An Interaction Profile-based Classification for Twitter Users

Jonathan Debure, Stephan Brunesseaux, Camelia Constantin, Cédric Du Mouza

An Interaction Profile-based Classification for Twitter Users

Jonathan Debure
AIRBUS & CNAM
Paris, France
jonathan.debure@airbus.com

Stephan Brunessaux
AIRBUS
Paris, France
stephan.brunessaux@airbus.com

Camelia Constantin
Sorbonne University
Paris, France
camelia.constantin@lip6.fr

Cédric Du Mouza
CNAM
Paris, France
dumouza@cnam.fr

***Abstract*—Social networks have become a primary communication tool and are used by hundreds of millions of users daily. They bring together a wide variety of people, individuals, companies, public figures, media, influencers, etc. Users have different behaviours on social networks, such as different publication frequencies, number of followers or different user interactions. In the Twitter social network, for instance, users do not reply, quote or use mentions in the same way. Our intuition is that these interactions may characterise different user types and we consequently present in this work a non-supervised classification method based on interaction scores. We propose and experimentally compare different score estimations, leading our experiments to confirm the relevance of our approach.**

***Index Terms*—Social Network; Clustering; Behaviours; PageRank.**

I. Introduction

Nowadays, Online Social Networks (OSN) are omnipresent. There are different kinds of OSN, providing different services. Among the most famous ones, we can mention Twitter, which allows users to share short messages, media (videos and photos) and private messaging, Facebook, which proposes to share with friends photos, videos, messages and even to sell items or services, LinkedIn, which targets professional users and proposes a recruitment service and YouTube that hosts entertainement videos that users can stream. For several years, these social networks have been analyzed in different contexts. For instance, the content of some messages is analyzed to deduce users sentiments regarding a company or a product for marketing purposes. Recommendation systems use sociological studies in attempt to understand users behaviour for advertising purposes or recommendations of friends connexions. Other applications analyze content and/or user connections to detect inappropriate content or criminal activity.

Community detection in social network analysis is also gaining increasing attention as shown by the current tremendous amount of researches in this area. Existing approaches generally rely on the underlying social network graphs and attempt to group highly connected or frequently communicating users. Our goal here is quite different since we group people according to their type of profile: individuals, media, influencers, etc. The underlying assumption of this classification is that users react differently to messages contents, depending on their profile. We make use of interaction analysis to classify user accounts and to automate users classification.

The rest of the paper is structured as follows. In the Section II, we present a state of the art on online social networks researches. In the Section III, we present our datasets. Then, in Section IV, we explain our data model. In the Section IV, we present an analysis of our model on a global relational graph from our dataset and the obtained results. Then, in the Section VI, we present the same analysis, but this time on specific relational graphs and the obtained results. Finally, in the Section VII, we summarise the different analysis and future work.

II. Related work

Twitter users classification is mostly based on messages content: some studies use linguistic content [11] to classify users by their political orientation [7] or ethnicity. [3] proposes six different approaches to classify tweets content based on different symbols, keywords, categories or interacting messages into different groups, such as “Information”, “Conversation”, “Broadcast”, or “Other”. Content analysis is also employed by [2] which identifies and classifies users in three categories: “Bot”, “Human” or “Cyborg” based on message structures of entities such as URL, images, mentions, etc. [13] and [1] identify users sub-graphs (*i.e.*, community detection) by using a PageRank-based clustering that spreads computation scores through a random walk computation on the graph structure. Network structure-based users clustering and community detection are proposed by [10] and by [8]. The detected communities mainly reflect users’ connectivity and messages spreading across the network.

Several approaches have been proposed to perform community detection in social graphs, based on the follower/followee graph. Users exchange information in a privileged way inside the detected communities. Some existing methods determine a measure of users authority inside a social network based on node degrees [9]. Other approaches are based on the betweenness centrality measure proposed by [4]. They compute node authority depending on the distance between nodes, therefore highlighting users who are in the middle of the network. There are also approaches which consider recommendation scores provided by a PageRank-like algorithm that considers incoming links of nodes and that takes into consideration user centrality. A node with an high score of PageRank is a popular user with a high probability to propagate messages.

