Incorporating Text Analysis into Evolution of Social Groups in Blogosphere

Bogdan Gliwa, Anna Zygmunt, Stanisław Podgórski
Department of Computer Science
AGH University of Science and Technology
30-059 Kraków, Poland
{bgliwa,azygmunt}@agh.edu.pl, pstanisl@student.agh.edu.pl

Abstract—**Data reflecting social and business relations has often form of network of connections between entities (called social network). In such network important and influential users can be identified as well as groups of strongly connected users. Finding such groups and observing their evolution becomes an increasingly important research problem. One of the significant problems is to develop method incorporating not only information about connections between entities but also information obtained from text written by the users. Method presented in this paper combine social network analysis and text mining in order to understand groups evolution.**

I. INTRODUCTION

NOWADAYS, elements of our everyday life move increasingly to the virtual reality: we write blogs or comment on someone else's posts, participate in discussions on forums, we exchange our opinions on fanpages of telecommunications companies and banks whose products we use. Everywhere we leave traces of our activity, which can be analyzed and ably combined with each other. Trading companies and banks might be interested in finding active or influential people in their environment and to offer them a new product in hope that it will be proposed by them to many others. Identification on time disgruntled people on banks or telecommunications companies fanpages will allow to respond quickly and prevent the spread of discontent.

The data about different types of dependencies can be modeled as a network of relationships and its structure can be analyzed using Social Network Analysis methods (e.g. for finding important nodes). Such a network, however, is not homogeneous and one can distinguish groups of people, for example, who more often exchange opinions. Such groups frequently are formed around important individuals and, for various reasons, the groups continue to exist or not, grow, shrink or can be joined with other groups. To understand causes of such events, which significantly affect the behavior of groups, it is important to include information that we can extract from the content of opinions or comments that are left behind. If the group talk about the same topics, does it affect its longer duration? Or, perhaps has a variety of discussed topics stronger impact on the duration of groups? How the themes of discussion are changing in a group? Is a small group with a strong leader more durable than a large one with few strong individuals?

Such knowledge derived from open sources can be combined, for example, with information about the history of bank transfers or loans as well as data about phone calls. Methods and algorithms proposed in the paper have been tested on one of the largest and highly dynamic polish blogosphere: salon24.pl, in which the main topic of discussion are political issues. However, they can be applied to other social media such as, for example, Twitter.

II. RELATED WORK

A. Blogosphere and Social network analysis

Internet social media (e.g. blogs, forums, media sharing systems, microblogging, social networking, wikis) has revolutionized the Internet and the way of communication between people. Among them, blogs play a special role in creating opinions and information propagation. Author gives opinions on some themes or describes interesting events and readers comment on these posts. Posts can be categorized by tags. A very important element of blogs is the possibility of adding comments, which allows discussions. Basic interactions between bloggers are writing comments in relation to posts or other comments. The relationships between bloggers are very dynamic and temporal: lifetime of a post is very short [1].

Based on blogs, posts and comments, we can build network, which can be analysed by Social Network Analysis (SNA) methods [2]. The SNA approach provides measures (SNA centrality measures) which make it possible to determine the most important or influential nodes (bloggers) in the network. Around such bloggers, the groups are forming, sharing similar interests.

B. Groups in social networks

Groups (or communities) are sets of nodes that are relatively densely connected to each other but sparsely connected to other nodes in network [3]. Many methods of finding groups exist in literature - one of the most popular ones is the CPM method (Clique Percolation Method) [4], which allows to extract overlapping groups i.e. groups that can have shared nodes with other groups.

Considering the dynamic nature of various social media, a growing interest in developing algorithms for extracting communities that take into account the dynamic aspect of the network has been observed.

A method of tracking groups over time was proposed in [5]. First, a division into time steps is carried out. At each step, the graph is created and groups are extracted. Groups from consecutive time steps are matched using the Jaccard index (value of this measure above predefined threshold means a continuation for analysed group). Palla et al. in [6] identified basic events that may occur in the life cycle of the group: growth, merging, birth, construction, splitting and death.

For further analysis, different characteristics, describing the communities and their transformation in time [7], are calculated, which concerns the comparison of the strength of internal relations of group members with their external connections with nodes outside the group, density of connections in the group or stability of the membership in time.

C. Methods of text mining

Text classification is one of major goals of Text Mining [8]. It involves extracting similar documents, inferring text topic and searching documents based on topic criteria.

