REINA at RepLab2013 Topic Detection Task: Community Detection*

José Luis Alonso Berrocal[1], Carlos G. Figuerola[1], Ángel Zazo Rodríguez[1]

[1] Science and Technology Institute. Research Group REINA. University of Salamanca
{berrocal,figue,zazo}@usal.es

Abstract. Social networks have become a large repository of comments which can extract multiple information. Twitter is one of the most widespread social networks and larger and is therefore an important source for detecting states of opinion, events and happenings before even the mainstream media. Topic detection is important to discover areas of interest that arise in the tweets. We have used classical systems for a similarity matrix and we have used community detection techniques. The results have been good and allows us to study new possibilities.

Keywords: Topic detection, Community detection

1 Introduction

Social networks, blogs, or any online forum internet have become a large repository of comments which can extract multiple information [1] as shown by numerous research being carried out in recent years

Twitter is one of the most widespread social networks and larger and is therefore an important source for detecting states of opinion, events and happenings before even the mainstream media [2], [3].

It can be used to share information and also to describe virtually any daily activity [4], because it allows users to express their opinions and interests, abbreviated and highly personalized in real time [5]. Its importance is shown where it is present in virtually all areas of life (social, economic, education ...) and covers any topic (sports, culture, entertainment, industry, science).

* Financed by the project of the Ministry of Science and Innovation FFI2011-27763

If we value the importance of this network in quantitative terms, it is necessary to refer to the volume of tweets generated every day, in June 2011, about 200 million, a number that is increasing- [6]. If assessment in qualitative terms we do need to consider how their influence is reflected in the many social events that retrasmiten real time in order to gain visibility as well as the large number of original messages that become spread (retweets), so you can consider even that Twitter does become niche opinion, since a message created by a person (either original or a fragment of another work as a newspaper headline or an extract of news) can be retweeted by another or others who in turn relay it again causing a diffusion effect in clusters.

It is true that much of the information provided is completely irrelevant tweets, it is also necessary to consider that in many cases the messages isolated from their context lose value, but also a very rich source of information because it compiles the relevant information condensed for users, whether individuals, institutions or companies that are the highlights news, opinions or feelings, information would be very difficult to collect by other channels and is therefore by analyzing twitter messages is being used as feedstock for multiple investigations ranging from the role played by different types of users in the dissemination of information [7] a sociological analysis [8], applications to classification [6] and information retrieval [5] [9], semantic analysis [10], etc.

In this sense, the monitoring of comments, messages and opinions that are poured into Twitter is useful from the point of view of digital reputation people and institutions. Early detection of issues and ratings on a particular subject can allow it to react appropriately and maintain a positive public image [11].

2 Topic detection

Our group has focused on the topic detection, starting some exploratory work in this field. Topic Detection and Tracking (TDT) is an area that began as track in the Text Retrieval Conferences (TREC) [12] and this year is celebrated in RepLab2013 [13]. However, the application of these techniques is relatively new to Twitter. Some notable works are those of [14], who applied clustering techniques, or Petrovic [15], which also was an experimental collection of tweets that has been used in other studies, the Edinburgh Twitter Corpus [16]. Mathioudakis and Koudas [17] proposed a system for detecting trending topics from a stream of tweets. Also on the detection of trending topics have worked Shariffi, Hotton and Kalita [18], like Cheong and Lee [19], although they work focuses on the temporal evolution of the trending topics.

The question is therefore how to determine the similarity between all pairs of tweets relating to a given entity. The similarity between two documents is one of the central problems in information retrieval and can be approached in various ways. One of the best known is to consider each document (tweet) as a bag of words and apply a classical scheme tf x idf.

The tweets, however, are documents with a number of special features that should be taken into account. Anta and colleagues [20] mention several of them. A issues 'classic' of using unigrams, bigrams, trigrams, etc. or stemming, in the case of tweets must add the emoticons, abbreviations, including a slang that medium itself, and as numerous abnormalities ortho-typographic. The brevity of the tweets is another important issue to consider [21]. In our specific case, we find texts at least two possible languages: English and Spanish.

Finally we will make the graph representation of the tweets of the entities and we apply Social Network Analysis to our information [22]. Social Network Analysis is a measurement tool allowing knowledge and structural analysis of the interactions between the actors of the analyzed network [23].

There is a wide range of indicators such as density, centrality, centralization, betweenness, closeness, etc.. that allow analysis of both network nodes as complete, although the detection of communities, groups. cliques, etc.., is a subject of great interest.

