotator agreement, or Cohen’s kappa coefficient, is described by the next equation [5]:

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Table 1: Example of event-tweet associations

Event Title	Tweet	AC	Event Popularity
Olly Murs	I'm at O2 Arena - @the_o2 for Olly Murs, Tich and Loveable Rogues (Greenwich, Greater London) w/ 8 others http://t.co/IDTRyUTCpb	0.75	23
Halestorm	I'm at @eballroomcamden for Halestorm, In This Moment and Sacred Mother Tongue	0.75	9
Beyonce	#NowWatching @Beyonce #LifeisButADream	0.75	9
The Script	O2 arena the script http://t.co/nlTxTURl88	0.75	54
Bastille	Seeing "Bastille". I am cool and with it. (@ O2 Shepherd's Bush Empire - @o2sbe w/ 7 others) http://t.co/BweCOsv4s5	0.75	106
Thursday Night	"#bigreunion concert on a Thursday night. Loving it!	0.25	98
Union	tweeting union council agenda avidly from @UKMStudentLive"	0.25	16
Over The Moon	Because of this, ive lost my faith in humanity! I'm done, *disappears to moon*	0.25	21
Everything on Red! -Columbia-Sabre Tooth Monk–Metropolis-Dogdaze	No red card will ever too that	0.535	118

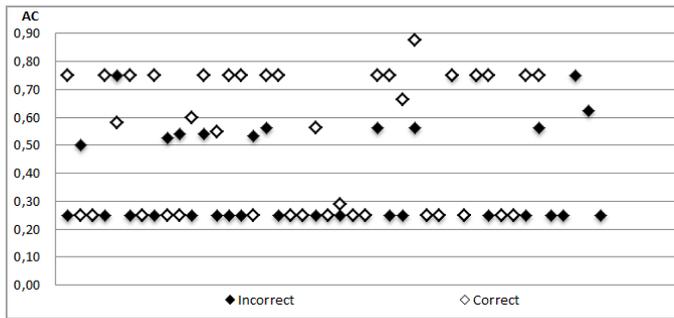


Figure 3: Values of AC for the associations of tweets and events evaluated

where $Pr(a)$ is the relative observed agreement among annotator, and $Pr(e)$ is the hypothetical probability of chance agreement. We obtained a Cohen's kappa coefficient equal to 0.661 for the two annotators, which can be considered as a "substantial" level of agreement [5].

The next step was to analyze the annotated associations, considering only those for which the annotators gave the same score. The analysis revealed that the average value of AC for the correct associations is higher than the average value of AC for the incorrect ones: 0.52 and 0.36, indicating that the AC defined is a fair metric for association. Further analysis of the performance of AC is done in Figure 3 and Figure 4. Figure 3 shows that the values of AC for correct associations are generally higher than the values of AC for incorrect ones.

Next, we have analyzed the precision of correct associations, calculated as the ratio between the number of correct associations and the total number of associations. As, expected, the precision of tweet event associations increases with the increase of AC, and a more detailed relation between AC and precision can be observed in Figure 4. Although the recall performance would have been an interesting measure to analyze, it was considered too expensive to be done manually. A larger dataset fully annotated would be more appropriated for such an analysis.

4. RELATED WORK

A large amount of research on Twitter messages has been performed and reported in literature and a full comparison of the work presented with other work is out of the scope of this paper. However, we would like to mention the work reported in [1] where a system for processing tweets in real time is introduced. The applications tackled refer to sentiments analysis and detection of term frequencies in real time. Although our method is not comparable in terms of complexity with the methods proposed in [1], the similarity can be found on the stream processing concept. Another similar problem is reported in [2], where the problem addressed is that of linking tweets with news articles. The authors propose a graph based latent variable model for enriching the short text of tweets in order to create a larger context for it. Finally, in [6], another study over open data sources for London city presents the results

of possible associations between social events, weather data and traffic.

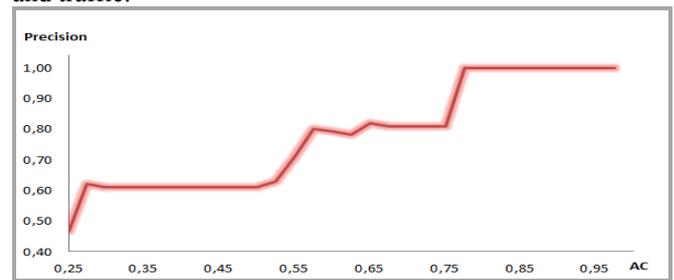


Figure 4: Precision performance for different values of AC

5. CONCLUSIONS

We have proposed and evaluated a method for discovering the popularity of social events happening in the city of London, based on tweets. The results show that the method proposed yields a positive outcome and is a valid solution for the problem addressed. The number of false positive associations of events and tweets can be decreased by setting a higher threshold for the AC coefficient. Further improvements of the method may be brought by including geo-location parameters in the AC equation, as well as by improving the preprocessing of data in terms of extension of the stop-word list or by including natural language processing techniques.

Acknowledgements

This work was supported by the Slovenian Research Agency and the ICT Programme of the EC under PlanetData (ICT-NoE-257641).

References

- [1] Bifet, A., Holmes, G., Pfahringer, B. MOA-TweetReader: Real-Time Analysis of Twitter Streaming Data. In: *Discovery Science*. Springer Berlin Heidelberg, 2011. p. 46-60.
- [2] Guo, W., Li, H., Ji, H., & Diab, M. Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media. In: *The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*.
- [3] Cugola, G., Margara, A.: Processing Flows of Information: From Data Stream to Complex Event Processing. *ACM Computing Surveys (CSUR)*, 2012, 44.3: 15.
- [4] EsperTech Inc, Esper Reference, Version 4.8.0, 2012, 676 pages. http://esper.codehaus.org/esper-4.8.0/doc/reference/en-US/pdf/esper_reference.pdf (last access date: 28.08.2013)
- [5] Cohen's kappa. http://en.wikipedia.org/wiki/Cohen's_kappa (last access date: 28.08.2013)
- [6] Moraru, A., Mladenici, D. Complex Event Processing and Data Mining for Smart Cities. In: *IS 2012*, 8-12 Oct 2012, Ljubljana, Slovenia.