Finding interest groups from Twitter lists

Mohamed Benabdelkrim, Jean Savinien, Céline Robardet

▶ **To cite this version:**

Mohamed Benabdelkrim, Jean Savinien, Céline Robardet. Finding interest groups from Twitter lists. SAC '20: The 35th ACM/SIGAPP Symposium on Applied Computing, Mar 2020, Brno Czech Republic, France. pp.1885-1887, 10.1145/3341105.3374077 . hal-03340772

HAL Id: hal-03340772
https://hal.science/hal-03340772v1

Submitted on 10 Sep 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Finding interest groups from Twitter lists

Mohamed Benabdelkrim
Université de Lyon,
LIRIS, UMR 5205, INSA LYON
69621 Villeurbanne, France
mohamed.benabdelkrim@insa-lyon.fr

Jean Savinien
em**lyon business school**
23 avenue Guy de Collongue
69130 Ecully, France
savinien@em-lyon.com

Céline Robardet
Université de Lyon, LIRIS, UMR 5205,
INSA LYON
69621 Villeurbanne, France
celine.robardet@insa-lyon.fr

ABSTRACT

Twitter lists enable users of the social network to organize people they follow into groups of interest (e.g. politicians or journalists they like, favorite artists or athletes, authoritative figures in a given field, and so on). For the analyst, lists are a means of access to the structure of interactions between Twitter users and can be used to identify main actors of a field of interest. In this work, we introduce a methodology for constructing an edge-attributed multilayer network of Twitter users based on their membership to Twitter lists. We propose and validate a new approach that identifies local communities of users and their common interests from the constructed graph. We provide evidences that our method performs in a better way than global community detection approaches, and faster with as good results as competitive local methods.

KEYWORDS

Local community detection, pattern mining, Twitter list analysis.

ACM Reference Format:
Mohamed Benabdelkrim, Jean Savinien, and Céline Robardet. 2020. Finding interest groups from Twitter lists. In *Proceedings of ACM SAC Conference (SAC'20).* ACM, New York, NY, USA, Article 4, 3 pages. https://doi.org/10.1145/3341105.3374077

INTRODUCTION

Social media make possible large-scale dissemination of user-generated content. The rapid diffusion and amplification of content are their major assets and make them a powerful tool. Different mechanisms enable a piece of information to find its audience: the popularity of its author, its number of views, or the tags or hashtags used to index the content and facilitate its access. On Twitter, these mechanisms are implemented through lists that group users identified by third parties as being concerned by the content broadcast on the lists. A Twitter list contains the data of the list creator (Twitter id, screen name, biography), the list title and description defined by its creator, and its date of creation.

These lists are valuable to infer the interests of Twitter users. As explained by [6], *"Twitter lists are unique in that when we look at a user and the names of the lists that he is in, those list names represent what other Twitter users think of that user."* Taken together, lists can be used to build a network of Twitter users. In fact, the co-membership of two Twitter users in several distinct lists means that several users have independently qualified them as pertaining to a similar field of interest. Moreover, the identification of lists related to the same specific field helps with detecting stakeholders in the field as identified by third parties (lists creators).

Few preliminary studies of Twitter lists analysis have been published [1, 6–9]. In [9], authors propose to label users based on a correlation analysis of Twitter lists, keywords and users. To analyze the content of tweets using the lists, [6] suggests to manually group lists by keyword categories and show the semantic coherence of their clusters. [7] proposes a method to recommend lists to Twitter users in order to help people in this task generally done manually. In [1], Bhattacharya et al. introduce a procedure to efficiently extract topical groups from Twitter. They use information on users and their lists membership to identify experts and seekers on multiple topics. In [2], same authors built up on their previous work and proposed a methodology for inferring interests of millions of Twitter users.

In this work, we start by introducing a procedure for constructing a field-specific network of Twitter users based on their membership to Twitter lists. Then, we present an exhaustive search algorithm for extracting local communities from the constructed network. Finally, we evaluate our method by comparing communities it extracts to communities obtained by other graph clustering methods from the literature.

