# Profiling Less Active Users in Online Communities

Alexandra Barysheva, Anna Golubtsova, and Rostislav Yavorskiy

Department of Data Analysis and Artificial Intelligence
Faculty of Computer Science
Higher School of Economics
Myasnitskaya 20, Moscow, Russia, 101000
{asbarysheva, annagolubtsova1, ryavorsky}@gmail.com

**Abstract.** Our research is focused on the study of social interactions of online community users, especially in business-oriented social networking services like LinkedIn or Habrahabr. The general aim of the work is to design methods for profiling of discussion participants within groups according to their interaction patterns. One of our goals is to make the approach independent from the language of communication, that is why we build our analysis on the comments graph and do not use information from the posts content. This paper suggest FCA based approach to profiling less active users for which not much data is available and statistical analysis is not applicable.

**Keywords:** online community, communication graph, user profiles

## 1 Introduction

Social Internet development unveiled great research potential for network analysis which includes the analysis of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities.

This paper focuses on the behaviour patterns of the members of social networks groups (communities) of interest. In these online groups, users on a regular basis can publish information or news that called posts, and interact with each other by commenting or liking them.

The general goal of this work is to provide a method of profiling group users by analysing the group interaction graph. An interaction graph is a graph where vertices correspond to users and edges represent relation "*user A comments or likes post of user B*".

Today almost every social media site provides an API for easy data retrieval. Application programming interface (API) is the set of routines, protocols, and tools for building software applications using the obtained data. In order to retrieve the graph of social interaction we use data sets collected from business-oriented social networking service LinkedIn and Habrahabr (leading Russian blog on Information Technology topics).

In this paper we continue research described in [1], which is also dedicated to the task of profiling online community users. The method proposed in that paper is based on clustering users according to statistical characteristics of their communication patterns.

Clustering based on statistical characteristics allows one to study the communication patterns, but it is not applicable to users with low activity for which not enough data is available. Our study shows that actively involved users constitute approximately 2% of a community, while more than half of the community member could be classified as "observers" (see [1]). That motivates us on designing a separate technique for profiling less active members of an online community.

The paper is organized as follows. Section 2 contains the review of relevant related works organized according to the used approach. Section 3 describes the data set. Section 4 summarizes achieved and anticipated results. In Section 5 we conclude and discuss the possible applications of our work.

## 2 Related work

Relationship is a central concept of the science of Social Network Analysis. Our race, ethnicity, background and personality — all influence our behaviour and interact. Thus, the behavioural patterns analysis in online communities can provide the information about the user that is not explicit in his or her profile page (and obviously cast some light on the principles of social behaviour in online networks).

There are different types of relationships between people: friendship, trust, influence, or conflict, dislike etc. In [2] authors provide several types of relationships in social networks including (1) binary and valued relationships, (2) symmetric and asymmetric relationships, and (3) multimodal relationships. Examples of binary and valued relationships are *"Sam follows Ann on Facebook"* and *"Alex retweeted 4 tweets from Mary"* respectively. Following or reposting on Twitter, Facebook or LinkedIn are asymmetric relationships by definition, but a follow-back tie can exist, thus symmetrizing them. An example of symmetric relation is *"Ann and Bob have common interests"*. Multimodal relationships are interactions between actors of different types  people possess information, group adds people, and so on. In our study, we analyze all these three types of relationships between group users.

The task of user profile modelling consists of many subtasks and approaches, such as content-based methods [3], the island method [2], researching of users friend- or following-connections, or tracing user activity [4] etc. Typically, the majority of proposed profiling methods combine different techniques.

An example of Twitter users profiling is presented in [5]. Authors study demographic estimation algorithms based on users tweets and community relationships. They propose a hybrid community-based and text-based method where demographics of Twitter users are estimated by tracking the tweet history and clustering the followers/followings. The method estimates wide varieties of demographics such as gender, age, area etc. The authors also consider users with few tweets such as followers of corporate accounts.

