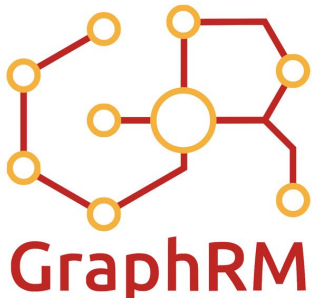
Who are the top influencers and what characterizes them?

Nicola Procopio 19-04-2022

{codemotion}
community



About me

Nicola Procopio

Senior Data Scientist

Contacts

nicola.procopio@healthwareinternational.com

https://it.linkedin.com/in/nicolaprocopio

https://github.com/nickprock

https://www.slideshare.net/NicolaProcopio

Actually @

healthware international

Background

integris

ICAR CNR

Affare

UNIVERSITÀ DELLA CALABRIA

KDD Lab

Community

AI DEMOCRATIZE

tag Talent Garden Cosenza

iaml ITALIAN ASSOCIATION FOR MACHINE LEARNING

GDG

SC STARTUP CALABRIA

Mission

healthware international

The full-service healthcare agency of Healthware Group

We play at the intersection of science, creativity, boundless curiosity, and our understanding of human needs. That's how we design transformational healthcare experiences that engage, simplify and empower people's lives.

We are digital natives and multi-talented coders, connected and passionate to learn and innovate.

Our mission is to design and develop successful solutions and digital products.

“One machine can do the work of fifty ordinary men. No machine can do the work of one extraordinary man.”

-Elbert Hubbard-

Classical Metrics

Some metrics to identify influencers:

- *indegree influence*: # followers
- *retweet(or mention) influence*: # RT (or @)
- follower/following
- *PageRank (or TwitterRank)*
- *HITS*

They consider neither if a user is active on a matter of interest, or her opinion.

Classical Metrics

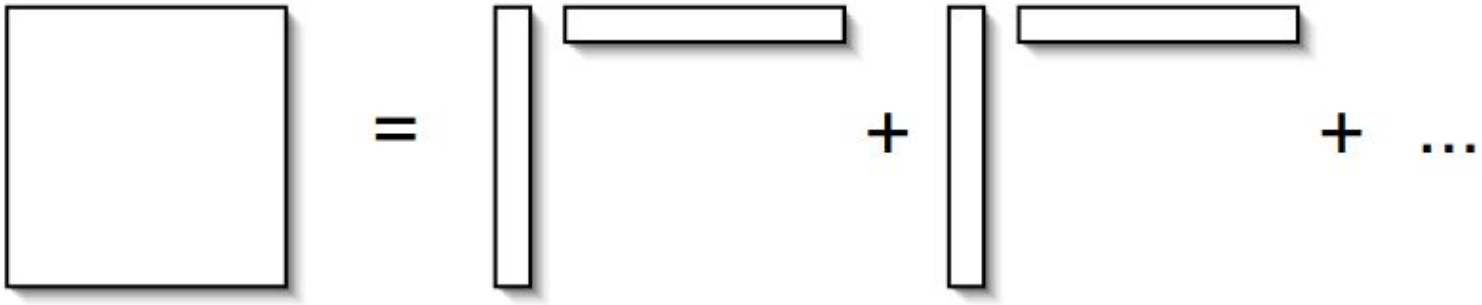
- ***in-degree*** corresponds to the popularity of a user
- ***mentions*** represent the name value of a user and measure the capability of that user to attract other users in a topic discussion
- ***retweets*** express the importance of the user's tweet content and measure the ability of that user to spawn interesting arguments
- ***TwitterRank*** (improves PageRank) is an approach to measure twitterer influence by taking into account the link structure of followers/following of individual users and the topical similarity between these users

HITS

The algorithm HITS (Hypertext Induced Topic Selection), assigns a score to web pages by exploiting the principal singular vectors of the adjacency matrix of the subgraph extracted from the web.

The algorithm introduces the concepts of **hub** and **authority**.

