Measuring community influence: a data mining approach to Eastern European blogosphere

Mclean N. (Commonwealth of Australia)

Структурно-алгоритмический подход к исследованию европейской блогосферы

Маклин Н. (Австралийский Союз)

Маклин Наталия / Mclean Natalia - – аспирант,
факультет информационных технологий,
Университет Сиднея, г. Сидней, Австралийский Союз

Abstract: *we use data mining techniques (in particular, social network analysis approach) to measure the influence and 'community spirit' coefficient of various online groups in the Easter European blogosphere. The purpose of this paper is to demonstrate how we collect, clean and model inter-member relationships data of a large number of groups hosted on an Eastern European SNB Diary.ru in order to measure the cohesion of the group – i.e., how strongly members of a group as a whole are bonded with each other. Translating the notions of friendship and communication into the realm of computational analytics and assigning a number to every group will arm us with important scientific data that can be used in a variety of subsequent experiments and research projects (for example, creating a classification of groups or discovering links between a group cohesion coefficient and a central theme or interaction style of the group).*

Аннотация: *в статье описан структурно-алгоритмический подход к анализу групп пользователей европейской блогосферы. Статья уделяет особое внимание методам сбора, очистки и моделирования данных для измерения коэффициента структурной целостности группы.*

Keywords*: social network analysis, data mining, linked web data, social networks, online communities, computational sociology, blogosphere.*

Ключевые слова*: анализ данных, структурный анализ, социальные сети, блогосфера, алгоритмическая социология.*

Groups and communities within SNBs

Blogs [1] provides a low-cost platform of self-publishing and self-promotion where authors can broadcast either their professional narrative or life stories to potentially unlimited audience [2]. "[I]n cyberspace, increasingly, the dream is not just 'owning a house' – it's living in the right neighborhood" [3, p. 14], and this search for right web neighbourhood finds its reflection in blog statistics today: less than 20% of bloggers choose stand-alone blogs, while the rest (more than 80%) prefer to host their blogs on social network blog sites [4].

Social network sites (SNS) are "on-line environments in which people create a self-descriptive profile and then make links to other people they know on the site, creating a network of personal connections" [1,5]. Social networked blogs (SNB) are constellations of personal blogs that are united by a larger digital platform with SNS functionality. SNBs Diary.ru and Livejournal were pioneers in social network blogging. They have added another dimension to a blog: a *profile* that represents an author of a blog, and a publicly visible of articulated connections between bloggers (knows as "friend list"). By doing so, Diary.ru has revolutionised the practice of blogging making relationships between bloggers as significant as the content of their blogs [2,6].

This study is based on Diary.ru, but the methodological framework can be successfully translated to analysis of Livejournal or any other SNB.

Groups (also known as 'communities' on some SNBs) are clearly defined symbolical public social places emerged between blogs [3, p. 14,7]. Groups have strict boundaries and a list of members, and serve as gathering and discussion spaces for bloggers and facilitate community engagement [2,4]. Readers of a group can be authors, and vice versa [2]. A group aggregates members with a purpose of forming a community around a shared topic.

A number of web sociologists and researchers in computer-mediated communications [8-11] even suggest that the concept of the third spaces can be easily transferred from offline, physical world to online, virtual world. Therefore, online environments which satisfy the outlined characteristics of a third place can be conceptualised as *virtual third places* [8]. Computer-mediated social context like forums, MUDs, chats, and of course, *groups* – "[n]either work nor home, such public and quasi-public spaces provide forums for sociability and interconnection with others" [9, p. 142], similar to physical third places, like coffee shops or exhibitions [8,11]. Not every group provides an atmosphere for its members to get to know each other better and form closer relationships. While some groups have a tightly knit core and a lot of activities, while others fail to unite their members, or at least introduce them to each other.