TWITTER NETWORK CONSTRUCTION

The procedure for constructing a field-specific network of Twitter users $\mathcal{G}_u$ is a growing procedure that starts with the choice of a set of seed Twitter accounts U central to the field of interest. It, then, identifies lists related to U and adds them to $\mathcal{G}_u$ as fully connected layers of users.
Let us denote by $\mathcal{L}$ the set of Twitter lists, $\mathcal{V}$ the set of Twitter users, and $\mathcal{K}$ the set of keywords used in the title and description of lists in $\mathcal{L}$. The construction process follows four steps:

(i) "seeds": select a set $U \subset \mathcal{V}$ of seed users related to the field of interest;
(ii) "first corona": select the set $\mathcal{L}_1(U) \subset \mathcal{L}$ of lists containing one of the seed users;
(iii) "scores": compute the scores $f(L)$ of lists $L \in \mathcal{L}_1(U)$, and the median m of these scores distribution;
(iv) "second corona": select the set $\mathcal{L}_2(U) \subset \mathcal{L}$ of all lists L containing one of the users in a list in $\mathcal{L}_1(U)$, and whose scores satisfies $f(L) \geq m$.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SAC'20, March 30-April 3, 2020, Brno, Czech Republic
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-6866-7/20/03…$15.00
https://doi.org/10.1145/3341105.3374077

The list score in stages (iii)-(iv) measures the users and keywords frequencies relative to $\mathcal{L}_1(S)$:

$$f(L) = \sum_{u \in V_L} \frac{|\{L \in \mathcal{L}_1(U); u \in V_L\}|}{|\mathcal{L}_1(U)|} + \sum_{k \in K_L} \frac{|\{L \in \mathcal{L}_1(U); k \in K_L\}|}{|\mathcal{L}_1(U)|}$$

where V_L and K_L are respectively the set of users and keywords that appear in list L.

In particular in stage (iv), if a user or a keyword does not belong to a list in $\mathcal{L}_1(U)$, its frequency is zero. This allows us to keep in the second corona $\mathcal{L}_2(U)$ only those lists whose users and keywords are strongly related to the first corona $\mathcal{L}_1(U)$, and eventually disregard lists unrelated to the seeds.

LOCAL COMMUNITIES EXTRACTION

Our objective is to extract local communities of structurally and semantically close users from the Twitter multilayer network defined above. We are looking for a subset of layers $S \subset \mathcal{L}$ such that the number of common vertices/users and keywords is large against the number of vertices and keywords in each single layer.

We propose to evaluate the quality of a set S by means of the following measure:

$$Q(S) = \alpha\, Q_{\mathrm{v}}(S) + (1-\alpha)\, Q_{\mathrm{k}}(S)$$

$$\text{where} \quad Q_{\mathrm{v}}(S) = \frac{|\bigcap_{L \in S} V_L|}{\max_{L \in S} |V_L|}, \quad \text{and} \quad Q_{\mathrm{k}}(S) = \frac{|\bigcap_{L \in S} K_L|}{\max_{L \in S} |K_L|}$$

The measure evaluates the interest of a set of lists by the number of members and keywords it shares normalized by the maximum number of users or keywords it may contain. A set of layers is considered to be a local community if its Q value is sufficiently large. The constraint P bound to Q and defined as $P \equiv Q(S) \geq \lambda$ is antimonotone meaning that when its is true for a set S it is true for all its subsets $S' \subseteq S$:

$$\forall X \subseteq Y, Q(Y) \geq \lambda \Rightarrow Q(X) \geq \lambda$$

Let ϕ be the mapping that associates the common nodes and keywords to a set of layers, and ψ its reciprocal map, namely:

- $\phi(S) = \left(\bigcap_{L \in S} V_L, \bigcap_{L \in S} K_L\right)$,
- $\psi(V, K) = \{L \in \mathcal{L} \mid \forall v \in V,\ v \in V_L \text{ and } \forall k \in K,\ k \in K_L\}$.

Maps $\sigma(S) = \psi(\phi(S))$ and $\sigma'(V, K) = \phi(\psi(V, K))$ denote closure operators.

We enumerate closed combinations of lists recursively in a depth-first search manner. Given a set of layers S that is currently explored and an index i of the next layer to consider, ExhaustiveSearch algorithm returns all its super-sets that are of high quality (given a parameter λ) and closed: $\{X \mid S \subset X \subseteq \mathcal{L},\ Q(X) \geq \lambda \text{ and } \sigma(X) = X\}$. For its first call, S is the empty-set of layers and $i = 1$.