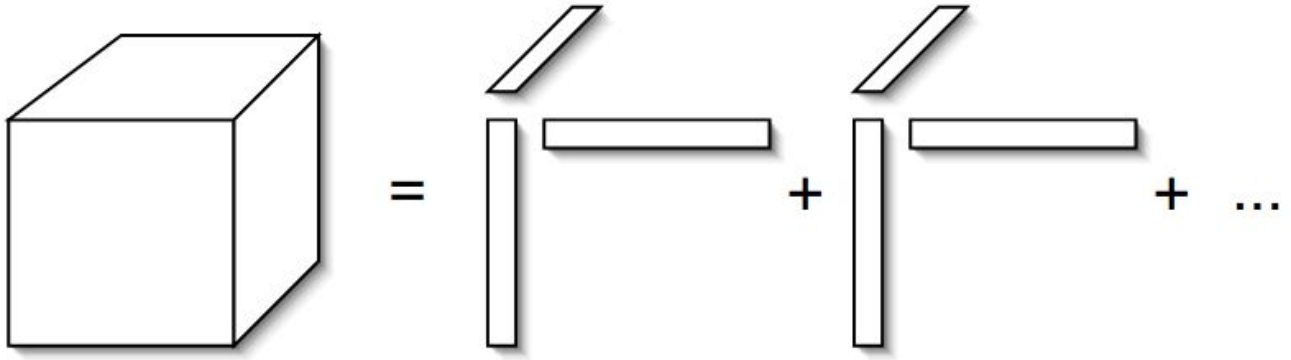
In HITS model, the SVD provides a 2-way decomposition that yields authority and hub scores.



TOPHITS

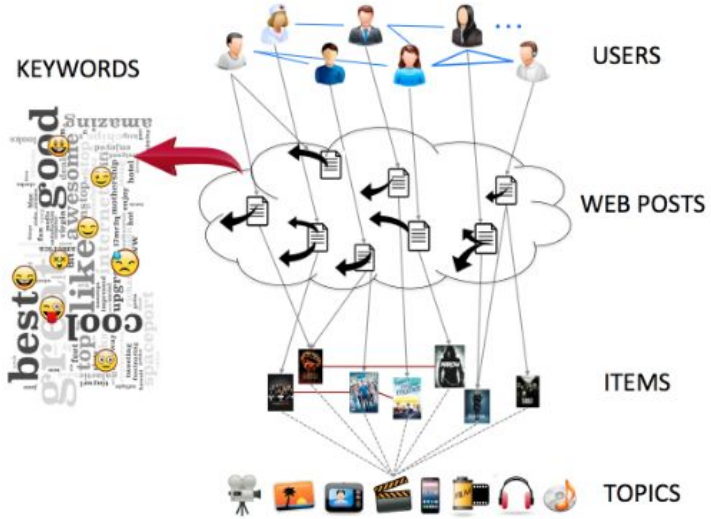
The algorithm TOPHITS extends HITS and produces set of triplets ***{u, v, w}***, where the *u* and *v* vectors contain hub and authority scores for the web pages, and the *w* vector contains topic scores for the terms.

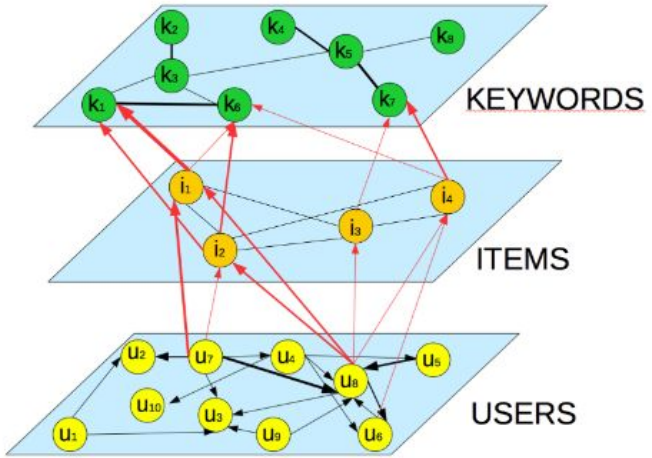
In TOPHITS, the PARAFAC model provides a 3-way decomposition that yields authority, hub, and topic scores.