In [6] authors suggest a generic model for user classification in social media with application to Twitter. Analyzing the users behaviour, linguistic content and the network structure of the users Twitter feed they develop the method of automatic inferring the values of user attributes such as political orientation or ethnicity. Machine learning approach is used relying on four general feature classes: user profile, user tweeting behaviour, linguistic content of user messages and user social network features. The paper presents experimental results on 3 tasks with different characteristics: ethnicity identification, political affiliation detection and detecting affinity for a particular business.

A weakly supervised approach to user profile extraction from Twitter is also suggested in [7]. In addition to traditional linguistic features, this approach also takes into account network information, offered by social media. Authors use users profiles from social media websites such as Facebook or Google Plus as a distant source of supervision for extraction of their attributes from user-generated text. They test the algorithm on three attribute domains including spouse, job and education and results demonstrate accurate predictions for users attributes based on tweets.

Unlike previous mentioned works, article [4] focus just on user activity, ignoring the content of messages a user exchanged. Authors take into consideration both social interactions and tweeting patterns of microblogging integrating service Twitter, which allow profiling users according to their activity patterns. According to the investigation, there are 75 % of the users in their appropriate cluster, which can be classified with a 0.9 assignment probability. Clusters are characterized by a set of statistical features relating user activity, network structure and dynamic patterns. Furthermore, the authors propose three algorithms to analyze the impact of content posted by a user.

In [8] authors use modelling user profile to predict the profile of another user in the network. They gather fine-grained data from two social networks and try to infer user profile attributes. The article proposes a method of inferring user attributes that is inspired by previous approaches to detecting communities in social networks based on the fact, that users with common attributes are more likely to be friends and often form dense communities. Results show that certain user attributes can be inferred with high accuracy when given information on as little as 20% of the users.

The expertise retrieval task of user profiling is also mentioned in [9]. In this work, the topical profiling task is decomposed into two stages: (1) discovering and identifying possible knowledge areas, and (2) measuring the persons competency in each of these areas.

Our research has the same goal as the previous mentioned works — to provide a method of profiling users. Main task of this paper is to suggest an approach for profiling of less active users, which usually form the majority of any online community. Since we have little data for these members, statistical analysis of their behaviour is not possible. That is why we turn to FCA tools.

The idea to apply formal concept analysis to social network analysis is not new, see e.g. [10] or [11]. Usually the technique is used for network clustering and detecting communities. Our goal is slightly different, we assume that a community is already given. We target at detailed description of roles of different users in this community.

Method of retrieving groups of websites users with similar behaviour using Formal Concept Analysis presented in [12]. Authors propose to construct a taxonomy based on users visits of different pages of websites. The problem of big number of concepts is solved by applying the stability index [13] to the lattice concepts.

Extension of using the stability index (not only in terms of intent stability, but also in terms of extent stability) to taxonomy construction described in [14]. For instance, in this work authors study the dataset from research by Davis, Gardner and Gardner [15] which features ladies attending particular events in a small Mississippi town in the 1930s. By constructing stabilised lattice (according to extent) authors found the core members in groups.

## 3 Description of the data set

In our work we use two data sets. The first one is communication graph retrieved from Habrahabr blogging site for several most popular topics. The second includes communication graphs for several LinkedIn groups.

### 3.1 Habrahabr data

HabraHabr (http://habrahabr.ru/) is the most popular Russian blog service devoted to Information Technology. Currently we work with communication graphs for the most active topics including ”Algorithms”, ”Big Data”, ”High performance computing”, ”Information security” and others. For a given topic the dataset is a single table with the following columns:

- Post Id
- Post author
- Comment Id
- Comment author
- Parent comment Id
- Time stamp

Also, for conveniece we added some derived values like comment depth, number of child comments etc.

Data set for ”Big data” community in CSV format is available at the project page on GitHub, see https://github.com/ryavorsky/HabraGraph.