 ISBN: 978-1-61208-857-0

III. Presentation of our dataset

To build our dataset, we use the Twitter API Stream that allows us to collect 1% of all tweets published on the platform. We collected tweets during a 5-month period of observation. We filter them to obtain two datasets: the first dataset gathers tweets about COVID, and the second dataset is composed of all tweets about NBA (National Basketball Association). Our final datasets consist of around 24 millions tweets.

NBA Dataset

The NBA dataset consists of 5M tweets produced by 2M unique users. From this 5M tweets, we identified 4.9M interactions (`Retweet`, `Quote`, `Reply` and `Mention`). It is important to note that not all tweets correspond to interactions while, at the same time various tweets may contain several interactions (such as retweets and/or quotes). To build the interaction graph used in our experiments, we only kept users that performed at least two interactions. Then, we computed the largest connected component (we used the `NetworkX` Python library [5]). This pre-processing step avoids to get a small sub-graph with isolated nodes that can reduce the global PageRank score of graph nodes.

The main characteristics of the NBA dataset are presented in Table I.

COVID Dataset

The COVID dataset consists of 21M tweets which allow to build an interaction graph of 6 million unique users and 17 million interactions. The extraction of the largest connected component during the pre-processing step produces a graph with 2,789,316 users.

The main characteristics of the COVID dataset are presented in Table II.

IV. The data model

We introduce in this section our notations and our data model. We consider the Twitter platform and its underlying directed graph of interactions $\mathcal{G} = (\mathcal{U}, \mathcal{E})$ where $\mathcal{U}$ denotes the set of nodes, *i.e.* users, $\mathcal{E} \subseteq \mathcal{U} \times \mathcal{U}$ is the set of edges, such that $(u_1, u_2) \in \mathcal{E}$ means that user u_2 performed an action on the tweets of user u_1. We denote $\mathcal{A}$ the set of possible interactions that a user can execute on another user tweet. In the following, we consider that $\mathcal{A} = \{a_{rt}, a_{qt}, a_{rp}, a_{mt}\}$, which corresponds respectively to the actions of `Retweet`, `Quote`, `Reply` and `Mention`.

The restriction of the interaction graph $\mathcal{G}$ to a given action $a \in \mathcal{A}$ denoted $\mathcal{G}_a$ is the graph $\mathcal{G}_a = (\mathcal{U}_a, \mathcal{E}_a)$ with $\mathcal{U}_a \subseteq \mathcal{U}$ and $\mathcal{E}_a \subseteq \mathcal{E}$ such as $(u_1, u_2) \in \mathcal{E}_a$ if u_2 performed an interaction of type a on a tweet of u_1.

Obviously, all edges from an interaction graph do not represent the same level of interaction between users. Some interactions may happen frequently, while others may happen rarely. To capture this notion, we define an interaction weight ω as follows:

Definition 1 (Global interaction weight): The global interaction weight ω is a function $\omega : \mathcal{E} \to \mathbb{R}$ that takes into account all interactions between user couples.

Definition 2 (Specific interaction weight): The specific interaction weight ω for an action $a \in \mathcal{A}$ is a function $\omega : \mathcal{E} \times \mathcal{A} \to \mathbb{R}$. This score is mainly based on the interactions of a type $a \in \mathcal{A}$ and gives less (or no) importance to other types of interaction.

Finally, we assume the existence of a function $count : \mathcal{E} \times \mathcal{A} \to \mathbb{N}$, such as $count((u_1, u_2), a)$ is the number of interactions of type a that u_2 performed on tweets of u_1.