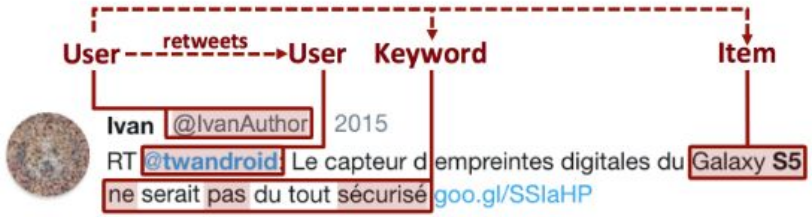
SocialAU

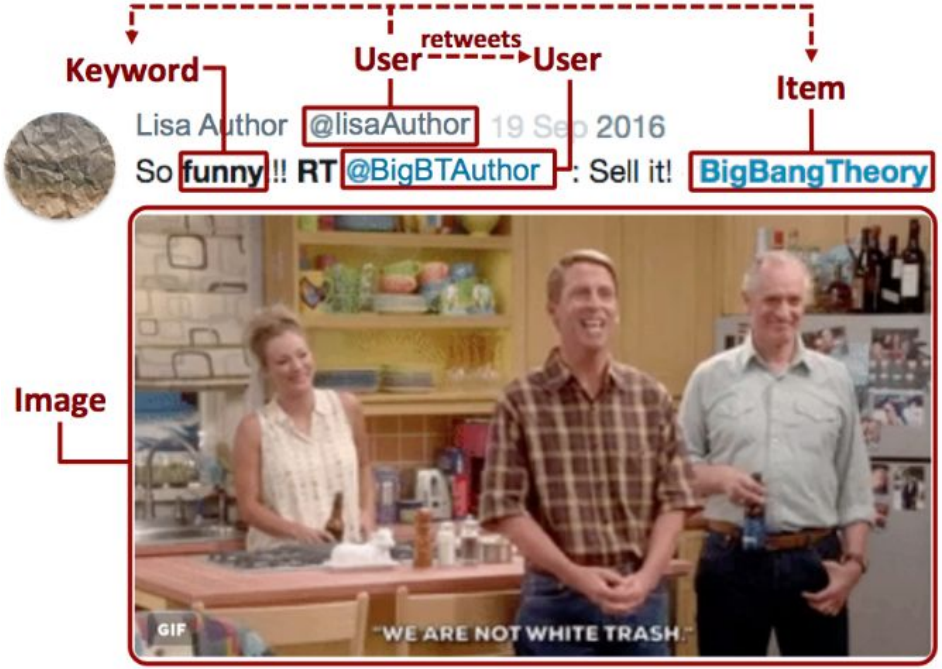
Social media Authoritative User extracts from user textual messages, the items related to the selected topic and the keywords used to express opinions on these items, and models this information with a three-layer network.





SocialAU: Data Structure





We build 4 graphs:

- 3 layer graphs represented by adjacency matrix
- An inter-layer graph built by triplets *(u, i, k)*

SocialAU and TOPHITS

Main similarities:

- SocialAU extends the TOPHITS technique to identify topics and the associated authoritative web pages (users for SocialAU).
- Analogously to TOPHITS, it employs the greedy PARAFAC procedure to obtain authority and hub scores of the three-layer network.

Main differences:

- SocialAU employs a multilayer network while TOPHITS uses a multiplex network, a particular case of multilayer network where the set of nodes is shared by all the layers, and cross-layer connections are only between a node in a layer and the counterpart in another layer.
- SocialAU modifies the PARAFAC greedy algorithm to take into account the scores computed on each layer by the HITS method.

SocialAU: How it works

Input: 2-dimensional adjacency matrices $\mathbf{M}_U$, $\mathbf{M}_I$, $\mathbf{M}_K$: of the graphs $G_U = (X_U, E_U)$, $G_I = (X_I, E_I)$, and $G_K = (X_K, E_K)$ modeling users, items, and keywords, respectively.
3-dimensional adjacency tensor $\mathbf{A}$ modeling the three-layer interconnections.
Output: Rank-1 approximation of $\mathbf{A}$ as triplet $(\mathbf{h}^{(1)}, \mathbf{a}^{(1)}, \mathbf{w}^{(1)})$ defining dominant users $\mathbf{h}^{(1)}$ which are also authoritative in the network G_U, dominant items $\mathbf{a}^{(1)}$ in G_I and dominant keywords $\mathbf{w}^{(1)}$ in G_K.
Method: Perform the following steps:

