Detecting influential users in customer-oriented online communities

Ivan Nuzhdenko, Amir Uteuov and Klavdiya Bochenina
ITMO University, Saint Petersburg, Russia
ivanbor38@niuitmo.ru

Abstract. Every year the activity of users in various social networks is increasing. Different business entities can analyze in more detail the behavior of the audience and adapt their products and services to its needs. Social network data allow not only to find the influential individuals according to their local topological properties, but also to investigate their preferences, and thus to personalize strategies of interaction with opinion leaders. However, information channels of organizations (e.g., community of a bank in a social network) include not only target audience but also employees and fake accounts. This lowers the applicability of network-based methods of identifying influential nodes. In this study, we propose an algorithm of discovering influential nodes which combines topological metrics with the individual characteristics of users' profiles and measures of their activities. The algorithm is used along with preliminary clustering procedure, which is aimed at the identification of groups of users with different roles, and with the algorithm of profiling the interests of users according to their subscriptions. The applicability of approach is tested using the data from a community of large Russian bank in the vk.com social network. Our results show that: (i) it is important to consider user's role in the leader detection algorithm, (ii) the roles of poorly described users may be effectively identified using roles of its neighbors, (iii) proposed approach allows for finding users with high values of actual informational influence and for distinguishing their key interests.

Keywords: social network analysis, opinion leaders, topic modeling, opinion mining

1 Introduction

The popularity of online social networks (OSNs) has led to the development of a wide variety of algorithms and tools to analyze different aspects of human activities and behavior in online context. Digital traces of individuals can give us useful insights on their habits, preferences, emotional state, structure and dynamics of social contacts and their involvement in information spreading. One of the most studied fields along with modeling cascades of information messages (review in [1]) is the detection of influential users in OSN (e.g. [2][3]).

The vast majority of studies on finding influential users is focused on topological properties of nodes in a network. The restrictions of pure network-based methods lead to a necessity of data-driven algorithms development which combine topological and individual characteristics aiming to solve a domain problem. In this study, we show an example of such approach for customer relationship management via specialized enterprise communities in OSN. We are focused on what we call customer-oriented online communities – communities which are created by a representative of an enter-

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_82

prise to inform the clients about the news, to answer their questions and to provide them any desirable support.

There are different ways to find opinion leaders in a social graph (i.e. to determine the users which information messages may have a strong impact on their audience). These approaches are: (i) using degree centrality, (ii) grouping local topological measures (e.g. different centrality measures [2]), (iii) using connectivity properties [4], (iv) exploiting both topology and semantic (for instance, [3] combines PageRank and sentiment analysis), (v) exploiting history of users' feedback (e.g. [5]). Existing algorithms mostly do not account for the functional role of a user within a community; however, there exist communities in which this role clearly influences the level of involvement of users within the context. In this study we propose an algorithm which exploits different kinds of data available in customer-oriented online communities to identify influential users who are not affiliated with community owner and belong to the target audience of a domain community. The paper is organized as follows. Section 2 gives a description of a problem and of proposed method. Section 3 describes a dataset. Finally, Section 4 demonstrates the results.

2 Method

By the customer-oriented community in online social network we mean a group in a social network which is owned by a particular stakeholder (e.g. a bank or a retailer) to inform the clients about news of organization, to provide real-time support, to answer the questions, to promote campaigns etc. In the customer-oriented community (in contrast to, e.g., entertainment community) the goal of the stakeholder is to interact only with the target audience to increase their loyalty and lifetime value.

The definition of a leader in a customer-oriented community also changes compared to ordinary communities. The leader of the customer-oriented online community is not only a person with high impact on their audience, but also the one who belongs to the target audience of community stakeholder (for a bank it will be a client or potential customer, not an employee and not a bot). After the opinion leaders in the customer-oriented community are detected, a stakeholder may suggest them personalized offers with account of their interests (this is the case that we consider in frames of this study) or to consider the interests of groups of influential users while developing strategies of customer relationship management. In this study, we use information from community, topology of network of subscribers and user profile data to identify the influential users among the people with a required role ("customer").