V. Global interaction score-based clustering

This approach is a little different from our original goal, which aimed to identify users with the same "profile" (role) within the social network using different interactions. Our intuition is, that clustering based on diff Better interactions between users provide more relevant clusters.

A. Global interaction occurrences-based clustering

Definition 3 (Occurrences-based global interaction score): The occurrences-based global interaction score σ_u^g for a user u is defined as:

$$\begin{gathered} \forall v \in \mathcal{U}, \omega(v,u) = \Sigma_{a \in \mathcal{A}}\ count((v,u),a) \\ \sigma_u^g = \log\left(\frac{\Sigma_{v \in \mathcal{U}}\ \omega(v,u)}{max_{w \in \mathcal{U}}(\Sigma_{v \in \mathcal{U}}\ \omega(w,v))}\right) \end{gathered} \tag{1}$$

Note the normalization of the score and the usage of the log function to smooth the differences between accounts.

According to this interaction score, since data is not tagged, we decided to use a non-supervised clustering. More precisely, we chose the K-Means clustering algorithm for its scalability and because it is known to give good clustering results. To determine the number of K-clusters, we rely on the Silhouette Score [12].

For the occurrences-based global interaction score approach, we observe that the Silhouette score is increasing with the number of clusters (see Fig. 1). It illustrates that no clusters number appears to be better than another (except maybe clusters with a single user). Moreover, the manual analysis of a clustering, for example with $K = 4$ or $K = 5$ reveals that the clusters obtained contain very heterogeneous classes of users.

B. Global interaction PageRank-based clustering

It has been shown that PageRank can accurately compute influence ranks since it is not influenced by the number of followers but by the user interactions [6]. Consequently, we expect that a PageRank-based global interaction score will provide a better user classification. The PageRank score for a user $u_i \in \mathcal{U}$ is estimated by the following formula:

$$PR(u_i) = (1-\alpha) + \alpha \sum_{u_j \in In(u_i)} \frac{PR(u_j)}{Out(u_j)} \tag{2}$$

 ISBN: 978-1-61208-857-0

TABLE I
NBA DATASET: STATISTICS

	Followers	Friends	# Tweets	# Quotes	# Retweets	# Mentions	# Replies
Value Count	882494	882494	1935124	561041	472376	644758	211985
Mean	4000.47	1096.57	2.49	1.25	1.93	2.27	0.45
Median	328	445	1	1	1	1	0
Std Dev	200434.26	4632.68	10.78	3.21	6.69	7.66	2.32
Min	0	0	1	0	0	0	0
Max	87244738	1480293	9171	1394	1478	2002	748

TABLE II
COVID DATASET: STATISTICS

	Followers	Friends	# Tweets	# Quotes	# Retweets	# Mentions	# Replies
Value Count	2789316	2789316	6278280	1783237	1699905	1945609	588131
Mean	3832.08	1128.18	2.64	1.31	2.49	2.01	0.37
Median	287	435	8348	1	1	1	0
Std Dev	128751.36	4644.14	8.27	3.09	8.76	6.74	2.37
Min	0	0	1	0	0	0	0
Max	85941911	1907480	8768	929	7910	2823	1480