### 3.2 LinkedIn data

Business-oriented social networking service LinkedIn, http://www.linkedin.com, allows users to create profiles and interact with each other in an online social network, which may represent real-world professional relationships. LinkedIn also supports the formation of interest groups that are, generally, employment related, although the majority of topics are covered mainly around professional and career issues. Currently we work with communication graphs for the largest groups related to the topic "Bioinformatics". The dataset has the same structure: Post Id, Post author, Comment Id, Comment author, Like author, Time stamp, and some derived characteristics, such as the number of child comments and its depth in the thread.

## 4 Users profiling

### 4.1 Clustering according to the statistical characteristics

In this paper we continue research on profiling online community users described in [1]. Firstly, a set of the user attributes that can be computed using community comment graph and post comment graphs is listed. They are: the number of people that leave comment to the user and were commented by the user, the number of posts the user wrote and commented, average depth of the user's comment and how often the user's comment was a terminal. These attributes reflect the user's communication style in online discussions.

The clustering allowed us to figure out the following user types:

1. **Silent stars** (2 users). Authors of popular posts who do not participate in the discussions.
2. **Communicative stars** (2 users). Authors of popular posts who are actively involved in the discussions.
3. **Active chatters** (2% of users). Participants who leave many comments, and reply to almost every comment on their posts.
4. **Idle chatters** (2% of users). People who write few comments, but usually their comments support the subsequent debate.
5. **Socializers** (5% of users). Users who do not produce many comments, although the number of people their talk with is notably high.
6. **Investigators** (15% of users). Participants who communicate with many people within very narrow discussion (few blog posts).
7. **Concluders**(22% of users). Participants, who produce little comments and quite often their comment is the last one in the discussion branch.
8. There is also one more type of user - **observers** (more than 50%) who are the most inactive users: each one leaved no more than 3 comments.

It can be seen that less active users represent bigger part of community. That is why the goal of the current work is to provide method of detection of dependencies between users with different activity rate. In other words we want to know to whom among the active users the less active users are similar.

### 4.2 Profiling of less active users

As it was mentioned above, the task of analyzing and profiling of user behaviour is rather straightforward for more active users, when a lot of data is available.

For the other part of online community, the majority of less active members, we suggest to describe the user profile in terms of similarity to one or several preselected benchmarks, key users of the community with well-known behavioural patterns.

In more details the suggested procedure is the following. First, select a small number of key users of the studied community. Second, build the object-attribute table, in which rows (objects) are all group users and columns (properties) are benchmark profiles (key users). Then use use FCA tools to compute the lattice of formal concepts. Finally, conclude that activity pattern of user user1 could be described in terms of intersection of few benchmarks, e. g. core_user1 and core_user3.

### 4.3 Core users

There are many different ways to determine the set of benchmarks. In our work we use the notion of communication graph core [16,17]. The picture on fig. 1 shows 3-core for users-posts graph (that is the largest subset of users and posts, in which each user left comments in at least three posts and each post has comments from at least three users) corresponding to "big data" community at Habrahabr.ru platform.

Restricting our focus with the 3-core helps us to filter out blog posts which are not very relevant to the main community topic, and also figure out users, which play central role in the group communications.

Consider for example an irrelevant post, which produced quite intensive discussion. We can detect its irrelevance by the fact that core users did not participated in the thread. Formally, to classify a post as a core one we require that at least 3 core users participate in the discussion.

Similarly, there might be a user, who went into long comments exchange in a single thread or left few remarks in some rather irrelevant discussions. User with such a behaviour usually is interpreted as a casual visitor, not a core one. To be included into the community core we require that the user should participate in at least 3 core discussions.

### 4.4 Why FCA

As it was already mentioned above, the main goal of this work is to design an approach for describing profiles for majority of less active members of an online community. Usually we have very little data for such users, a couple of comments or so. That is why classification according to numerical characteristics suggested in [1] hardly makes sense. The users will be classified as "inactive" and that's it.

In this paper we suggest to use information about the particular topics, which attracted a user. That data has "object-property" type, so we turn to FCA.