Calculating group cohesion using social network analysis approach

We translate every group into a structural social network by identifying two key elements of SNA – nodes and links (or ties). Every blogger who belongs to a group is a node. If blogger b_i has a reciprocal friendship with a blogger b_j, then a link L (b_i b_j) exists. Combination of nodes and links form a social graph of a group. Every network is unique and has a different structure. To be able to describe the qualities and behaviour of a network, SNA-scientist use structural analysis. [12, p. 205-15]

To measure a group cohesion, we use a measurement of a network called **clustering coefficient**. Clustering – also known as transitivity in sociology [16] - is very important for describing the "texture of a population" [17] of a group. Clustering coefficient measures the degree to which people in a network tend to cluster together, i.e. gather into tightly knit groups with a relatively high density of ties. If a friend of a friend knows a friend of a friend, and so on, it means that those people form a cluster with high density of links [18]. Mathematically, a clustering coefficient is counted as "the average fraction of pairs of neighbors of a node that are also neighbors of each other" [18], therefore, the bigger is the largest cluster (group of friends) in a group, and the more clusters (groups of friends) there are in a group, the bigger is the clustering coefficient. And the larger the clustering coefficient of a group, the more well bound and friendly a group is.

Data collection

As previously discussed, the data is being collected on the largest Eastern European SNB Diary.ru. It is very helpful for the purposes of sampling that Diary.ru has a catalogue of the most popular groups (see Figure 1) which saves us the labour of discovering them among millions of existing blogs.

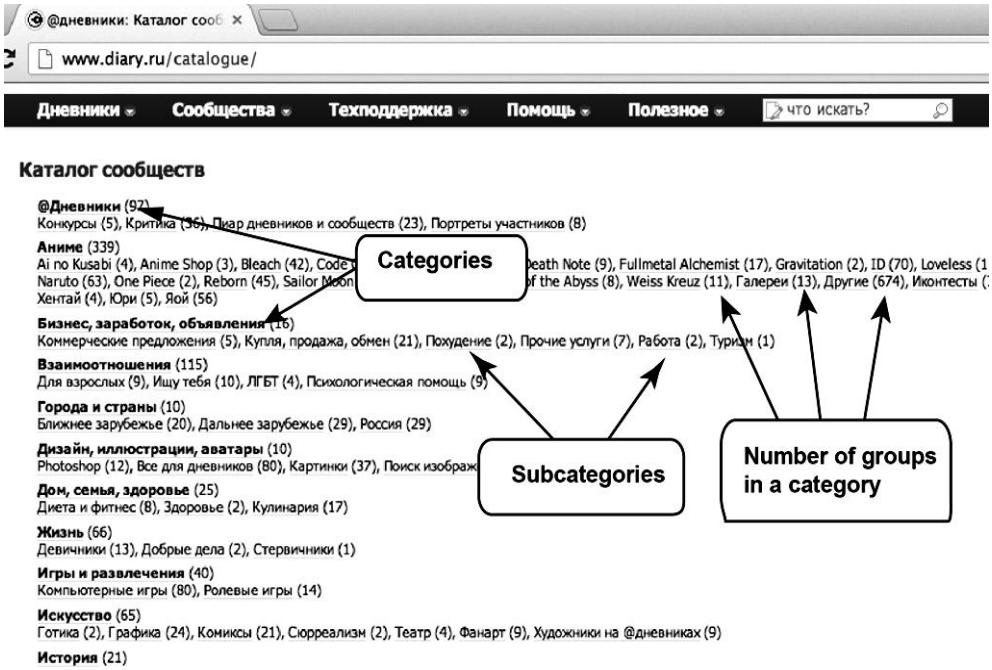

Fig. 1. A catalogue of most popular public groups on a SNB Diary.ru

Sampling

For a pre-study and sampling, we had to collect the information about all the groups in a catalogue. A special software – web crawler [19] – was designed and coded for this purpose using programming language Python. The program started its data-mining journey from the main page of the catalogue. It opened all the categories and sub-categories and collected the links to every group mentioned under each category. Then the program processed each link by connecting to the group, opening its main page and its profile, and collecting the main data about the group: name, ID, date of last update and URL. Finally, the program extracted a list of members and a list of readers (those who receive updates from the group but are not members and do not participate in group activities) from the profile of a group. Then the program compared both lists and selected only those who are members and readers at the same time, thus receiving the number of active members of the group. All this data was stored in data file for further processing.