The strategy adopted in this work has been the application of techniques of Social Network Analysis, in particular communities detection techniques. In a social network G = (V,E) a community is a subgraph of entities $Vc \subseteq V$ that are associated with common elements of interest. The elements that are part of the community can be topics, real-life people, places, events, etc. These techniques are based on detecting, in a network node groups with strong bonded between them. In our case the tweets would be the nodes of the network; a semantic similarity between two tweets mean a link between network nodes.

There are many techniques for detecting communities [24] [25-27] as hierarchical clustering algorithms, methods based on cliques, grouping cuts, Girvan-Newman algorithm, etc.
One widely used method is the analysis of modularity [28] (the number of links between groups is small, within groups is high), highlighting the Louvain algorithm [29].

One method that is effective is showing the VOS clustering algorithm (visualization of similarities) and some jobs are proving more effective compared to other systems, especially for better performance than systems based on modularity in detecting small clusters [30]. It is a modification of the algorithm based on modularity where the weights are maximized differently [31].

Regarding VOS clustering technique, we can use the mapping to visualization VOS is very effective compared to other methods, adding a plus detection systems communities [32]. In this map, the colors indicate the density within each community, ranging from blue (low density) to red (high density). We can see the most important communities and placed in relation to each other.

3 Our approach

Since this is the first time we participate in this work, our focus has been simple and without too many refinements. We have considered each tweet (within each entity) as a document whose basic features are the words it contains, and we have applied after heavy classic scheme tf x idf and cosine to construct a similarity matrix [33]. Some specific issues applied in our work have been:

- we have not made any distinction between languages of the tweets, possibly there are notable differences in the treatment applicable according to the language it is one or other one [20, 34] but in our case we have performed uniform lexical analysis all tweets
- we applied a simple s-stemmer
- we removed the words with less than 4 characters

Additionally, we have considered discarded emoticons. We have considered hashtags and entities terms particularly interesting.

On the other hand, in numerous tweets appear weblinks, we have considered especially interesting, if two tweets have links to the same website we think that dealing with very similar issues. Thus, the URL of these links are considered equally as important characteristics of terms or the tweets.

Given the small number of terms present in a tweet, the co-occurrence of URLs, hashtags and entities are especially significant. Some studies apply techniques designed to increase the number of terms per tweet [35], following the links and adding to the features of that tweet the words of the website referenced. Anta and colleagues [20], however, report the amount of noise that this technique produces.

Other refinements possible, as the use of Wikipedia [36] for additional information and produce more accurate results have not been applied by us on this occasion.

Once the network weighted with the weights of the similarity, we proceeded to generate individualized networks for each of the entities under study. We obtained the number of communities (by VOS Clustering algorithm) of each of the entities, we have individualized the communities and thereafter we performed calculation on the density of each of the communities.

When we boarded the density term relationships and social networks we refer to a widespread concept. This can be defined as the proportion of links in a network relative to the total possible links (sparse versus dense networks). Other authors density defined as the interface between network members. The density is an indicator of social network analysis allows us to measure the extent to which a network is connected.
We can say further that a dense network nodes have a very close relationship between them, confirming the theory that "the performance of a network has a positive association with the high density of the network"

With these data we created two tasks:

1. reina_1: Topics were assigned to all tweets, depending on the community to which they belonged. Topic was assigned to all tweets, even if the community consisted of few documents.
2. reina_2: Filter was performed according to the density of each of the entities. We considered only communities with a density greater than 0.5. Topic was assigned only tweets belonging to these communities.

4 Results

The results of our two task (reina_1 and reina_2) were as follows:

Table 1. Measure F and ratio of processed tweets

RUN	Rel.	Sen.	F	Ratio
re-plab2013_UNED_ORM_topic_detection_2	0.46	0.33	0.32	0.98
reina_2	0.31	0.43	0.29	0.79
lia_topic_detection_3	0.22	0.35	0.25	0.99
lia_topic_detection_2	0.23	0.27	0.24	0.99
re-plab2013_UNED_ORM_topic_detection_7	0.30	0.22	0.24	0.99
UAMCLYR_topic_detection_07	0.35	0.50	0.24	0.97
re-plab2013_UNED_ORM_topic_detection_3	0.42	0.21	0.23	0.99
re-plab2013_UNED_ORM_topic_detection_4	0.42	0.21	0.23	0.99
re-plab2013_UNED_ORM_topic_detection_5	0.42	.,21	0.23	0.99
lia_topic_detection_1	0.38	0.17	0.23	0.99
reina_1	0.16	0.52	0.23	0.99

The result of the measure F[37] (table 1), as we can see has given better results the task reina_2 and furthermore their behavior with respect to the rest of the task has been very good. Note that the ratio obtained for this task is the lowest of all the set, (density filter). This filtering requires a revision in the threshold used to improve the ratio of tweets.

