An Interaction Profile-based Classification for Twitter Users

Jonathan Debure [1,2] Cédric Du Mouza [2] Camellia Constantin [3]
Stéphan Brunessaux [1]

[1]AIRBUS, Paris, France

[2]CNAM, Paris, France

[3]Sorbonne University, Paris, France

May 30, 2021

Presenter

Jonathan Debure

- PhD Student in Conservatoire National des Arts et Métiers de Paris (2018-2021)
- Data Scientist in Airbus Defence & Space
- Research interest
  - Social network analysis
  - Machine Learning

Outline

Introduction

- Social network for communication, entertainment and marketing
- Share texts, images, audios, videos
- Importance of influencers for advertisers, medias and security

Related work

- Influence scoring : PageRank, Klout score...
- Community detection
- Presidential election prediction, fake news, sentiment analysis
- Bot detection

The data model

Notations: Graph

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.

The data model

Notations: Graph

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.

Notations: Interaction

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`.

Interaction based graph

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.

Account Ranking

Weighted PageRank

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

Account Ranking

Weighted PageRank

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

Weight

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

$$\mathcal{W}^{out}_{(i,j)} = \frac{\sum_{a \in \mathcal{A}} count((u_i, u_j), a)}{\sum_{v \in Out(u_i)} \sum_{a \in \mathcal{A}} count((u_i, v), a)} \quad (2)$$

Graph statistics

- We collect a dataset of 21 Million Tweets about COVID
- We built a graph of 6 Million nodes and 17 million edges
- We extract the largest connected component : 2.7 Million nodes and 6.2 Million edges

Graph statistics

Table 1: 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

Global interactions clustering

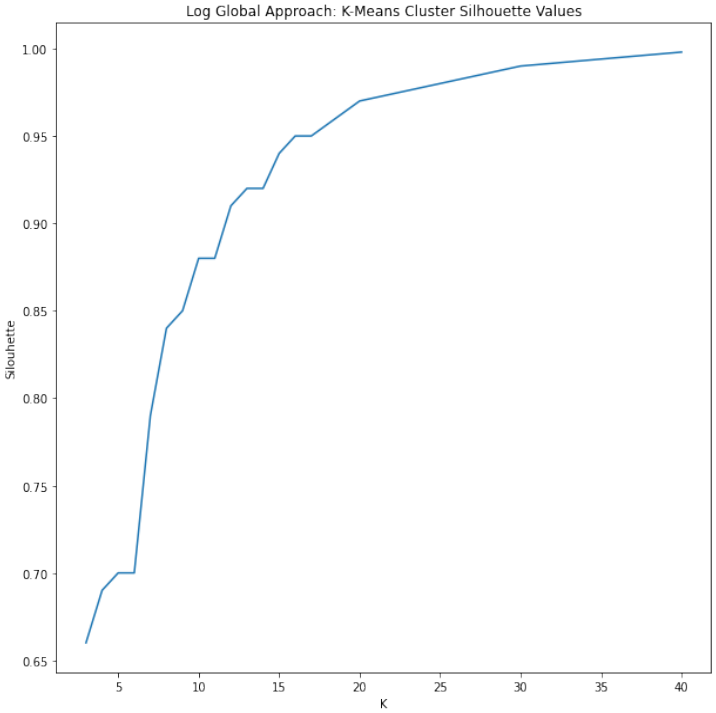
First experiment : using the global interaction graph

The global interactions weight

The global interaction edge weight ω is a function $\omega : \mathcal{E} \rightarrow \mathbb{R}$ that takes into account all interactions between user couples.

