Twitter User Clustering Based on Their Preferences and the Louvain Algorithm

Daniel López Sánchez, Jorge Revuelta, Fernando De la Prieta, Ana B. Gil-González and Cach Dang

Abstract In this paper, a novel agent-based platform for Twitter user clustering is proposed. We describe how our system tracks the activity for a given topic in the social network and how to detect communities of users with similar political preferences by means of the Louvain Modularity. The quality of this clustering method is evaluated against a subset of human-labeled user profiles. Finally, we propose combining community detection with a force-directed graph algorithm to produce a visual representation of the political communities.

Keywords Clustering · Data mining · Community detection · Visualization

1 Introduction

During the last years, the number of users in social networks has grown exponentially. Many individuals from all around the world share publicly their opinions, likes, dislikes and interests. In this context, the technological challenge consists of designing and deploying new information systems that enable us to mine the massive amount of information available to successfully extract knowledge from it.

In addition to mining the information that the users provide in the social media, it is also interesting to analyze the way users aggregate and form communities.

D.L. Sánchez · J. Revuelta · F. De la Prieta(✉) · A.B. Gil-González
Department of Computer Science and Automation Control, University of Salamanca, Plaza de la merced s/n, 37007 Salamanca, Spain
e-mail: {lope,jrevuelta,fer,abg}@usal.es

C. Dang
HoChiMinh City University of Transport (UT-HCMC), Ho Chi Minh City, Vietnam
e-mail: tucach@hcmutrans.edu.vn
http://www.hcmutrans.edu.vn/en/

F. de la Prieta et al. (eds.), *Trends in Pract. Appl. of Scalable Multi-Agent Syst., the PAAMS Collection*, Advances in Intelligent Systems and Computing 473,
DOI: 10.1007/978-3-319-40159-1_29

For instance, if it is designing a political strategy it might be of interest to analyze the different communities of users that support and detract a specific topic. It could be possible to track their activity independently to design specific actuations.

In this article, it is proposed a novel agent-based platform to identify communities of users with the same opinion alignment in Twitter. Section 2 explains the state of the art about information retrieval. In section 4, we present the summary of the platform and the approach that it is used to detect communities in this network. Section 4 presents the results of the experiment we conducted to determine the degree of political aggregation of the users. Finally, in section 5 we explain how to integrate the results of the Louvain algorithm with a force-directed graph algorithm to produce a consistent visualization of the communities in a 2D surface.

2 Data Extraction, State of the Art

The tremendous growth and development of the Web produce a collapse of information to any user interested in access to quality information [12]. Additionally, another difficulty is the necessity to automatically extract information from different and heterogeneous sources [3]. To cope with this challenges, there are research oriented studies and application specially oriented to retrieve and organize information, especially nowadays in social networks [11][13].

The vast amount of content available on the Internet, provokes that the end-users can not exploit all its usefulness if they do not have the adequate tools to retrieve and display useful information organized. Much of the needs of Web information recovery for users is solved with the usage of search engines [8]. However, there are still open challenges on the organization of information in quantities priori impossible to handle for users [6].

Formally, the term information retrieval usually refers to the query and data mining both structured and unstructured data. The solutions developed by search engines or conventional web browsers are very effective to recover the visible contents of the Web. They make use of web crawlers or spiders [10]. These spiders crawl websites and recursively follow hyperlinks in the documents. The data extracted by the spiders are treated differently by the various search engines. The difficulty that exists in dealing with all the huge amount of information stored on the web is the homogeneity of the sources, which terribly difficult their harvesting. Virtually every website has its own way of representing information: i.e. social networks like Facebook or Twitter have their own APIs for accessing data, blogs and online newspapers have the most diverse information structures and navigation, etc.

At the end, crude extraction of the contents of the web and social networks is all stages of analysis which is more developed and involves the least of the problems web content. The difficulty increases when we talk about the treatment of the information obtained to get useful information. To deal with this challenge, specially in social networks, this article presents a novel agent-based platform to harvest information and analyze it.

3 Agent-Based Platform

The proposed architecture is based on organizational aspects and, therefore, it is necessary to identify the organizational structure to be used. For this reason, the first step has been to identify its components, which allows for the interaction model based on the analysis of the needs of potential users of the system. Subsequently, from this analysis it has been possible to deduce the roles of users and components involved in the system and how they will exchange information. Fig. 1. show the platform obtained from this approach including organizations and agents.