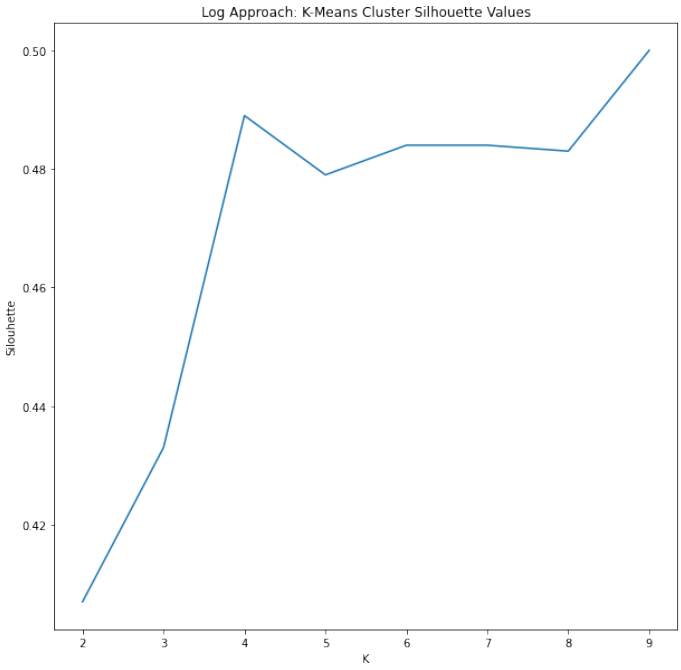

Fig. 1. Log Basic Approach: K-Means silhouette

where $In(u_i)$ denotes the set of users that have an interaction with u_i (*i.e.*, $\{u_j \in \mathcal{U}, (u_i, u_j) \in \mathcal{E}\}$), $Out(u_j)$ the out-degree of user u_j, α is a dumping factor.

We take into consideration multiple occurrences of the same interaction between two users by supposing that they illustrate a strong interaction between those users. This is modeled by the edges weights of the interaction graph $\mathcal{G}$. The weight of an edge between two users is the total number of interactions between them. In order to compute a PageRank score using edges weight, we use the Weighted PageRank (WPR) Algorithm [14]. WPR assigns higher scores to more important nodes instead of dividing the score between their neighbours. Nodes will get a value proportional to their number of in-interactions (interactions a user had with his tweets) and out-interactions (interactions a user had with tweets of other users). Consequently, we adopt the following definition:

Definition 4 (Interaction weights):
The in-interactions weight $\mathcal{W}^{in}_{(i,j)}$ and out-interactions weight $\mathcal{W}^{out}_{(i,j)}$ for an edge $(u_i, u_j) \in \mathcal{E}$ are estimated as:

$$\mathcal{W}^{in}_{(i,j)} = \frac{\sum\limits_{a \in \mathcal{A}} count((u_j, u_i), a)}{\sum\limits_{v \in In(u_i)} \sum\limits_{a \in \mathcal{A}} count((v, u_i), a)}$$

$$\mathcal{W}^{out}_{(i,j)} = \frac{\sum\limits_{a \in \mathcal{A}} count((u_i, u_j), a)}{\sum\limits_{v \in Out(u_i)} \sum\limits_{a \in \mathcal{A}} count((u_i, v), a)} \tag{3}$$

Finally, we adapt the Weighted PageRank proposed in [14] to take into consideration interaction weights on edges.

Definition 5 (Weighted interaction PageRank score):
Using the previous PageRank formula and the interaction weights defined above, we estimate:

$$WPR(u_i) = (1 - \alpha) + \alpha \sum_{p_j \in In(u_i)} WPR(u_j) \times \mathcal{W}^{in}_{(j,i)} \times \mathcal{W}^{out}_{j,i)} \tag{4}$$

VI. INTERACTION PROFILES-BASED CLUSTERING

By considering all interactions as similar, we are masking differences in the user behaviours. Indeed, the analysis of a

 ISBN: 978-1-61208-857-0

few accounts seems to reveal that some users appear to favour certain interactions over others and this could be a relevant classification criterion. To verify this intuition, we built an interaction profile for the users that is defined as follows:

Definition 6 (Interaction profile): The interaction profile of a user u is a quadruplet $\sigma_u^p(\sigma_{rt}(u), \sigma_{qt}(u), \sigma_{rp}(u), \sigma_{mt}(u))$ where each dimension σ_a is the specific interaction score determined on the graph restriction $\mathcal{G}_a$.

As for the global approach, we compare the straightforward approach with specific interaction scores estimated with the number of interactions of the corresponding action, and the PageRank one.