We obtain the data from the largest Russian online social network vk.com (denoted as VK). It supports automated collection of: (i) a list of community subscribers, (ii) a list of subscribers' friends, (iii) reactions of users to community posts (likes, shares, comments), and (iv) for each user – sex, age, place of work. Moreover, users in VK have personal page including the following additional information: (i) "wall", which means the chronologically sorted list of user posts and shares (information from a wall may also be commented and shared by the friends of a user), (ii) list of user subscriptions which may be used to estimate their interests.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_82

All these data may be used to estimate target properties of opinion leaders (Table 1, Algorithm 1). By target property we mean the parameter of a leader which is important in a context of customer-oriented community. For example, the level of user involvement in a community "life" influences the potential informational influence on this user by messages generated by this community (if a user has several hundred subscriptions, it will likely miss the information from a single one).

Table 1 – Algorithms: (i) influential users' detection, (ii) user affiliation detection

(i) Algorithm 1. Influential users detection	(ii) Algorithm 2. User affiliation detection
Input: Followers (list of followers) 1 **for** f **in** Followers**:** 2 **if** f**.**workplace **!=** bank 3 **if** f**.**friends_number > M 4 **if** f**.**friends_inside_comm > N 5 **if** f**.**posts_number > P 6 **if** f.avg_likes_per_post > Q 7 then**:** 8 f.is_influential = **True** 9 **if** f.is_influential 10 interests = f.get_interests**()**	Input: G (network of subscribers) 1 **for** cluster **in** G.get_clusters(): 2 label = cluster.workplace.most_common() 3 **if** label.frequency() > threshold 4 then: 5 **for** follower **in** cluster 6 follower.affiliation = label

Algorithm 1 is organized as a system of filters. It takes into account the potential information impact and behavioral characteristics forming a certain level of trust in users. Values *M*, *N*, *P* and *Q* may be chosen as *k*-th (e.g. 0.9) quantile of the corresponding distribution or to be set according to a desired threshold. The majority of users do not specify information about their place of work in the OSN which is needed by Algorithm 1. To address this, we use procedure of user affiliation detection described in Algorithm 2. To identify the interests of users (Algorithm 1, step 10), the information on users' subscriptions was processed (see the details in Section 4.2). The output of Algorithm 1 is a list of opinion leaders with their personal interests.

3 Data description

To perform the study, we collected the data of the community of a large Russian bank in VK OSN for a period 26.04.2016–20.09.2017. The page of a community contains 400 posts with 29139 followers, 10682 likes, 1850 comments and 1330 shares in total. For each of the followers we also collected data required by Algorithms 1, 2 (see Table 1). To check the actual information impact of influential users, we also collected data on likes, comments and shares on all posts from users' personal pages ("walls"). The size of the resulting dataset amounted to 23 GB. It contained 973 000 posts with information about likes, shares, comments and date of the publication (Table 2).

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_82

Table 2 – Descriptive statistics of dataset with personal pages of community followers

Target property	Mean	Median	Max
Average likes per post	3.6	0.48	850
Average shares per post	0.12	0	65.5
Posts number	59.7	15	16599
Friends number	337.8	170	10000
Friends inside community	2.6	1	708
Subscriptions number	244.7	102	5008

To study local topological properties characterizing user behavior a network of community subscribers was created. A node in the network represents community user and an edge represents friendly link between users. Thus, a resulting network contains only friend relationships between subscribers and do not contain relationships with people who are not members of a community. The resulting graph has 17580 nodes and 38593 edges, and average clustering coefficient equals to 0.09.

4 Results

4.1 User affiliation recovery

Figure 1 represents a visualization of the giant component of a network of subscribers (15000 nodes) after clustering (see Algorithm 2). Different colors represent different clusters, a size of a circle represents a number of friends inside the community. As it was expected, the information about user affiliation was available for a low percentage of users (9 %). In total, 2083 different occupations were found. As a result, from 2524 users with known occupation 199 (8 %) were labeled as affiliated with a bank.