The algorithm takes advantage of the anti-monotonicity of Q. This property is used in line 4 to stop the search process when no more community is expected in an unpromising enumeration branch. To avoid generating a pattern several times, we use an arbitrary order $\ll$ on $\mathcal{L}$ such that $L_i \ll L_j$ iff $i \leq j$. We generalize this order relation on sets: $\forall X, Y \subseteq \mathcal{L}, X \ll Y \Leftrightarrow \forall x \in X$ and $\forall y \in Y \setminus X,\ x \ll y$. We also take advantage of the closure operator to accelerate the enumeration process by skipping index i corresponding to layers already added by closure: the function nextIndex returns the index of the first layer in $\ll$ order that is greater than all the layers of S and that does not belong to S'.

Algorithm 1: ExhaustiveSearch

Input: S, i, R
Output: $R = \{X \mid S \subseteq X,\ \sigma(S) = S \text{ and } Q(S) \geq \lambda\}$

```
1 if (i = |L| + 1) and (Q(S) ≥ λ) then
2 |   R ← R ∪ {S}
3 else
4 |   if Q(S ∪ L_i) ≥ λ then
5 |   |   S' ← σ(S ∪ L_i)
6 |   |   if S ≪ S' then
7 |   |   |   j ← nextIndex(S, S')
8 |   |   |   ExhaustiveSearch(S', j, R)
9 |   |   ExhaustiveSearch(S, i + 1, R)
```

The optimal value for λ is such that the generated communities cover a maximal number of users with minimal redundancy. The cover of users by patterns, denoted $co(\lambda)$, is calculated as the ratio of users appearing in the set of communities C over the total number of users in $\mathcal{G}_U$. The redundancy, $q_3(\lambda)$ and λ, is estimated by taking the third quartile of the distribution of the Jaccard similarity calculated on pairs of communities of C. We assign scores balancing the cover and the redundancy to values of λ as: $sc(\lambda) = \log \frac{co(\lambda)}{q_3(\lambda)}$. The optimal value $\lambda^\star$ is the one for which sc is maximal.

EVALUATION

We experimentally evaluate our method using field example of the 'French political scene'. We apply the network construction procedure using three seed users U who cover the spectrum of French political parties. The resulting multilayer network is made of 38,306 vertices, 7.845,652 edges over 2.836 layers.

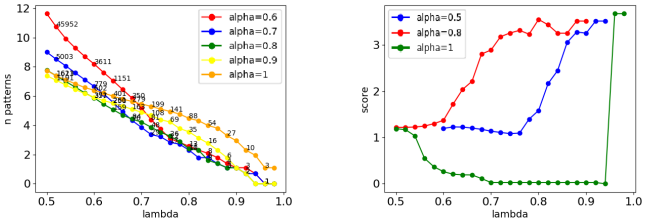

Figure 1: Logarithm of number of distinct local communities of French politics returned by exhaustive enumeration (left), and scores $sc(\lambda)$ (right), as a function of λ with different values for α (see legend).

From figure 1, we can observe that (1) the number of discovered patterns drops exponentially with λ, (2) for small values of α and λ, there is an explosion of the number of patterns due to the large combinatorics between vertices and keywords, (3) for $\alpha = 0.8$, the optimal value for λ is $\lambda^\star = 0.8$.

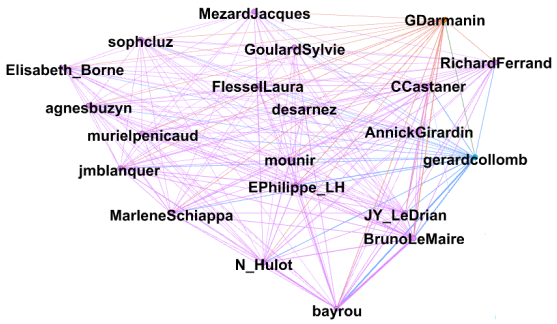

Figure 2: Example of extracted community with $Q = 0.827$ (ExhaustiveSearch method).

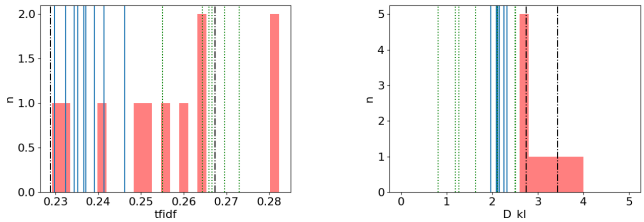

Figure 3: Distribution of mean TF-IDFs (left) and Kullback-Leibler divergence (right) of most occurring words in the biographies of users gathered in local communities extracted by ExhaustiveSearch. Vertical lines display the same quantity for Louvain (green), ML-LCD (black) and LPV (blue).