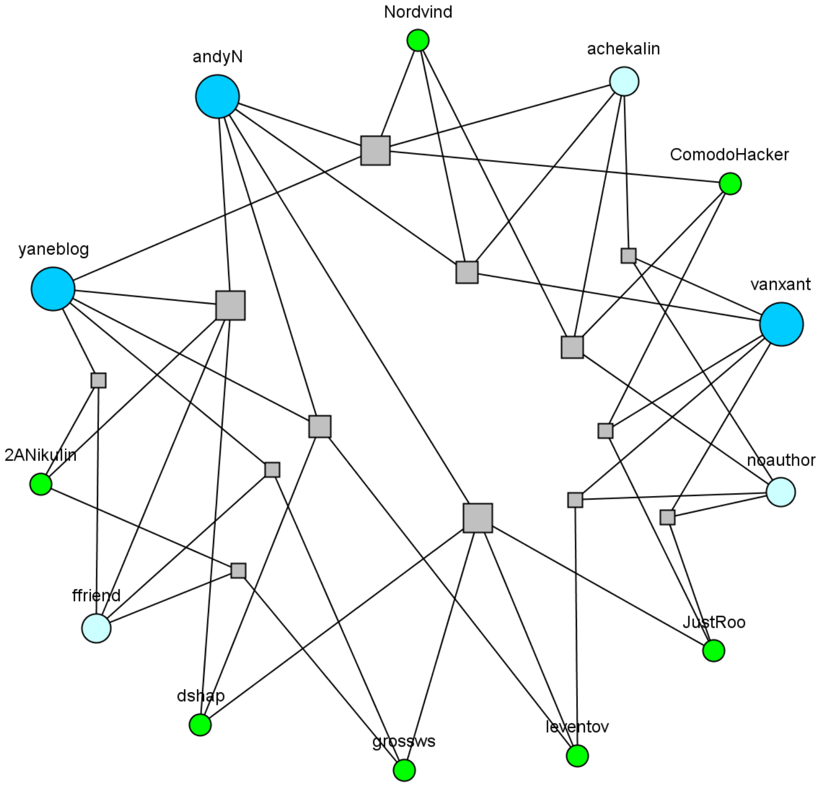

**Fig. 1.** 3-core of for graph corresponding to "big data" community at Habrahabr.ru

The formal context is defined as follows. Assuming user is a group user and benchmark is a core user we say that object user has property benchmark if users user and benchmark together participated in a post discussion. For the example of "big data" community mentioned above the formal context table has 13 properties (the number of core users) and hundreds of objects (for all the other community members).

We use FCA tools (Concept Explorer [18] and FCArt [19,20]) to build the lattice of the formal context. As a result, the set of formal concepts is given (see fig. 2). Each formal concept has a set of objects (extent). These users are similar to each other and their profile could be specified in terms of the benchmarks, the set of core users.

By combining the 3-core graph and the lattice we can get a visual map of the community, see fig. 3.

Also, in applications we can use the resulting formal concepts for introducing the corresponding links between the user profiles. Indeed, for a user with low activity the profile will be almost empty due to lack of statistics. The links from