Most text mining methods focus on text preprocessing (eg. stop words removal, words stemming and lemmatization) and converting input into structural representation [9]. Each word is represented as a separate entity assigned with a weight of the word importance, and thus it also allows to easily extract keywords. Algorithm TF-IDF (Term Frequency - Inverted Document Frequency) is one of the most popular weighting method [9]. It is based on the assumption that the importance of a word is proportional to number of occurrences of this word in a document, and inversely proportional to number of documents in which the word occurred. However, using only keywords to classify texts fails to find connection between semantically convergent documents that utilize different vocabularies, and more complex methods need to be applied such as Topic Modeling [10].

Topic Modeling [11] is a statistical technique that uncovers abstract "topics" that occur in a collection of documents. "Topic" is a set of words that tend to co-occur in multiple documents, and, therefore, they are assumed to have similar semantics. Main benefit of this model is that instead of using words from pattern to search for similar documents, words from topic are used, and therefore similar texts can be discovered even if they use different vocabulary.

Entirely different approach to uncovering documents semantics involves human input and it is called tagging [12]. Tags can be assigned either by author or by community in a process called crowdsourcing [13]. Number of tags assigned to a document may be large, and, therefore, it is imperative that a proper grouping and selection mechanism is implemented.

D. Text mining in the context of social network analysis

Existing research utilizing both SNA and Text Mining are mainly focused on very narrow cases. Aggarwal and Wang in [14] provided broad overview of text mining methods useful for social networks analysis. Tuulos and Tirri in [15] analysed IRC (Internet Relay Chat) communication network to discover and verify chat channels topics. Agrawal et al.

in [16] used text mining methods to split social group into protagonists and antagonists. Caverlee and Webb in [17] used automatic classification methods based on keywords extraction and Topic Modeling to confirm personal information provided by Myspace users.

III. ANALYSIS TOPICS OF GROUPS AND THEIR IMPACT ON GROUP BEHAVIOUR

In this section we provide the concept of methods used to further analysis. The social network from whole data range is divided into series of time slots and each time slot contains static snapshot of network from defined period of time. In every time slot we extract groups and then find their dynamics in time. Irrespectively, we also discover topics in texts of comments and posts. Afterwards, we try to match topics for groups based on topics of comments and posts written by members of groups between themselves. Next, we analyse relations between topics of groups and behaviour of groups.

A. Groups in dynamic social network

Groups in each time slot were detected using the CPM [4] method (directed version of CPM from CFinder [1]). Groups from neighbouring time slots can be matched in order to find continuation of groups from different time. For this purpose, the SGCI (Stable Group Changes Identification) [18] method was employed. The algorihm consists of four main steps: identification of short-lived groups in each separated time slot; identification of group continuation (using modified Jaccard measure), separation of the stable groups (lasting for a certain time interval) and the identification of types of group changes (transition between the states of the stable group). The SGCI method identifies following event types:

- **split**, occurs when group divides into several groups in next time slot,
- **deletion**, similar to split, but it happens when small group detaches from significantly bigger one,
- **merge**, when several groups in the previous time slot join together and create larger group,
- **addition**, similar to merge, but it takes place when small group attaches to significantly bigger group,
- **split_merge**, when for the predecessor group the event is split and for the succesor group of given transition the event is merge in the same time,
- **decay**, the total disintegration of the group - the group does not exist in the next time slot,
- **constancy** means simple transition without significant change of the group size,
- **change_size** - simple transition with the change of the group size.

More detailed description of this method is provided in [18].

B. Finding topics of groups

For texts of posts and comments we employed methods of text mining in order to discover topics. Topics were extracted using 3 different methods:

---

[1]http://www.cfinder.org/

- TF-IDF keywords - words with the highest TF-IDF scores,
- Topic Modeling - topics extracted with LDA algorithm,
- Tags provided by post authors.

Keywords set for Topic Model is assumed to be a set of the most significant words for topics inferenced for messages.

We compared these methods between themselves using *similarity* measure:

$$similarity(S_1, S_2) = \frac{|S_1 \cup S_2|}{\min(|S_1|, |S_2|)}$$

where: S – keywords set, $|S|$ – number of elements in S.

For each group we can also assign set of topics discussed by its members. The topics are inferred based on posts and comments written by members of groups. We focused mostly on topic modelling as this method provides the highest level of abstraction from presented methods. Only topics that were present in more than 5% messages for groups were taken into consideration.