In Figure 2, we represent an example of local communities output by ExhaustiveSearch with $\alpha = 0.8$ and $\lambda = \lambda^\star = 0.8$. The community contains 21 users from three lists and it has a score $Q = 0.827$. It exclusively contains members and collaborators of the actual political party of the governmental majority in France. This shows the ability of our method to extract more focused and specific clusters compared to Louvain partitioning. This is also the case for communities constructed locally with ML-LCD [5] which are also targeted and of small size. In fact, for this example of French politics, communities of ML-LCD are also returned by ExhaustiveSearch but in a much faster way (around 1,000 times faster).

We propose to assess the quality of local communities extracted with our method with semantic means using the biographies of users associated with the vertices of the local network. Notice that biographies of users provide us with an independent source of data which is neither used for the local network construction nor in any of the extraction methods.

Given an extracted local community c, let D_c denote the document made of the biographies of users in $c$and P_c the target probability distribution of words of D_c. Let $\bar{D}_c$ denote the document made of the biographies of users of $\mathcal{G}_U$ which are not in c and P te probability distribution of words of the whole corpus $D_c \cup \bar{D}_c$.

We compute the mean TF-IDF of words in D_c over $D_c \cup \bar{D}_c$ and the Kullback-Leibler divergence $K(P_c \| P)$. The higher the TF-IDF and the divergence, the more exceptional the extracted community c is as far as semantics are concerned.

In Figure 3, we compare patterns extracted by ExhaustiveSearch with $\alpha = 0.8$ and $\lambda = 0.8$ (pink histogram) to communities returned by other methods:

- Louvain: the purely topological community detection algorithm [3] applied on the graph obtained by aggregating all the layers (green dotted lines);
- ML-LCD: the recent method proposed in [5] for local community detection in multi-layer networks (black dashdot lines);
- LPV: the method by Greene et al. [4] for topical Twitter communities extraction (blue solid lines).

CONCLUSION

In this work, we introduced an exhaustive enumeration method that makes it possible to extract targeted and overlapping communities from a multi-layer local network. Communities obtained by our approach are semantically homogeneous as shown by their TF-IDF and Kullback-Leibler divergence scores when compared to those of clusters computed with Louvain or LPV algorithms, and are of similar quality to the communities identified by ML-LCD while being computed much more quickly.. They are also more fine-grained as a single Louvain cluster generally contains hundreds of different local communities extracted by our approach.

As Twitter lists have a date of creation, in future research we will study the temporal evolution of the field-specific Twitter networks to analyze changes that occur in their local communities.

REFERENCES

[1] P. Bhattacharya, S. Ghosh, J. Kulshrestha, M. Mondal, M. B. Zafar, N. Ganguly, and K. P. Gummadi. 2014. Deep Twitter Diving: Exploring Topical Groups in Microblogs at Scale. In *CSCW '14*. ACM, 197–210.

[2] P. Bhattacharya, M. B. Zafar, N. Ganguly, S. Ghosh, and K. P. Gummadi. 2014. Inferring User Interests in the Twitter Social Network. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*. ACM, New York, NY, USA, 357–360. https://doi.org/10.1145/2645710.2645765

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (2008), P10008.

[4] D. Greene, D. O'Callaghan, and P. Cunningham. 2012. Identifying Topical Twitter Communities via User List Aggregation. In *Proc. 2nd International Workshop on Mining Communities and People Recommenders (COMMPER 2012)*. 41–48.

[5] R. Interdonato, A. Tagarelli, D. Ienco, A. Sallaberry, and P. Poncelet. 2017. Local community detection in multilayer networks. *Data Min. Knowl. Discov.* 31, 5 (2017), 1444–1479.

[6] D. Kim, Y. Jo, I. chul Moon, and A. Oh. 2010. Analysis of Twitter Lists as a Potential Source for Discovering Latent Characteristics of Users. In *CHI 2010 Workshop on Microblogging: What and How Can We Learn From It.*

[7] V. Rakesh, D. Singh, B. Vinzamuri, and C. K. Reddy. 2014. Personalized Recommendation of Twitter Lists using Content and Network Information. In *Conference on Weblogs and Social Media, ICWSM.*

[8] S. Velichety and S. Ram. 2013. Examining Lists on Twitter to Uncover Relationships Between Following, Membership and Subscription. In *World Wide Web – WWW*. ACM, 673–676.

[9] Y. Yamaguchi, T. Amagasa, and H. Kitagawa. 2011. Tag-based User Topic Discovery Using Twitter Lists. In *ASONAM*. IEEE Computer Society, 13–20.