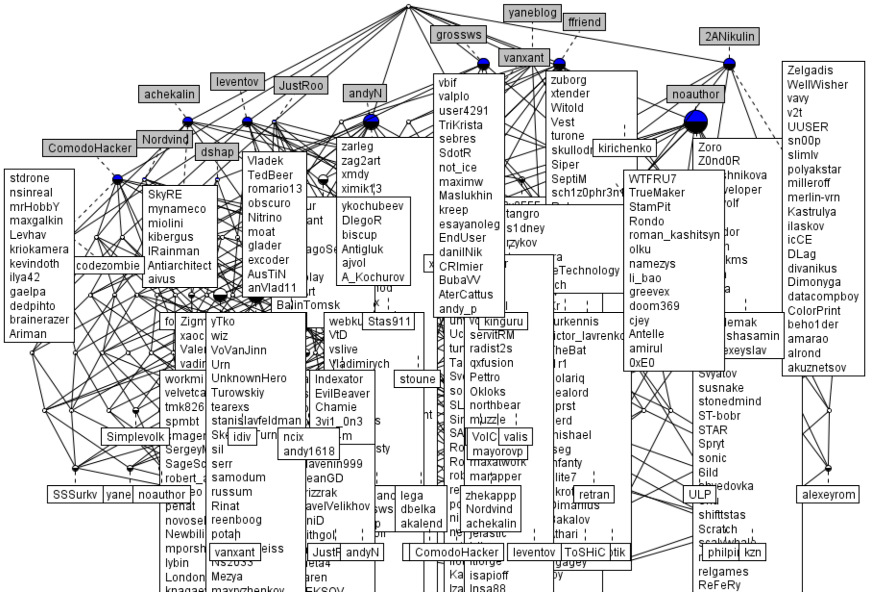

**Fig. 2.** Lattice of users-benchmarks formal context corresponding to "big data" community on Habrahabr

this empty profile to more detailed profiles of most similar core users will help to get at least some information about the user interests.

## 5 Conclusion

The paper describes work in progress on developing a universal tool for automated building of profiles for online community users. The proposed method is based on the user activity in the process of posting, liking and commenting group posts.

To make the approach suitable for the analysis of different online communities, the approach does not use information from the user profile or content analysis. Thus, it is based on user activity and his/her skills to interact with other group participants.

In order to classify less active online group members the method based on retrieving formal concepts with core users as attributes is suggested. The results can be used to extend the functionality of the groups with the detailed description of the profile of participants and the nature of their interaction, which in turn should help to understand users behaviour.

The developed method can be applied to any online community.