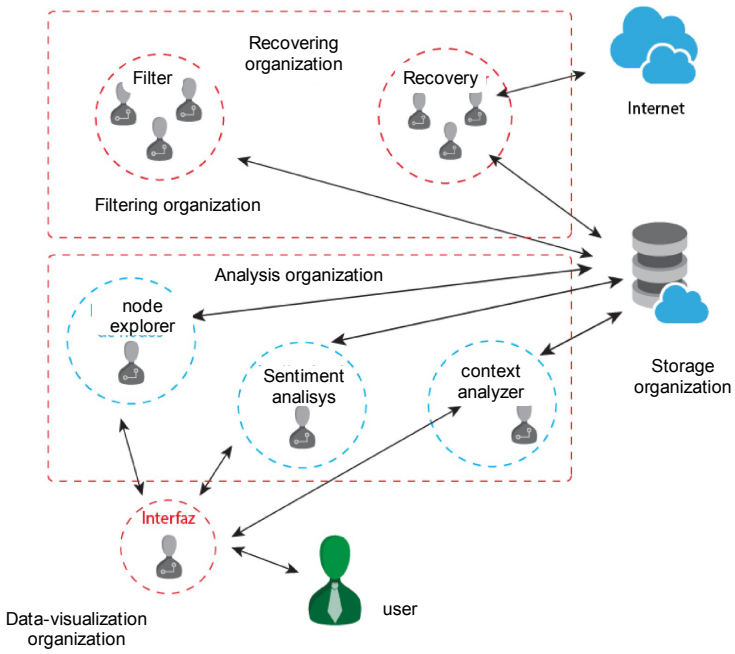

Fig. 1 Agent-based platform

Recovery agents use the stream API of Twitter to extract all the tweets published in Spanish that contain a specific list of keywords; for each recovered tweet, they perform a request to the REST API of Twitter to obtain information concerning the user who published the tweet (e.g. followers, retweets, followings…).

Filter agents keep the extracted information up to date: as several attributes of social users may vary over time, it is necessary to keep polling Twitter periodically to update fields like the followers and followings of social users.

The agents of the Analysis organization perform the different data mining algorithms described in the following sections. Although our platform is capable of performing automatic sentiment analysis (via the sentiment analysis agent) this feature is not applied to the case study of this paper.

The main objective of this article is not to describe the proposed architecture, more details about it, can be read in [14]. Following sections describe the tasks performed by agents of the analysis organization.

3.1 Louvain Modularity for User Clustering

The Louvain Modularity [1] is one of the most widely used methods to extract communities from networks of any kind. It is especially interesting when other methods for community extraction are not applicable due to the size of the network, both in terms of the number of nodes and links. The computational complexity of the algorithm is not known, but it has been shown empirically that it can be computed over a network with n nodes at the cost of time $O(n \log(n))$.

The method follows a greedy optimization strategy, trying to optimize the modularity of a partition of the network. The modularity [7] is a metric that takes values inside the interval [-1, 1]; it measures the density of links inside communities compared to links between communities.

The modularity is optimized by means of two phases or steps that are repeated iteratively:

1. each node of the network is assigned to its own community and the modularity is optimized locally.
2. nodes in the same community are grouped and a new network is build where nodes are communities from the previous step.

The links between nodes of the same community are represented with a self-loop on the community node of the new network and links between nodes in different communities are represented as weighted links between community nodes. Figure 2 shows a sample iteration of the algorithm.

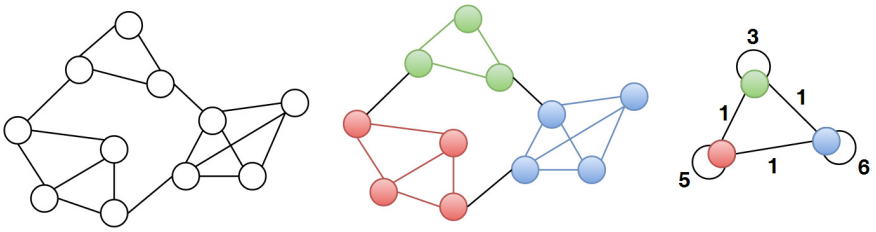

Fig. 2 First iteration of the Louvain algorithm on a simple network

The proposed system, uses the Louvain method to detect communities based on the information extracted from twitter. Note that, the Louvain method does not consider the direction of links within the network to detect communities. In the other hand, the relations between twitter users have a semantical direction: if one user *follows* another one, not necessary the second user follows the fist.
We defined two manners of transforming twitter *following* relations to links in the network that will be segmented in communities by means of the Louvain method. In the first approach, we settle a link between two users (i.e. nodes of the network) if and only if they follow each other, we call this a *hard link*. In the second approach, we place a link between two users if one of them follows the other, or when both of them follow each other; we call this a *soft link*.