We defined *topic exploitation* for given topic and group as a ratio between number of group messages on certain topic and all messages for this group:

$$topicExploitation_k = \frac{|T_k|}{\sum\limits_{i=1}^{n} |T_i|}$$

where: T_k – set of messages (posts and comments) for which topic with number k was inferenced, n – number of all topics, $|T_x|$ – amount of elements in T_x.

C. Topics changes in groups

To describe topics changes during transition between groups, we introduced following metrics:

- *Change in topic exploitation* for m-th group after transition t from time slot n to $n+1$ is calculated as:

$$c_{m,n,t} = \sum_i \sum_k [g_{m,n,i} - g_{k,n+1,i} \cdot f(m, n, k, t)]$$

where: i is a number of topic, $g_{m,n,i}$ is the topic exploitation of i-th Topic for m-th group in n-th time slot, f is function returning 1 if k-th group in slot $n+1$ is a continuation of m-th group from slot n and this transition has event type t.

- *Maximal positive change of single topic* (how much a topic gained) for m-th group after transition t from time slot n to $n+1$ is defined as:

$$mpc_{m,n,t} = \max_i \sum_k [\epsilon(g_{m,n,i} - g_{k,n+1,i} \cdot f(m, n, k, t))]$$

where: ϵ is a function returning the argument when the argument is **negative**, otherwise 0; other symbols were explained for *Change in topic exploitation* measure.

- *Maximal negative change of single topic* (how much a topic lost) for m-th group after transition from time slot n to $n+1$ was calculated as:

$$mnc_{m,n,t} = \max_i \sum_k [\theta(g_{m,n,i} - g_{k,n+1,i} \cdot f(m, n, k, t))]$$

where: θ is a function returning the argument when the argument is **positive**, otherwise 0; other symbols were explained for *Change in topic exploitation* measure.

Using above metrics we can analyse influence of different evolution types on topics change. Therefore, for each evolution type the average values of above defined measures for all groups are evaluated and we refer to them as *Average overall change in topic exploitation*, *Average maximal positive change of single topic* and *Average maximal negative change of single topic* respectively.

For above metrics, evolution events were taken into consideration only if there were at least 10 such events in selected time period.

D. Migrations of users depending on topics

To analyse difference in topics between given user and given group, we defined *topic divergence*, which has the following form:

$$m_t = t_{group} - t_{user} = \sum_{i=1}^{n} |(topic_{i,user} - topic_{i,group})|$$

where: n is a number of all topics in model (350), t_{group} is set of weights of each topic for given group, $topic_{i,group}$ – weight of i-th topic for given group, t_{user} is set of weights of each topic for given user, $topic_{i,user}$ is weight of i-th topic for given user.

It's worth noting that minimal value of m_t is 0.0 when user and a group has identical weight for every topic and maximal value is 2.0 when they are totally different. Maximum value of 2.0 is connected with the fact that group might cover topic X in 100% and user might cover topic Y in 100%, and therefore difference between group and user on topic X is 100% and on topic Y is also 100% which adds up to 200%.

Using this measure, we are trying to investigate relations between *topic divergence* and migrations of users (leaving and joining to groups). For this purpose the following measures are utilized:

- *Probability of leaving the group*. We assumed that potentially any member can leave the group. This value is calculated as:

$$P_l(m) = \frac{|leavers_m \cap candidates_m|}{|candidates_m|}$$

where: $leavers_m$ are users that in fact left any group and had the value of *topic divergence* measure equals m; $candidates_m$ are members of groups that have *topic divergence* = m.

- *Probability of joining the group*. When considering topic measure we assumed that candidates for joining are all users that were active in previous time slot. This value is calculated as:

$$P_j(m) = \frac{|joiners_m \cap candidates_m|}{|candidates_m|}$$

where: $joiners_m$ are users that in fact joined any group and had the value of *topic divergence* measure equals m;

$candidates_m$ – users active in previous time slot with topic divergence = m.

While calculating joiners and leavers sets we considered all group continuations to be a single group. The reason for that is to prevent *deletion* event to distort results - if a group splits into multiple small groups and we are assuming that anyone from the group can leave, then we will get very high accuracy from each event when huge group changes into small group.

It is worth noting that only both values - probability and histogram with migrations can provide us with complete information. Probability alone strongly depends on test case - if only 1 user had measure value = X and this user migrates then probability of migration for measure=X will be 100%, even if 100 different users migrated but they all had measure value=Y just as rest 10000 users, and thus probability for measure value=Y will be 1%. Without histogram we could not tell if any of those cases are marginal.