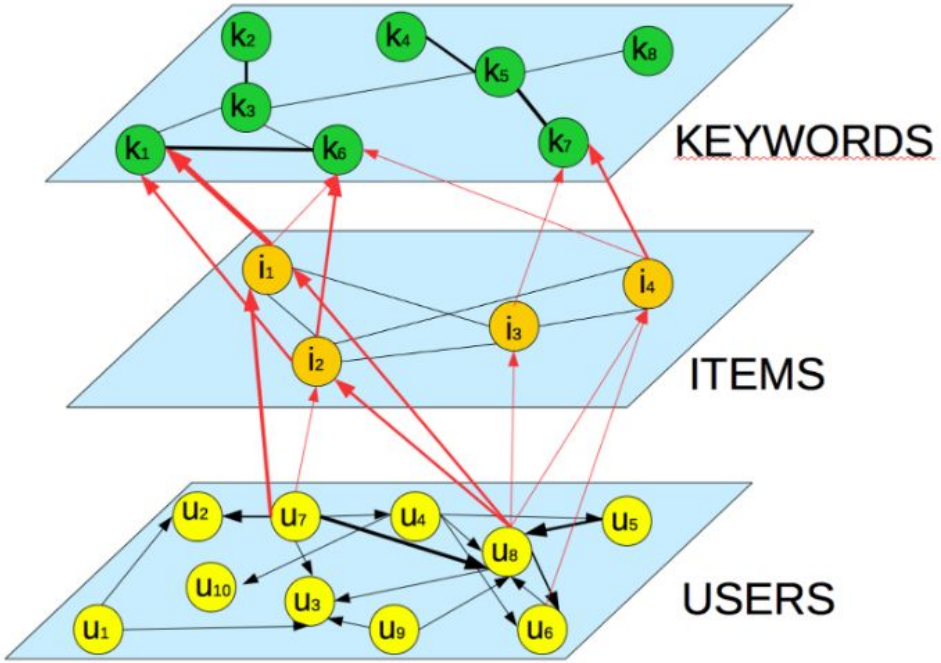
1) set t=1, Initialize $\mathbf{a_U}^t$, $\mathbf{h_U}^t$, $\mathbf{a}^t$ to all ones vectors of size n
   initialize $\mathbf{a_I}^t$, $\mathbf{h_I}^t$, $\mathbf{h}^t$ all ones vectors of size m
   initialize $\mathbf{a_k}^t$, $\mathbf{h_k}^t$, $\mathbf{w}^t$ all ones vectors of size r
2) $\lambda = 0$, set ε to a small value
3) **while** not termination
4) $\mathbf{h_U}^{t+1} = \mathbf{M}_U * \mathbf{a_U}^t$
5) $\mathbf{a_U}^{t+1} = \mathbf{M}_U^T * \mathbf{h_U}^{t+1}$
6) $\mathbf{h_I}^{t+1} = \mathbf{M}_I * \mathbf{a_I}^t$
7) $\mathbf{a_I}^{t+1} = \mathbf{M}_I^T * \mathbf{h_I}^{t+1}$
8) $\mathbf{h_k}^{t+1} = \mathbf{M}_K * \mathbf{a_k}^t$
9) $\mathbf{a_k}^{t+1} = \mathbf{M}_K^T * \mathbf{h_k}^{t+1}$
10) $\mathbf{h}^{(t+1)} = \mathbf{A} \bar{\times}_2 \mathbf{a}^{(t)} \bar{\times}_3 \mathbf{w}^{(t)} + \mathbf{h_U}^{t+1} + \mathbf{a_U}^{t+1}$
11) $\mathbf{a}^{(t+1)} = \mathbf{A} \bar{\times}_1 \mathbf{h}^{(t+1)} \bar{\times}_3 \mathbf{w}^{(t)}$
12) $\mathbf{w}^{(t+1)} = \mathbf{A} \bar{\times}_1 \mathbf{h}^{(t+1)} \bar{\times}_2 \mathbf{a}^{(t)} + \mathbf{a_k}^{t+1}$
13) $\lambda_1 = \|\mathbf{h}\| \; \|\mathbf{a}\| \; \|\mathbf{w}\|$
14) normalize all vectors
15) **if** $\lambda_1 - \lambda \leq \varepsilon$
16) termination=true
17) **else** $\lambda = \lambda_1$
18) **end while**
19) **return** $\mathbf{h}^{(1)} = \mathbf{h}^t$, $\mathbf{a}^{(1)} = \mathbf{a}^t$, $\mathbf{w}^{(1)} = \mathbf{w}^t$, $\sigma^{(1)} = \lambda$

SocialAU is an iterative algorithm, each iteration cost is *O(N)* like TOPHITS.