Table 2. Amount of improved systems

SYSTEM	Amount of improved systems (UIR>0.2)
UAMCLYR_topic_detection_07	12
replab2013_UNED_ORM_topic_detection_2	11
reina_2	9

Concerning the amount of improved systems (table 2), we can see that again reina_2 task behavior is better than reina_1 and also offers good results with respect to total tasks.

Table 3. System pair, improvements and UIR

System Pair		Amount of cases in which A improves B for both measures	Amount of cases in which B improves A for both measures	UIR
lia_topic_detection_3	replab2013_UNED_ORM_topic_detection_1	51	0	0.84
lia_topic_detection_3	BASELINE	51	1	0.82
replab2013_UNED_ORM_topic_detection_2	replab2013_UNED_ORM_topic_detection_1	47	0	0.77
lia_topic_detection_2	BASELINE	45	0	0.74
replab2013_UNED_ORM_topic_detection_2	BASELINE	45	0	0.74
lia_topic_detection_2	replab2013_UNED_ORM_topic_detection_1	44	0	0.72
replab2013_UNED_ORM_topic_detection_2	lia_topic_detection_4	43	0	0.70
reina_2	replab2013_UNED_ORM_topic_detection_1	43	1	0.69
reina_2	BASELINE	42	1	0.67

In the comparison of the system pairs and UIR [38] (table 3) reina_2 improves to reina_1 and continues to maintain good results with respect to total tasks of track.

With the working method we can visualize detected communities for a given entity. Show two different views (Fig. 1 and Fig 2)of the entity RL2013D01E002, allowing us to obtain a representation of these communities, which eventually become the detection of specific topics in the entity.

This working method allows us to simultaneously perform a complete entity reduction in their various communities and establish the relationships between these communities, which offers a mechanism for relations between communities, and therefore the relationship among topics (Fig. 3).