A. Occurrences-based interaction profiles

For this approach, we consider that a specific interaction weight for an interaction a is estimated on the restricted graph $\mathcal{G}_a$ as:

$$\forall (u, v, a) \in \mathcal{U}^2 \times \mathcal{A}, \omega(u, v, a) = count((u, v), a) \quad (5)$$

Consequently, our interaction profile scores are estimated as follows:

Definition 7 (Occurrences-based interaction profiles scores): The scores for the occurrences-based interaction approach σ_u^p for a user u is defined as:

$$\forall a \in \mathcal{A}, \sigma_a(u) = \log\left(\frac{\Sigma_{v \in \mathcal{U}} \omega((v, u), a)}{max_{w \in \mathcal{U}}(\Sigma_{v \in \mathcal{U}} \omega((w, v), a)}\right) \quad (6)$$

Once these scores are computed, we perform the K-Means non-supervised clustering. To evaluate the clustering quality, we have performed a human validation which consists of manually analyzing a sample of 50 accounts randomly chosen inside each cluster.

We observe that, with the occurrences-based interaction profile approach, our clusters remain heterogeneous, as well as with the occurrences-based global approach: all kinds of users are present in each cluster. This phenomenon can be explained by the fact that this method only uses in-degree values. However, we aim at classifying users based on interactions on their messages. It has been shown that messages shared by popular or central users of the graph can be spread efficiently [15].

B. PageRank-based interaction profiles

Instead of a straightforward interaction scores computation, based on the number of occurrences, we estimate them using a PageRank approach. Therefore, we considered the graph restriction $\mathcal{G}_a$ of each interaction and performed the weighted PageRank algorithm to compute the associated dimension of the interaction profile. Our intuition is that capturing the "influence" of a user on a given interaction (*i.e.* his capacity to generate a given interaction on the network) better characterizes a user behaviour.

Definition 8 (PageRank-based interaction profiles scores): The scores for the PageRank-based interaction approach σ_u^p for a user u is defined as:

$$\forall a \in \mathcal{A}, \sigma_a(u) = WPR_{\mathcal{G}_a}(u) \quad (7)$$

TABLE III
WEIGHTED PAGERANK: CLUSTERS SUMMARY

	Weighted PageRank Value
Cluster 1	Reply PageRank is in average **9.10% smaller**, Retweets PageRank is in average **26.56% smaller**, Quote PageRank is in average **29.39% smaller**.
Cluster 2	Reply PageRank is in average **70.84% greater**, Mentions PageRank is in average **20.30% smaller**, Quote PageRank is in average **13.99% smaller**.
Cluster 3	Retweet PageRank is in average **520% greater**, Quote PageRank is in average **230% greater**, Mention PageRank is in average **204% greater**.
Cluster 4	Reply PageRank is in average **182% greater**, Mention PageRank is in average **181% greater**, Retweet PageRank is in average **84.71% greater**.
Cluster Outliers	Reply PageRank is in average **493% greater**, Retweet PageRank is in average **3155% greater**, Quote PageRank is in average **3857% greater**.

where $WPR_{\mathcal{G}_a}(u)$ is the Weighted PageRank score computed for u on the graph $\mathcal{G}_a$, and $\mathcal{G}_a$ is the reduction of the graph $\mathcal{G}$ for the interaction a.

Once these scores are computed, we also perform the K-Means non-supervised clustering. The clustering produces clusters with very different characteristics (see Table III). Globally, we see that `Reply` actions are what is mostly done by real individual users. On the contrary, entities (companies, media, etc) generate more `Retweet` actions. `Mention` can be generated by both human and entities. As there is also a correlation between popularity and `Retweet` actions, we can consider `Quote` as a kind of `Retweet`. Using weighted PageRank on the different interaction graphs to estimate the interaction scores allows to demonstrate the importance of the nodes that interact with the user. This is why we obtain homogeneous clusters with a large majority of similar users.