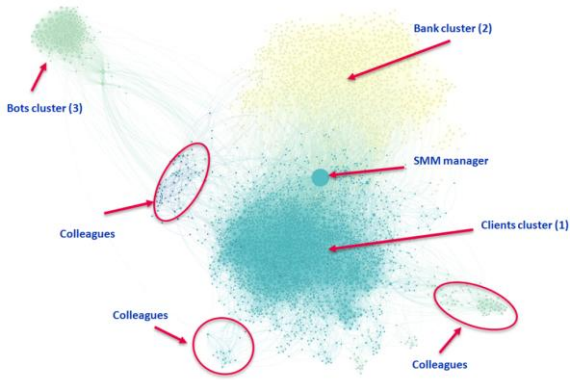


Figure 1 – Giant component of a community of a bank
(cluster 1 – aquamarine, cluster 2 – yellow, cluster 3 – green)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_82

To assign labels to clusters, we calculate a percentage of users with a bank and non-bank affiliation for each cluster. Although the percentage of users with known workplace for clusters 1 and 2 is similar (11% and 15%, respectively), the concentration of bank employees in a cluster 2 is extremely high (49%). At the same time, cluster 2 contains only 2 % of users with known occupation from the bank. This suggests that cluster 1 is a cluster of clients, and cluster 2 is a cluster of bank employees. Cluster 3 was marked as cluster of bots by analysis of the similarity network of users. The smaller clusters of users from the same organizations can also be distinguished (the examples are shown in Figure 1).

4.2 Detection of influential users and their interests

For this chapter, we used the following values of filters: a) more than 200 total friends, b) more than 20 friends inside the community, c) less than 100 subscriptions to other groups, d) more than 10 posts on a personal page. After the filtration process, the algorithm divided extracted followers on four clusters which we denote as bank_possible, bank_verified (true positive), ok_possible, ok_verified (true negative), see Table 3. Bank_verified means that the follower has an affiliation with a bank verified by the workplace field on the personal page and the detection algorithm assigns this follower to the bank cluster. Bank_possible means that the algorithm assigns a follower to the bank cluster but there is no information about the workplace on his or her page. The same logic holds for ok_verified and ok_possible clusters. There was only one false negative from 45 users, and there were no false positive cases.

Table 3. Characteristics of groups (user affiliation detection algorithm): 1 – avg friends count (total), 2 – avg friends count (inside community), 3 – avg likes count, 4 – avg shares count, 5 – avg friends count from “bank” cluster, 6 – avg friends count from “non-bank” cluster

Cluster (size)	1	2	3	4	5	6
bank_possible (23)	372	32	14.15	0.2	30	2
bank_verified (10)	413	40	11.42	0.14	35	5
ok_verified (6)	3152	26.5	33.54	3	2.16	24.16
ok_possible (5)	2364	38.8	31.1	1.48	1.2	39

Table 3 demonstrates that ok_verified cluster has very close parameters to ok_possible while bank_verified and bank_possible are similar to each other. From the 23 users within the cluster, 16 (69.5 %) were classified as actual and former employees of the bank, 4 (17.4 %) were classified as “unknown”, 2 (8.7 %) were classified as “non-bank” and one user has deleted the profile. The low count of false positive cases (only 2 of 23) suggests the good quality of affiliation detection algorithm.

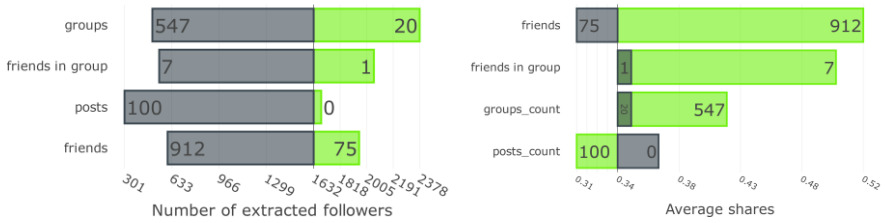
To show the importance of user affiliation recovery procedure for leader detection in customer-oriented communities, we compare the results of our algorithm with basic topological-based approaches. Table 4 show the results for top-500 influential users. 33 % of important nodes determined by degree centrality were affiliated with a bank and 14 % was marked as bots. In case of betweenness centrality – 12 % of followers

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_82

were in the bank cluster. 20 % of users provided by PageRank was affiliated with bank and 5 % was marked as bots. It means that these algorithms may provide from 10 to 50 percent of non-target audience in the resulting list of the opinion leaders.