4 Community Purity Evaluation

Our goal is to obtain a split of the extracted network such that the preference of the users is homogeneous inside communities. This is possible since many social networks exhibit similar properties [5] (i.e. users prefer to connect to those more like themselves). To evaluate the proposed system, it is used a small set of users whose political alignment (i.e. left or right alignment) has been labeled manually by a human expert. Then we apply the purity metric to evaluate how well our community detection matches that gold standard.

Purity is an external criteria of clustering evaluation. It is computed by assigning to each community the class or label that is most frequent in that community. Then the number of correctly assigned nodes is counted and divided by the total number of nodes in the network. Formally:

$$Purity = \frac{\sum_{\mathrm{i}} max_j \, n_{ij}}{\sum_{\mathrm{ij}} n_{ij}} \tag{1}$$

Where i is the index for communities, j is the index for ground truth labels and n_{ij} is the number of nodes with label j assigned to community i. An example of this can be seen in figure Fig. 3.

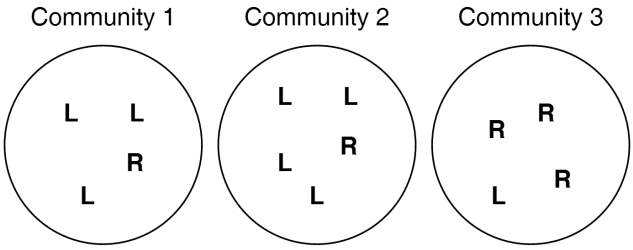

Fig. 3 Most frequent label and number of nodes of the most frequent label for the three communities are: L, 3 (community 1); L, 4 (community 2); R, 3 (community 3). Purity is $(1/13) \cdot (3 + 4 + 3) \approx 0.76$

To evaluate whether if using *hard links* or *soft links* produces a better community detection the following experiment was conducted. The activity concerning a current controversial topic on Twitter is tracked, i.e. the political situation of Catalonia (Spain). A total of 1575 users tweeted about the topic during a period of two days, it was extracted a total of 8448 *hard links* between those users. A human expert evaluated a number of user profiles, labelling then as left-aligned, right-aligned or ambiguous; this yield a total of one hundred non-ambiguous user profiles. This expert based knowledge was used as a gold standard and computed the purity of clustering this one hundred manually-labelled users according to de communities detected by our system in the complete network with 1575 users. Using the *hard link* approach lead to a purity of 0.95 whereas using the *soft link* criteria the purity was of 0.98.

This evidences that the *soft link* approach produces a community detection where users are more accurately grouped by their political alignment.

4.1 Visual Representation: Force-Directed Graph

It is very useful to generate a visual representation of the communities detected by the Louvain algorithm. For this purpose we used a force-directed graph layout as implemented by D3 library [2].

Force-directed graph algorithms are a family of algorithms designed to create graph visual representations in an aesthetically pleasing way. They perform a physical simulation to decide the final location of each node in a 2D surface; an attractive force is simulated for each pair of linked nodes, as well as a repulsive force between nodes. The attractive force is often simulated according to Hooke's law, while repulsive force considers nodes as infinitesimal points with equal charge and mass and thus the repulsion is computed according to Coulomb's law. Additionally, D3 implements a pseudo-gravity force that keeps nodes centered in the visible area and avoids expulsion of disconnected subgraphs.

Simulating such an n-body system would have a computational complexity of $O(n^2)$, to overcome this problem D3 uses the Barnes–Hut [9] approximation algorithm. In this, a quadtree is applied to accelerate the charge interactions between the particles, reducing the computational complexity to $O(n \log(n))$.

To ensure that the nodes belonging to the same community are drawn together and that the communities do not overlap in the representation, we modified the force-directed graph algorithm. In our version, a link existing between nodes of the same community (as detected by the Louvain algorithm) produces an attractive force ten times stronger that those links between nodes in different communities.

Figures 4 and 5 show an example community detection, specifically the one used in the previous section for our experiments. The images show that the communities detected when considering *soft links* are more fine-grained, which explains why the purity was higher in that case.

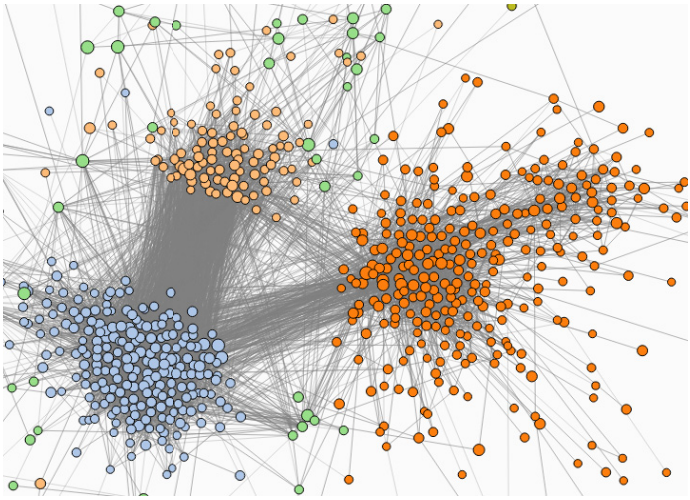

Fig. 4 Community visualization (*hard links*)