As for the occurrences-based interaction profile approach, we evaluated clustering quality with a human validation by manually analyzing (i.e Read users timelines, descriptions, photos) a sample of 50 accounts randomly chosen inside each cluster. Since the clusters are more homogeneous, it was possible to qualify the different classes of users we identified. Table IV presents the results of our analysis where Types are defined from our manual analysis of each account.

Finally, we perform a last experiment to validate our clustering. We consider 100 new users we manually "tagged" with a cluster id, according to the cluster composition we observed with our initial dataset. Then, we use our clustering algorithm to allocate them in a cluster. For these new users, we obtain that 96% of them were tagged with the good cluster id, which means that the clusters we obtained correspond to well-identified classes of users.

 ISBN: 978-1-61208-857-0

TABLE IV
WEIGHTED PAGERANK: CLUSTERS COMPOSITION

	Size	Composition	Types
Cluster 1	92.63%	100% composed from common users	Common users
Cluster 2	5.44%	55% composed from common users and 45% popular users (more than 4000 followers)	Moderately popular users, local celebrities, doctors, media specialists and active community users
Cluster 3	0.59%	55% composed from entities and 45% human users but mainly above 10 000 followers	Entities, professional users, brands, hospital, city and feed/news accounts
Cluster 4	0.66%	60% composed from popular user more than 4000 followers and 35% users with more than 10 000 followers	Influencers, writers, journalist, attorneys
Cluster Outliers	0.68%	60% human users, 40% entities. With 45% users with more than 100 000 followers and 40% with more than 10 000 followers	Celebrities, international news, politicians and brands

VII. CONCLUSION

This article presents a method to cluster Twitter users based on the interactions on their tweets. Based on interaction graphs and Weighted PageRank computation, we determine the user interaction profiles. Then, we perform a K-means non-supervised clustering which groups users with similar interaction profiles. Our experiments and manual validation confirm that this approach provides relevant clusters. As future work, we intend to study the parameters that influence the differences between the number of followers within the same cluster. We will also consider the graph dynamicity to propose an adaptive cluster re-computation on a sliding window.

REFERENCES

[1] R. Andersen, F. Chung, and K. Lang. Local partitioning for directed graphs using pagerank. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 166–178. Springer, 2007.

[2] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter: human, bot, or cyborg? In *Proceedings of the 26th annual computer security applications conference*, pages 21–30, 2010.

[3] S. Dann. Twitter content classification. *First Monday*, 2010.

[4] L. C. Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[5] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using networkx. In G. Varoquaux, T. Vaught, and J. Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

[6] B. Hajian and T. White. Modelling influence in a social network: Metrics and evaluation. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 497–500. IEEE, 2011.

[7] O. Hanteer, L. Rossi, D. V. D'Aurelio, and M. Magnani. From interaction to participation: The role of the imagined audience in social media community detection and an application to political communication on twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 531–534, 2018.

[8] B. S. Khan and M. A. Niazi. Network community detection: A review and visual survey. *arXiv preprint arXiv:1708.00977*, 2017.

[9] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600, 2010.

[10] S. Nandanwar and M. N. Murty. Structural neighborhood based classification of nodes in a network. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1085–1094, 2016.

[11] M. Pennacchiotti and A.-M. Popescu. A machine learning approach to twitter user classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, 2011.

[12] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.

[13] S. A. Tabrizi, A. Shakery, M. Asadpour, M. Abbasi, and M. A. Tavallaie. Personalized pagerank clustering: A graph clustering algorithm based on random walks. *Physica A: Statistical Mechanics and its Applications*, 392(22):5772–5785, 2013.

[14] W. Xing and A. Ghorbani. Weighted pagerank algorithm. In *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.*, pages 305–314. IEEE, 2004.

[15] T. R. Zaman, R. Herbrich, J. Van Gael, and D. Stern. Predicting information spreading in twitter. In *Workshop on computational social science and the wisdom of crowds, nips*, volume 104, pages 17599–601. Citeseer, 2010.

 ISBN: 978-1-61208-857-0