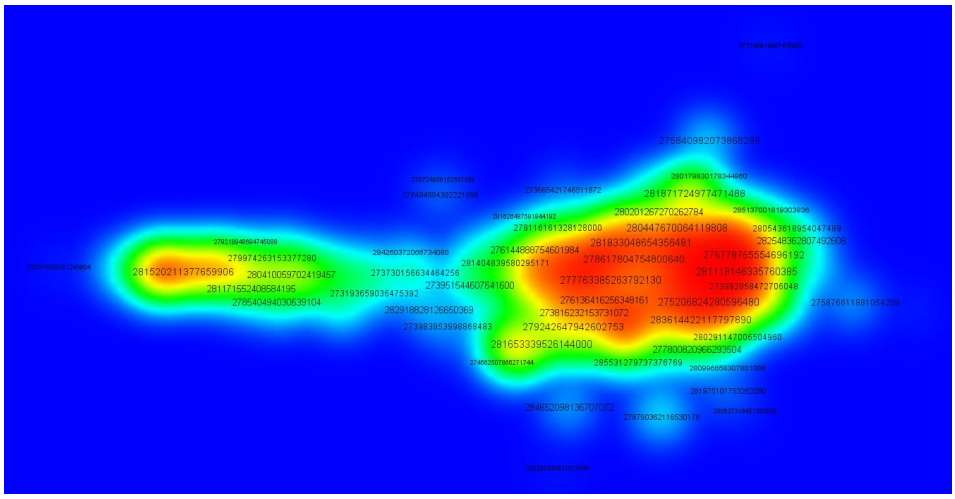

Fig. 1. Density of the communities (VOS mapping). Entity RL2013D01E002

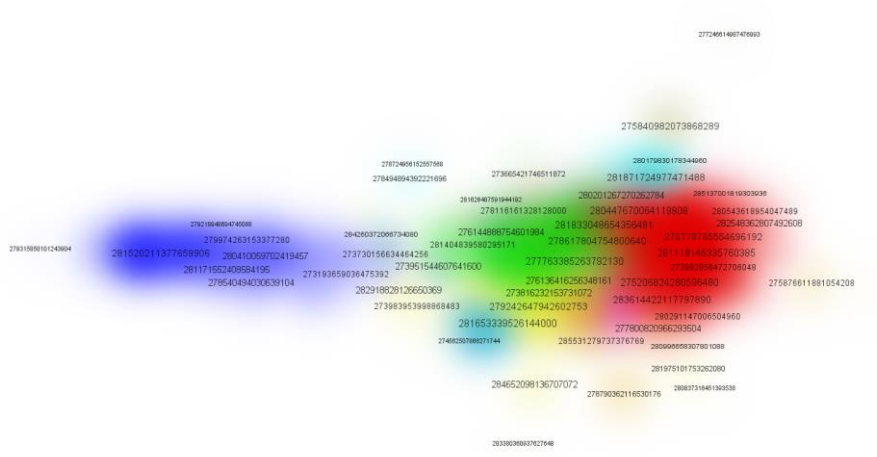

Fig. 2. Detected communities. Entity RL2013D01E002

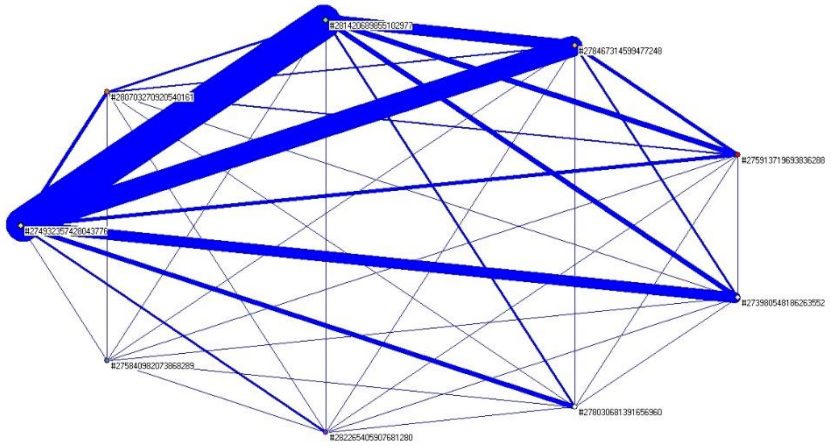

Fig. 3. The most related communities (Topics). Entity RL2013D01E002

5 Results

We have raised a system of detection of topics differently but that has given a few good enough results.

Mixing basic scheme for generating the similarity matrix, and detection of communities promising results.

The use of a filtering network density has better result than without filtering.

The threshold used in filtering lowered the ratio of processed tweets.

In the future we need to try different schemes when generating the similarity matrix, try different community detection algorithms and use other filtering techniques.

6 Bibliography

1. Thelwall, M., K. Buckley, and G. Paltoglou, *Sentiment in Twitter events.* Journal of the American Society for Information Science and Technology, 2011. **62**(2): p. 406-418.
2. O'Brien, T., *Twitter breaks news of plane crash in the Hudson.* 2009.
3. Kwak, H., et al. *What is Twitter, a social network or a news media?* in *Proceedings of the 19th international conference on World wide web*. ACM.