Analogically histogram itself can tell us only for which value there are the most migrations.

IV. Description of experiments

A. Data set

The analysed data about blogs was retrieved from the portal www.salon24.pl, which is dedicated especially to political discussions, but also subjects from other domains may be brought up. The data consists of 26 722 users (11 084 of them have their own blog), 285 532 posts and 4 173 457 comments within the period 1.01.2008 - 31.03.2012. Presented results were conducted on whole dataset - from 1.01.2008 to 31.03.2012. The analyzed period was divided into time slots, each lasting 7 days and neighboring slots overlap each other by 1 days. In the examined period there are 259 time slots. In each slot we used the comments model, introduced in [19] - the users are nodes and relations between them are built in the following way: from user who wrote the comment to the user who was commented on or if the user whose comment was commented on is not explicitly referenced in the comment (by using @ and name of author of comment) the target of the relation is the author of post.

B. Number of groups

The number of communities, with given size, for different value of k (parameter for CPM algorithm) is presented in table I. The k parameter determines the minimum group size (e.g. k equals 3 means that groups should consist of 3 or more members). The larger value of k, the smaller size of the biggest group. As we can notice, small groups outnumber other ones for each k. Furthermore, for k equals 6 the quantity of groups is much lower than for other values of k parameter.

C. Evolution events

Table II contains number of different evolution events in dataset for different values of k. We can observe for k equal 4 or 5 that the most popular events are *addition* and *deletion*, but for k equal 6, the most frequent events are *merge* and *split* (events similar to *addition* and *deletion*). The reason is that

TABLE I
NUMBERS OF GROUPS WITH DEFINED SIZE.

size	k=4	k=5	k=6
< 5	1596	0	0
5 − 6	384	2372	0
6 − 7	207	632	584
7 − 8	113	255	149
8 − 9	88	139	86
9 − 10	50	63	39
10 − 50	289	332	199
50 − 100	25	54	30
100 − 200	59	96	6
> 200	172	17	0

for k equal 4 or 5, there is a lot of small groups and there are also very huge groups (which not happens for k equal 6). In further analysis, we focus on groups extracted for parameter k equal 5 from the CPM method.

TABLE II
NUMBERS OF EVOLUTION EVENTS.

type	k=4	k=5	k=6
change_size	699	470	195
constancy	257	100	42
merge	428	439	434
split	323	409	397
addition	1091	2070	197
deletion	1115	2040	188

D. Convergence of different message topic extraction methods

This experiment covered comparison of different messages topic inference methods: TF-IDF keywords, topic modelling and tags provided by users.
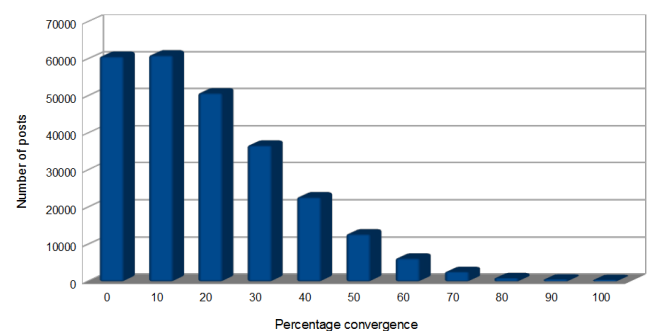

Fig. 1.   Convergence level of TF-IDF keywords and most significant Topic Model words for inferenced topics

Fig. 1 presents comparison between TF-IDF keywords and Topic Modelling. As it can be seen, for about 20% of documents there was not even a single matching word. Convergence rate above 50% was achieved by merely 6000 posts which is around 5% of all input data.
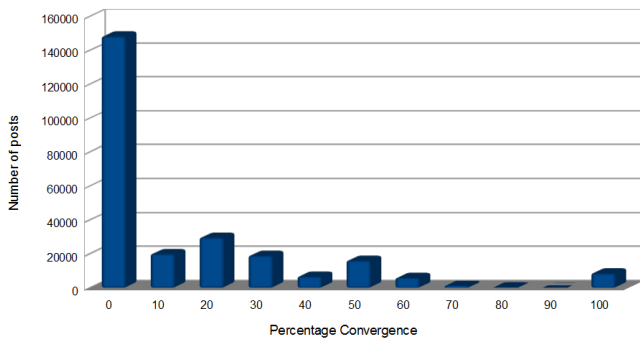
Fig. 2.   Convergence level of TF-IDF keywords and user provided Tags