As **TOPHITS** it calculates the most influential triplets but also takes into account a score of importance in their network using **HITS**.

The original ***SocialAU*** was written in MATLAB 2015b using the Tensor Toolbox.
Between 2020 and 2022 i rewrote the algorithm using **numpy** and **pytorch**, you can find the implementations on github

SocialAU: Toy Example



SocialAU		*TOPHITS*	
user	*score*	*user*	*score*
u_8	0.72635	u_7	0.71482
u_7	0.60013	u_8	0.69931
u_5	0.21736	u_6	0.0010271
u_4	0.14216	u_5	0.0010271
u_9	0.14189	u_1	0
u_2	0.11397	u_2	0
u_{10}	0.090703	u_3	0
u_3	0.042535	u_4	0
u_1	0.031082	u_9	0
u_6	0.026105	u_{10}	0

Evaluation Measures

To qualitatively evaluate the results of SocialAU some influence measures adopted in the literature for Twitter datasets are considered.
Moreover, new indexes are introduced to better understand the activity rate of a user and the capability of generating interesting contents that catch other users' attention.

- Followers/following Ratio
- Retweet Influence Ratio and Mention Influence Ratio
- Retweet and Mention Ratio (Normalized and User Normalized)
- Interaction Ratio (and Normalized)
- Social Network Potential
- User Activity

Examples

TV Series

- **20366 Tweets** from 4th to 14th January 2016
- **12 TV Series**
- **14207 users** with **17410 edges**
- **6123 keywords** with **72856 edges**
- **51534 triples** in a tensor 14207x12x6123

Smartphones

- **24834 Tweets** from 7th May to 27th July 2015
- **51 smartphones**
- **9028 users** with **9191 edges**
- **2706 keywords** with **29554 edges**
- a Tensor 9028x51x2706 with **26673 triples**

YELP

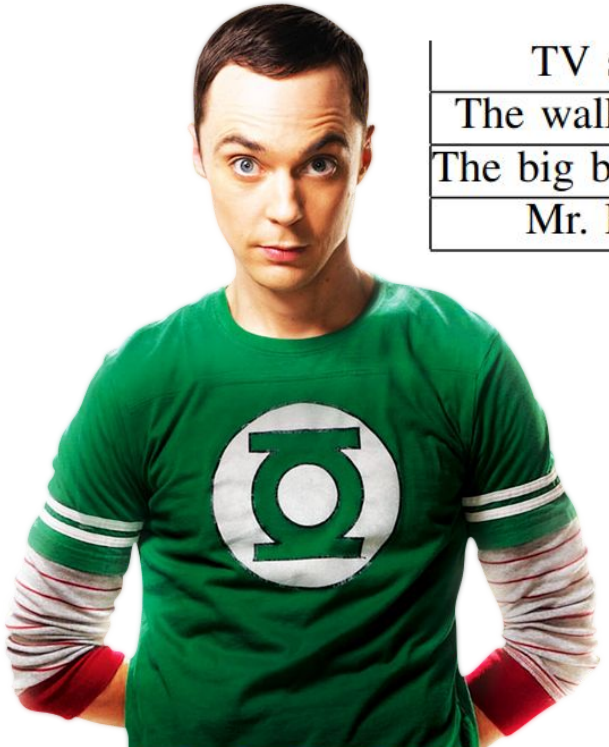
- has different category and subcategory, for each category a multilayer net was built
- for **activity:**
  - **5327 users** with **372571 edges**
  - **7495 items** and **262906 edges**
  - **19746 keywords** with **706595 edges**
  - a tensor with **158107 triples**