When this data was collected (January 2016), the catalogue contained 1620 groups. A number of filters were applied to the retrieved list of groups, and some types of the groups were removed from the working list of groups. These types are:

1. Inactive: groups that have less than 50 posts (the number is very small comparing to an average group)
2. Abandoned: groups that have not been updated for a long time (more than 100 days prior to data retrieval) or had less than 10 posts in 2012.

3. Groups with a really small number of members (1-5), as all of them appeared to be a blog with multiple (but fixed) authors, not a proper group.

After filtering, 1220 groups remained in the short list.

If we sort a list of group by the number of active members, from largest to smallest, and visualize it, the bar codes which show a number of members will form a power law curve [12, p. 195] (see Figure 2) .

Since both visual analysis and statistical testing confirmed that the distribution is power law, we will use appropriate sampling methodology: sampling equal amount of groups from both sides of the group with the median number of members, in a sorted list of groups [20]. The median number of active members was – 224. Therefore, a group with 224 active members was the centre of the sample. 60 groups were sampled round this number: 30 groups that have less than 224 active members and 30 groups which have 224 members and more (see Figure 2). Overall, we sampled 60 groups for a quick preliminary study.

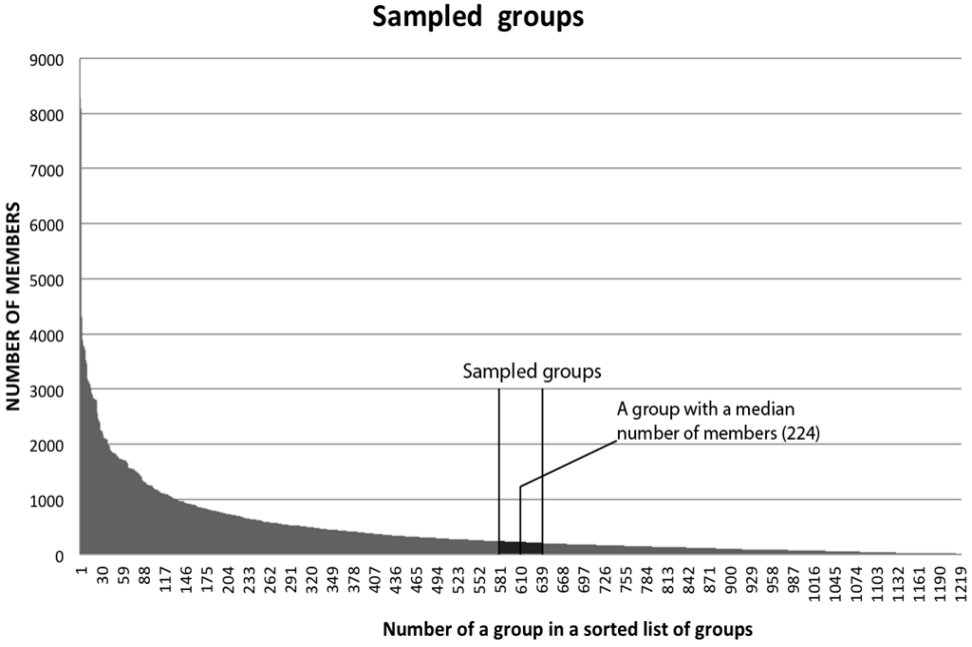

Fig. 2. Group sampling

Data modelling

As discussed above, to measure a group cohesion, we need to convert every group into a network and then calculate its clustering coefficient [18]. Clustering coefficient is the index of the inter-member bonding.

To collect the data about structure of a group, a special software – blog-spider - was designed and coded using programming language Python. For every group in the sample, the spider studied the group profile and extracted IDs and nicknames of active member of a group, forming a working list of members. After that, for every active member in the list, the spider studied their profiles and extracted information about their reciprocal friends. If the spider found out that one member of a group is a reciprocal friend with another member of the same group, the information about their connection was recorded into a database. It is important to note that, when recording the information, the spider used only codes of the members, not their actual nicknames of URLs, therefore the data stored in the database is completely anonymised, and the privacy of group members was fully protected.

Results and future work

In total, the information about 12 897 members and 18 088 inter-member connections were recorded for 60 group. After the structural information was collected, the program transformed the data for every group into a network and calculated average clustering coefficient, transitivity, network density and the proportion of the largest connected component using NetworkX – a special Python library for studying networks and graphs.