Table 4. Quality of target audience detection for topological-based approaches

Metric	Ok (% of total)	Bank (% of total)	Bots (% of total)	% of non-target audience
Degree	53 %	33 %	14 %	47 %
PageRank	75 %	20 %	5 %	25 %
Betweenness	87 %	12 %	1 %	13 %



a) b)

Figure 2 – Tornado chart for: a) extracted followers, b) average shares

Figure 2 gives the tornado diagrams for the number of extracted followers and average shares which show effect of input on output parameters. Median values were used to calculate a baseline. Border values corresponds to 25 and 75 percentiles. The most important parameters are number of friends within community/in total.

To detect leaders' interests, we applied topic modelling for their subscriptions (12K communities). In this study, ARTM model was used as it showed better results compared to LDA [6]. Topic extraction algorithm created 32 different clusters. After that, we have found the preferred topics for particular users according to their subscriptions (Figure 3). Two interests with high frequency can be distinguished for the leaders: beauty salons and theaters/museums.

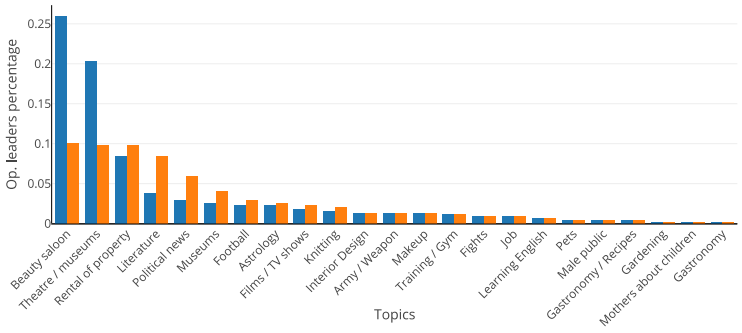


Figure 3 – Comparison of interests of leaders and ordinary users)
(blue – opinion leaders, orange – ordinary users)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_82

5 Conclusion

The recent trend in approaches to influential user detection in online social networks incorporates different types of available information while deciding on the potential impact of a user. In this study, as an example of such a problem we consider detection of leaders in customer-oriented online communities. The detection procedure is organized as a system of filters accounting for different target parameters of potential leaders. We account for different roles of users inside a community by introducing user affiliation recovery algorithm. Finally, we supplement leader detection algorithm with a tool for identifying interests of users according to their subscriptions. The experimental part of the study was conducted using the data on a banking community from the largest Russian social network, vk.com and showed the acceptable accuracy of affiliation detection and the restricted applicability of pure network-based approaches for customer-oriented communities.

Acknowledgements

This research is financially supported by The Russian Science Foundation, Agreement №17-71-30029 with co-financing of Bank Saint Petersburg.

References

[1] M. Li, X. Wang, K. Gao, and S. Zhang, "A Survey on Information Diffusion in Online Social Networks: Models and Methods," *Information*, vol. 8, no. 4, p. 118, 2017.

[2] "Identifying Key Opinion Leaders Using Social Network Analysis," *Cogniz. 20-20 Insights* |, no. june, 2015.

[3] M. Zhu, X. Lin, T. Lu, and H. Wang, "Identification of Opinion Leaders in Social Networks Based on Sentiment Analysis : Evidence from an Automotive Forum," *Adv. Comput. Sci. Res.*, vol. 58, no. Msota, pp. 412–416, 2016.

[4] N. A. Helal, R. M. Ismail, N. L. Badr, and M. G. M. Mostafa, "A novel social network mining approach for customer segmentation and viral marketing," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 6, no. 5, pp. 177–189, 2016.

[5] A. Sheikhahmadi, M. A. Nematbakhsh, and A. Zareie, "Identification of influential users by neighbors in online social networks," *Phys. A Stat. Mech. its Appl.*, vol. 486, pp. 517–534, 2017.

[6] K. Vorontsov and A. Potapenko, "Additive regularization of topic models," *Mach. Learn.*, vol. 101, no. 1–3, pp. 303–323, 2015.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93713-7_82