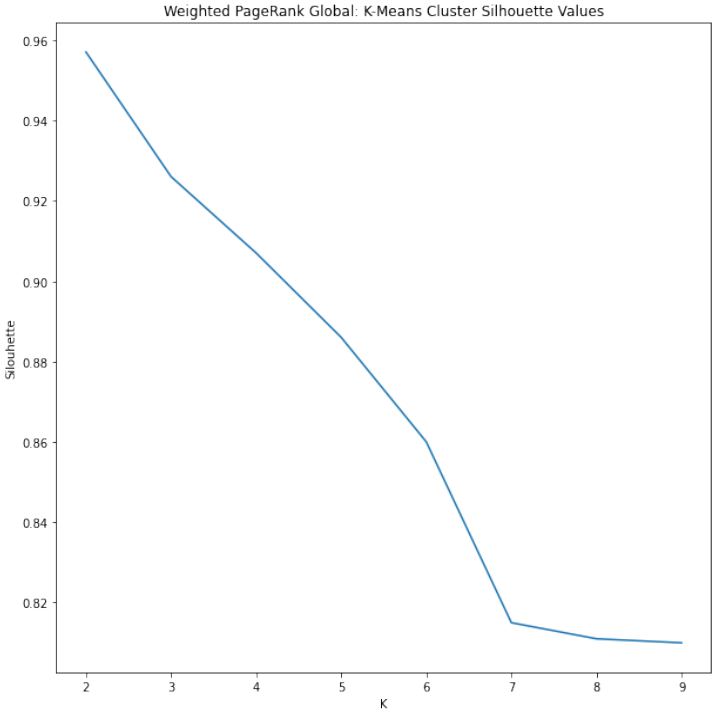
Global interactions clustering

First experiment : with log function



Global interactions clustering

First experiment : with Weighted PageRank



Occurrences-based interaction profiles

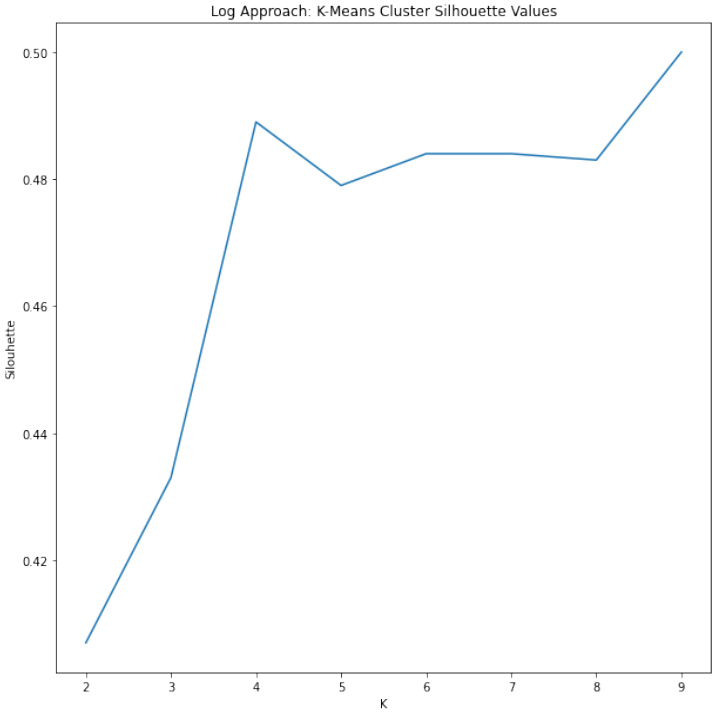
Occurrences-based interaction profiles

We consider that a specific interaction weight for an interaction a is estimated on the restricted graph $\mathcal{G}_a$ as:

$$\forall (u, v, a) \in \mathcal{U}^2 \times \mathcal{A}, \omega(u, v, a) = count((u, v), a) \tag{3}$$

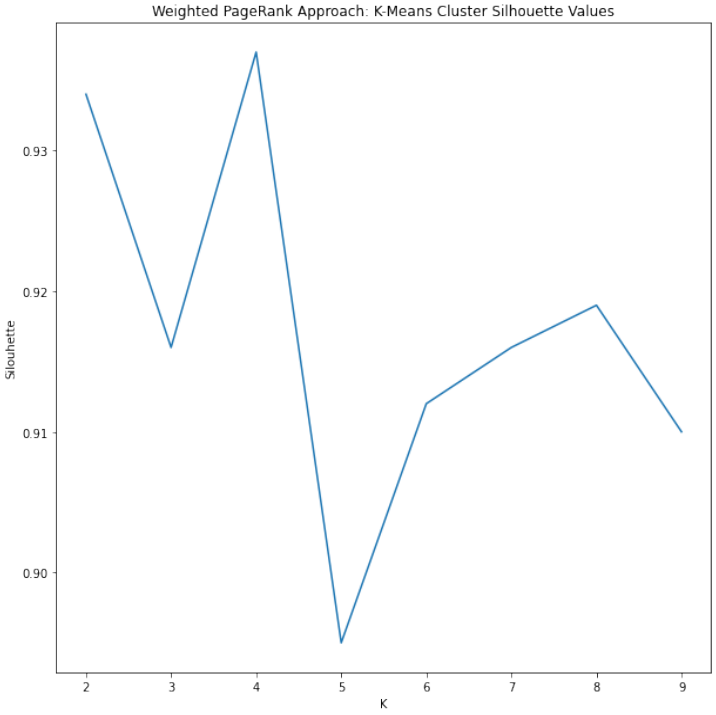
Occurrences-based interaction profiles

Second experiment : with log function



Occurrences-based interaction profiles

Second experiment : with Weighted PageRank



Results

Table 2: 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**.

Results

Table 3: 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

Conclusion

- We present an approach to cluster Twitter users
- Based on Weighted PageRank and users interactions
- We perform a K-means clustering and an manual validation to confirm that approach
- Several perspectives to complete this work :
  - Use mechanical turk to create a labeled and perform supervised clustering
  - Consider the graph dynamicity to propose an adaptive cluster re-computation on a sliding windows

Questions?

Thank you

Contact : jonathan.debure@airbus.com