Statistical data of the research output is presented in Table 1.

Table 1. Social network analysis characteristics of sampled groups

Network quality	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
Transitivity	0.02970	0.08333	0.12260	0.12740	0.15700	0.30690	0.06072
Avg. clustering	0.00256	0.00827	0.01599	0.01804	0.02371	0.05941	0.01236
Density	0.00018	0.00059	0.00079	0.00097	0.00112	0.00319	0.00062
Largest component	5.16%	16.35%	20.74%	19.37%	22.98%	29.07%	5.38%

The collected information will be used in the next rounds of our research, to measure the link between a number of group markers (such as centrality, topic, strictness of moderation, etc.) and a group cohesion. As a result, we plan to obtain a blueprint to designing a SNB group that facilitates a high level of inter-member bonding and influence.

References

1. *Nardi B. A. et al.* Blogging as Social Activity, or, Would You Let 900 Million People Read Your Diary? New York, New York, USA: ACM Press, 2004. Pp. 222–23110 pp.
2. *Karpf D., Karpf D.* Understanding blogspace. Taylor & Francis, 2008. Vol. 5, № 4. Pp. 369–38517 pp.
3. *Godwin M., Godwin M.* Nine Principles for Making Virtual Communities Work. 1994. Vol. 6, № 2. Pp. 12–143 pp.
4. *Lenhart A. et al.* Bloggers: A portrait of the Internet's new storytellers. PEW Internet & American Life Project, 2006.
5. *Donath J. et al.* Public Displays of Connection // Social Network Analysis and Mining. Kluwer Academic Publishers, 2004. Vol. 22, № 4. Pp. 71–8212 pp.
6. *Merry S. K. et al.* Living and lurking on LiveJournal: The benefits of active and non-active membership // Aslib Proceedings. Emerald Group Publishing Limited, 2012. Vol. 64, № 3. Pp. 241–26121 pp.
7. *Efimova L. et al.* Finding"the life between buildings": An approach for defining a weblog community // Internet Research. 2005. Vol. 6, № 1997. Pp. 1–1515 pp.
8. *Soukup C., Soukup C.* Computer-mediated communication as a virtual third place: building Oldenburg's great good places on the world wide web // New Media & Society. 2006. Vol. 8, № 3. Pp. 421–44020 pp.
9. *Kendall L. S., Kendall L. S.* Hanging out in the virtual pub: identity, Masculinities, and relationships online. University of California, Davis, 1998.
10. *Rheingold H., Rheingold H.* The virtual community: Homesteading on the electronic frontier. MIT press, 2000. № 28.
11. *Ducheneaut N. et al.* Virtual "Third Places": A Case Study of Sociability in Massively Multiplayer Games // Comput. Supported Coop. Work. Kluwer Academic Publishers, 2007. Vol. 16, № 1-2. Pp. 129–16638 pp.
12. *Rheingold H. et al.* Net Smart: How to Thrive Online. MIT Press (MA), 2012.
13. *Scott J., Scott J.* Social network analysis. Sage Publications, 1988.
14. *Knoke D. et al.* Social network analysis. Sage Publications Los Angeles, CA, 2008. Vol. 2.
15. *Jamali M. et al.* Different Aspects of Social Network Analysis. IEEE, 2006. Pp. 66–727 pp.
16. *Jin E. et al.* Structure of growing social networks // Physical Review E. 2001. Vol. 64, № 4. Pp. 64–707 pp.
17. *Hanneman R. et al.* Introduction to Social Network Methods. Riverside, CA: University of California, Riverside, 2005.
18. *Wang X. F., Chen G.* Complex networks: small-world, scale-free and beyond. IEEE, 2003. Vol. 3, № 1. Pp. 6–2015 pp.
19. *Thelwall M., Thelwall M.* A web crawler design for data mining // Journal of Information Science. Sage Publications, 2001. Vol. 27, № 5. Pp. 319–3257 pp.
20. *Mirkin B. G., Mirkin B.G.* Core Concepts in Data Analysis: Summarization, Correlation and Visualization. Springer London, 2011.