TV Series Results

SocialAU	User	*Fw*	*Fg*	NRMU	NRM	URM	$\mathcal{T}_u$	TOPHITS
1	ncis_cbs	1000000	76	0	1003	324	9	4326
2	shoresvkassie	500	56	0	45	7	1261	1
3	colors_infinity	20440	190	9	593	130	30	979
4	frizlj24	413	NA	20	2	2	3868	2
5	an0n88	NA	NA	161	0	0	256	40
6	rosalitamoog	302000	1284	72	0	0	208	3
7	whoismrrobot	217000	1181	0	323	185	0	12297
8	grantgust	1410000	660	3	420	394	3	3520
9	thewalkingdead	1190000	4680	1	480	427	2	940
10	m_weatherly	458000	187	0	189	114	0	12641
11	sawood69	706	892	64	0	0	49	6
12	donniewahlberg	1080000	7929	0	543	303	2	11209
13	tusharp75788052	317	425	128	0	0	134	524
14	walkingdead_amc	4450000	205	0	423	213	0	12732
15	h3ll0fri3nd1	607	218	43	30	17	51	532
16	ew	5510000	5740	0	301	298	2	6360
17	bigbang_cbs	2385	1833	0	553	491	3	2266
18	itsramimalek	197000	30	0	135	81	0	12296
19	sradhajena	2279	987	58	0	0	100	709
20	wheeler_forrest	10900	231	1	211	210	3	3527

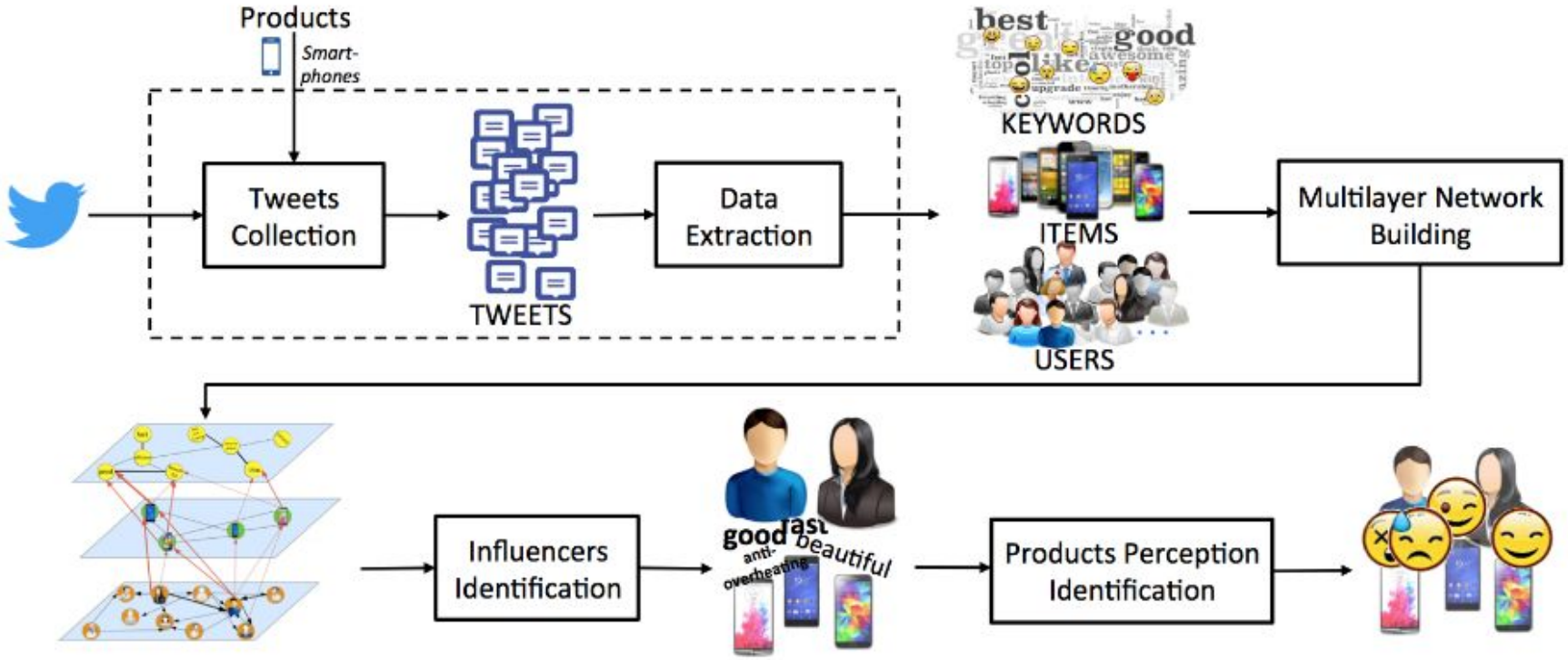
TOPHITS	*username*	*Fw*	*Fg*	*NRMU*	*NRM*	*URM*	$\mathcal{T}_u$	*SocialAU*
1	shoresvkassie	500	56	0	45	7	1261	2
2	frizlj24	413	NA	20	2	2	3868	4
3	rosalitamoog	302000	1284	72	0	0	208	6
4	janinfoster	392	40	1	1	1	85	35
5	walkingdead_ler	3140	2818	0	8	1	88	40
6	sawood69	706	892	64	0	0	49	11
7	walkingdeadbot	13300	8159	14	0	0	33	43
8	zombiemailman	3488	2930	13	0	0	24	28
9	frizman	49	7	0	0	0	92	75
10	coolstuff2get	3612	1431	0	2	2	41	108
11	ginatwdfan	NA	NA	5	0	0	17	38
12	ayedoukhay	1068	1497	8	0	0	17	39
13	jam_hirons	1087	763	13	0	0	30	79
14	vikingotwd	1662	3089	23	13	13	26	25
15	ftwdcollector	121	205	2	0	0	21	87
16	marian_banta	84	276	16	0	0	25	27
17	kyleabbot	61900	58100	0	0	0	170	94
18	lethahobbs141	531	1622	27	0	0	34	59
19	hughes6043	7064	6925	16	0	0	20	55
20	pjaycody1	2925	3096	16	0	0	24	78