4. Java, A., et al., *Why we twitter: An analysis of a microblogging community*, in *Advances in Web Mining and Web Usage Analysis*2009, Springer. p. 118-138.
5. Garcia Esparza, S., M.P. O'Mahony, and B. Smyth, *Mining the real-time web: a novel approach to product recommendation.* Knowledge-Based Systems, 2012. **29**: p. 3-11.
6. Lee, K., et al. *Twitter trending topic classification*. in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. 2011. IEEE.
7. Cha, M., et al., *The world of connections and information flow in Twitter.* Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 2012. **42**(4): p. 991-998.
8. Chen, S.-C., D.C. Yen, and M.I. Hwang, *Factors influencing the continuance intention to the usage of Web 2.0: An empirical study.* Computers in Human Behavior, 2012. **28**(3): p. 933-941.
9. Yerva, S.R., Z. Miklós, and K. Aberer, *Quality-aware similarity assessment for entity matching in Web data.* Information Systems, 2012. **37**(4): p. 336-351.
10. Narr, S., E.W. De Luca, and S. Albayrak. *Extracting semantic annotations from twitter*. in *Proceedings of the fourth workshop on Exploiting semantic annotations in information retrieval*. 2011. ACM.
11. Jansen, B.Z.M.S.K. and A. Chowdury, *Twitter power: Tweets as electronic word of mouth.* Journal of the American Society for Information Science and Technology, 2009. **60**(11): p. 2169-2188.
12. Allan, J., *Topic detection and tracking*, J. Allan, Editor 2002, Kluwer Academic Publishers. p. 1-16.
13. Amigó, E., et al. *Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems*. in *Fourth International Conference of the CLEF initiative, CLEF 2013, Valencia, Spain. Proceedings*.
14. Sankaranarayanan, J., et al. *TwitterStand: news in tweets*. in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM.
15. Petrović, S., M. Osborne, and V. Lavrenko. *Streaming first story detection with application to Twitter*. in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.
16. Petrović, S., M. Osborne, and V. Lavrenko. *The Edinburgh Twitter corpus*. in *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*. Association for Computational Linguistics.
17. Mathioudakis, M. and N. Koudas. *TwitterMonitor: trend detection over the twitter stream*. in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. ACM.
18. Sharifi, B., M.-A. Hutton, and J.K. Kalita. *Experiments in Microblog Summarization*. in *Proceedings of the 2010 IEEE Second International Conference on Social Computing*. IEEE Computer Society.

19. Cheong, M. and V. Lee. *Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base*. in *Proceedings of the 2nd ACM workshop on Social web search and mining*. ACM.
20. Anta, A.F., et al., *Sentiment Analysis and Topic Detection of Spanish Tweets: A Comparative Study of of NLP Techniques.* Procesamiento del Lenguaje Natural, 2012. **50**: p. 45-52.
21. Sriram, B., et al. *Short text classification in twitter to improve information filtering*. in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM.
22. Wasserman, S. and K. Faust, *Social Network analysis: methods and applications*1998, Cambridge: Cambridge University Press.
23. Molina, J.L., *El análisis de redes sociales: una introducción*2001, Barcelona: Bellaterra.
24. Porter, M.A., J.P. Onnela, and P.J. Mucha, *Communities in networks.* Notices of the American Mathematical Society, 2009. **56**(9): p. 1082-1097.
25. Fortunato, S., *Community detection in graphs.* Physics Reports, 2010. **486**(3): p. 75-174.
26. Papadopoulos, S., et al., *Community detection in social media.* Data Mining and Knowledge Discovery, 2012. **24**(3): p. 515-554.
27. Wang, Z., *Detection of overlapping communities in networks: a probabilistic approach.* 2012.
28. Blondel, V.D., et al., *Fast unfolding of communities in large networks.* Journal of Statistical Mechanics: Theory and Experiment, 2008. **2008**(10): p. P10008.
29. De Meo, P., et al. *Generalized louvain method for community detection in large networks*. in *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*. 2011. IEEE.
30. Van Eck, N.J., *Methodological advances in bibliometric mapping of science*2011: Erasmus University Rotterdam.
31. Waltman, L., N.J. van Eck, and E. Noyons, *A unified approach to mapping and clustering of bibliometric networks.* Journal of Informetrics, 2010. **4**(4): p. 629-635.
32. Van Eck, N.J., et al., *A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS.* Journal of the American Society for Information Science and Technology, 2010. **61**(12): p. 2405-2416.
33. Salton, G. and C. Buckley, *Term-weighting approaches in automatic text retrieval.* Information Processing and Management, 1988. **24**(5): p. 513-523.
34. Qureshi, M.A., C. O'Riordan, and G. Pasi. *Concept Term Expansion Approach for Monitoring Reputation of Companies on Twitter*. in *CLEF (Online Working Notes/Labs/Workshop)*.
35. Benhardus, J. and J. Kalita, *Streaming trend detection in twitter.* International Journal of Web Based Communities, 2013. **9**(1): p. 122-139.
36. Osborne, M., et al. *Bieber no more: First story detection using Twitter and Wikipedia*. in *SIGIR 2012 Workshop on Time-aware Information Access*.
37. Amigó, E., J. Gonzalo, and F. Verdejo. *A General Evaluation Measure for Document Organization Tasks*. in *Proceedings SIGIR 2013 /07*.

38. Amigó, E., et al., *Combining evaluation metrics via the unanimous improvement ratio and its application to clustering tasks.* Journal of Artificial Intelligence Research, 2011. **42**(1): p. 689-718.