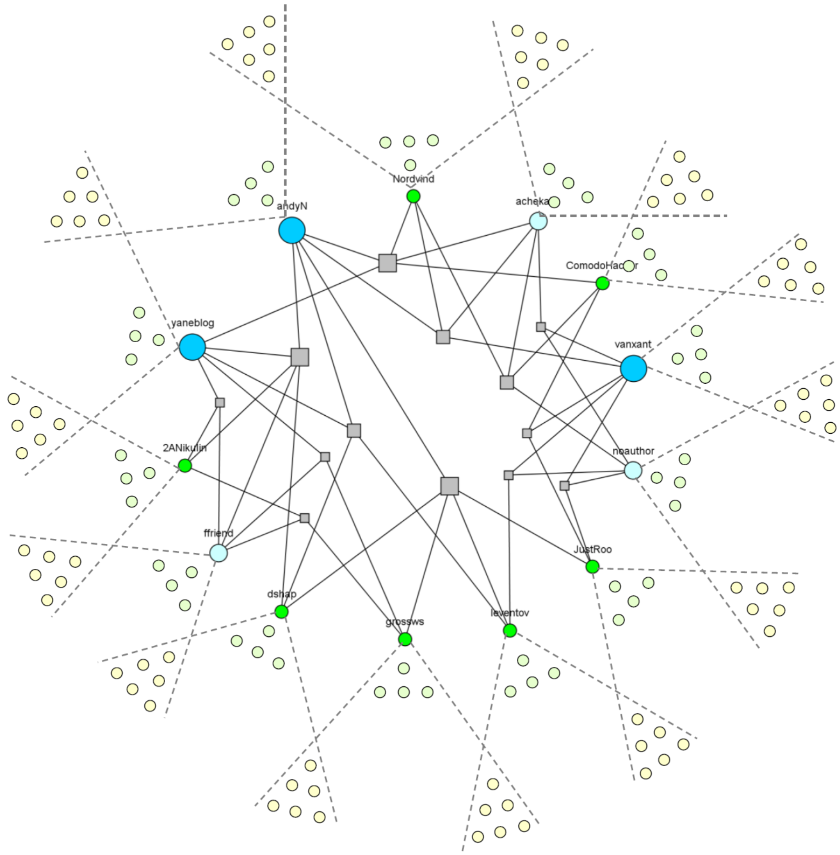

**Fig. 3.** The map of the Habrahabr "big data" community built from the 3-core graph

## References

1. Barysheva A., Yavorskiy R. *Building Profiles of Blog Users Based on Comment Graph Analysis*, Proceedings of AIST'2015, 4-th International Conference "Analysis of Images, Social Networks and Texts", Yekaterinburg, 9-11 April 2015. To appear in Springer CCIS.
2. Kouznetsov A., Tsvetovat M. **Social network analysis for startups** O'Reily, 2011.
3. Santosh R. *Author Profiling: Predicting Age and Gender from Blogs,* PAN at CLEF, 2013.
4. Rocha E. *User profiling on Twitter,* Semantic Web. Interoperability, Usability, Applicability, 2011.
5. Kazushi I. *Twitter user profiling based on text and community mining for market analysis* Knowledge-Based Systems, 2013, pp. 35-47.
6. Pennachiotti M. A. *Machine Learning Approach to Twitter User Classification,* Fifth International AAAI Conference on Weblogs and Social Media, 2011, p. 45.

7. Li J., Ritter A., Hovy E. *Weakly Supervised User Profile Extraction from Twitter* ACL, 2014.
8. Druschel P., Gummadi K. P., Mislove A., Viswanath B. *You Are Who You Know: Inferring User Profiles in Online Social Networks* ACM WSDM, 2010.
9. Balog K., Fang Y., de Rijke M., Serdyukov P., and Si. L. *Expertise Retrieval*, Foundations and Trends in Information Retrieval, 6 (2-3), 2012, pp. 127-256.
10. Snasel, Vaclav, Zdenek Horak, and Ajith Abraham. *Understanding social networks using formal concept analysis.* Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 03. IEEE Computer Society, 2008.
11. Gnatyshak, D., Ignatov, D. I., Semenov, A., Poelmans, J. (2012). *Gaining insight in social networks with biclustering and triclustering.* In Perspectives in Business Informatics Research (pp. 162-171). Springer Berlin Heidelberg.
12. Sergei O. Kuznetsov, D.I. Ignatov, Concept Stability for Constructing Taxonomies of Web-site users. In: S. Obiedkov, C. Roth, Eds., Proc. Social Network Analysis and Conceptual Structures: Exploring Opportunities, Clermont-Ferrand, 2007.
13. Kuznetsov, S.O.: On stability of a formal concept. In SanJuan, E., ed.: JIM, Metz, France (2003)
14. Sergei O. Kuznetsov, Sergei Obiedkov and Camille Roth, Reducing the Representation Complexity of Lattice-Based Taxonomies. In: U. Priss, S. Polovina, R. Hill, Eds., Proc. 15th International Conference on Conceptual Structures (ICCS 2007), Lecture Notes in Artificial Intelligence (Springer), Vol. 4604, pp. 241-254, 2007.
15. Davis, A., Gardner, B.B., Gardner, M.R.: Deep South. University of Chicago Press, Chicago (1941)
16. Batagelj V., Zaversnik M. *Generalized Cores*, arXiv:cs/0202039v1, 2002.
17. Seidman S. B. *Network structure and minimum degree* Social Networks, 5, 1983, pp. 269–287.
18. Yevtushenko, Serhiy A. *System of data analysis"Concept Explorer".* Proceedings of the 7th national conference on Artificial Intelligence KII. Vol. 2000. 2000.
19. Neznanov, Alexey, Dmitry Ilvovsky, and Andrey Parinov. *Advancing FCA Workflow in FCART System for Knowledge Discovery in Quantitative Data.* Procedia Computer Science 31 (2014): 201-210.
20. Neznanov, A. A., and A. A. Parinov. *FCA Analyst Session and Data Access Tools in FCART.* Artificial Intelligence: Methodology, Systems, and Applications. Springer International Publishing, 2014. 214-221.