TV Series Results



TV series	user	keyword
The walking dead	shoresvkassie	grave, new, comic, easy, sexy, bloody, flat
The big bang theory	frizlj24	new, fair, own, funny, only
Mr. Robot	colors_infinity	right, favourite, many, simple, iconic

SocialAU: Pipeline



Main Contributions

1. Content of web posts regarding a topic of interest posted by users is modeled by a three-layer network, completely different from the state-of-the-art approaches

2. **The greedy PARAFAC algorithm for computing the rank-1 approximation of the 3rd-order tensor representing inter-layer interactions has been extended to take into account the hub and authority scores determined by the HITS method on the users and keywords layers.**

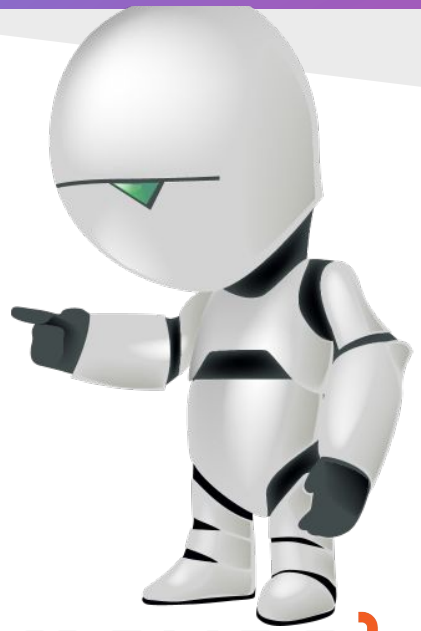
3. **SocialAU combines topological and context analysis to obtain influential users**

4. New evaluation measures are proposed to assess the capability of the approach to detect authoritative users expressing their point of view on the most discussed items by using the most dominant keywords

5. **Experiments** on TV series coming from Twitter, and a Yelp dataset reporting reviews on several categories, **show the ability of SocialAU to find users that are both authoritative in the user network, and very active in expressing their viewpoint**

Thank You!

Let's Chat!



GraphRM

{codemotion}
community

Some References

- *Oro E., Pizzuti C., Procopio N., Ruffolo M. Detecting Topic Authoritative Social Media Users: a Multilayer Network Approach. IEEE Transactions on Multimedia, 2017, 20.5: 1195-1208.*

- *Oro E., Pizzuti C. and Ruffolo M. (2018). **A Methodology for Identifying Influencers and their Products Perception on Twitter**.In Proceedings of the 20th International Conference on Enterprise Information Systems - Volume 1: ICEIS, ISBN 978-989-758-298-1, pages 577-584. DOI: 10.5220/0006675405770584*

- *Procopio, N. (2022). Influencer Centrality (Version 0.2.0) [Computer software].*