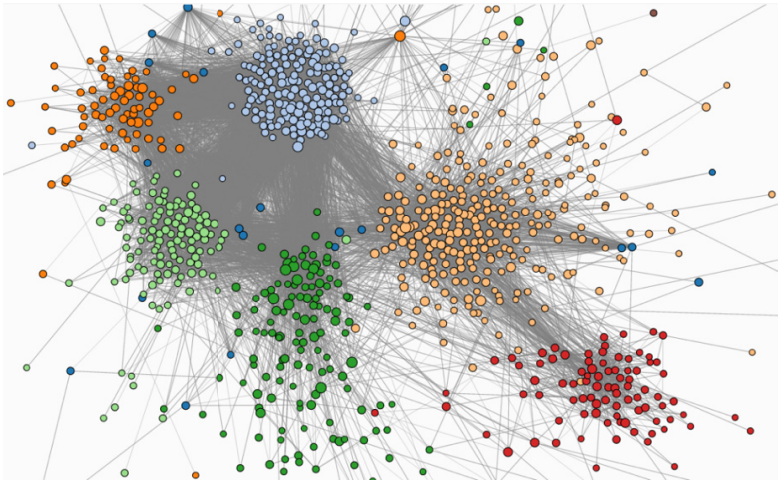

Fig. 5 Community visualization (*soft links*)

5 Discussion and Future Work

In this paper, a novel framework for twitter user clustering was proposed. The empirical results presented in section 4 suggest that *soft links* (i.e. unidirectional following relations) are more suited for political community detection than *hard links*. This technique can be applied to detect, analyze and visualize communities of users with the same political preferences.

The proposed setup could be modified to be used as a simple classifier that could predict the political alignment of a given user. If a few nodes from each community

are manually labelled by a human expert, then the resulting community map can be used as a multiclass classifier where each new user is assigned to the community (and therefore class) that contains most of the users he is following. It would be also interesting to perform a similar study considering the political party that users support, and not only their political alignment (left or right winged). This could be also applied to cybersecurity; detecting communities of users with extremist ideologies could help security forces to anticipate riots and other violent events. The recent advances in automatic sentiment analysis [4] could be integrated in the system to automatize the task of conflictive community vigilance.

Acknowledgments This work is supported by the Ministry of Industry, Energy and Turism, Project PIAR (TSI-100201-2013-20).

References

1. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10), P10008 (2008)
2. Bostock, M., Ogievetsky, V., Heer, J.: D^3 data-driven documents. IEEE Transactions on Visualization and Computer Graphics **17**(12), 2301–2309 (2011)
3. Klusch, M. (ed.): Intelligent information agents: agent-based information discovery and management on the Internet. Springer Science & Business Media (2012)
4. Liu, B.: Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies **5**(1), 1–167 (2012)
5. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual review of sociology, 415–444 (2001)
6. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, pp. 29–42. ACM, October 2007
7. Newman, M.E.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences **103**(23), 8577–8582 (2006)
8. Nguyen, D., Demeester, T., Trieschnigg, D., Hiemstra, D.: Federated search in the wild: the combined power of over a hundred search engines. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 1874–1878. ACM, October 2012
9. Pfalzner, S., Gibbon, P: Many-Body Tree Methods in Physics. Cambridge University Press (2005)
10. Schrenk, M.: Webbots, Spiders, and Screen Scrapers: A Guide to Developing Internet Agents with PHP/CURL. No Starch Press (2012)
11. Stefanidis, A., Crooks, A., Radzikowski, J.: Harvesting ambient geospatial information from social media feeds. GeoJournal **78**(2), 319–338 (2013)
12. Tapscott, D.: Grown Up Digital: How the Net Generation is Changing Your World HC. McGraw-Hill (2008)
13. Westerman, D., Spence, P.R., Van Der Heide, B.: Social media as information source: Recency of updates and credibility of information. Journal of Computer-Mediated Communication **19**(2), 171–183 (2014)
14. Sanchez Martin, A.J., de la Prieta Pintado, F., De Gasperis, G.:Fixing and evaluating texts: mixed text reconstruction method for data fusion environments. In: 2014 17th International Conference on Information Fusion (FUSION), pp. 1–6. IEEE, July 2014