In fig. 2 TF-IDF keywords are compared with user tags. One can see that for huge part of documents achieved convergence rate was 0%. There are also local maxima at 20, 30, 50 and 100% and their origin is connected with number of tags user provides, which in most cases is between 1 and 5 (when matched 1/5, 1/4, 1/3, 1/2 and 1/1 of keywords).
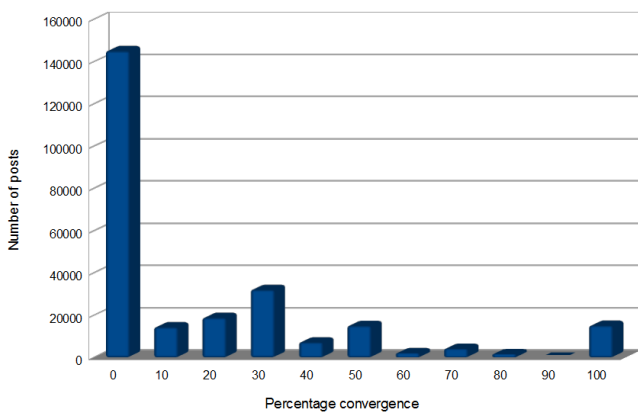

Fig. 3.   Convergence level of user provided Tags and most significant Topic Model words for inferenced topics

Fig. 3 displays user tags in comparison with topic modelling. This histogram shows that, similarly to fig. 2, very huge part of tested posts don't have even a single word in common. We can also notice similar maxima due to the same reasons.

According to presented data, convergence rate seems to be very low, and therefore it would be imperative to check which of presented methods gives correct results. However, in depth analysis of our results uncovered that in fact all three methods are properly describing posts semantic. The problem is with vocabulary that is being used to describe it. The reason for different vocabulary is connected with level of generalisation that is utilised by presented methods:

- keywords provides very specific and detailed description of given document,
- tags provides general summary of the document and are far less specific, but error-prone of misspelling (they are provided by post authors),

- topics provides very general and abstract idea behind the document.

For these reasons, we are focusing more on topic modelling.
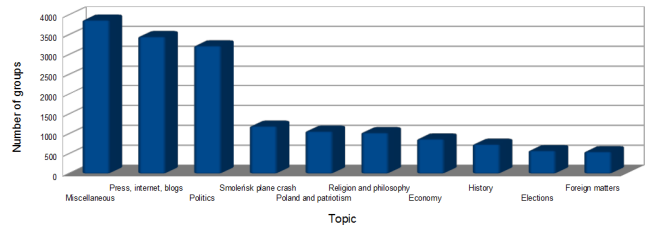
E. Topics coverage by groups

Fig. 4.   Number of groups exploiting given topic

Figure 4 shows number of groups that exploit given topic in more than 5% of total group members messages. This figure presents ten most popular topics. We can notice that most popular topic is *Miscellaneous* which is very general and in fact is a mix of many themes.

F. Influence of group size on covered topics

In this experiment we aimed to check if there is a correlation between groups size and topics this groups covers.
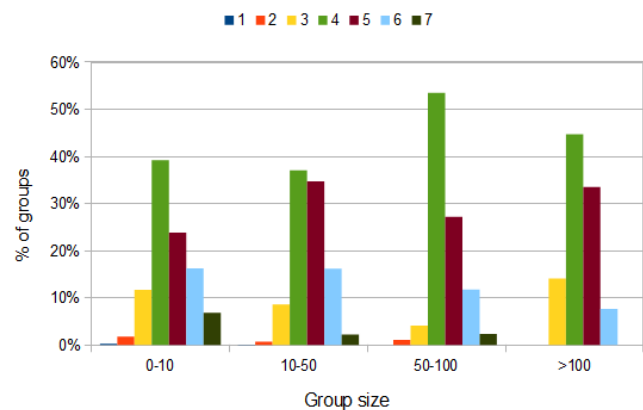

Fig. 5.   Number of topics covered by groups with certain size, k=5

Fig. 5 shows number of Topics covered by groups with certain sizes. Considering 5% threshold of topic importance we used to remove noise. Maximum value any group could achieve was twenty topics, however no group covered more than seven topics and most groups cover four or five topics. In fig. 5 we can observe that three to seven topics are in groups of any size, but there are some small and medium-size groups that discuss only about one or two topics, which not happens in large groups.

Fig. 6 presents that for some topics *topic exploitation* is very similar regardless of group size (e.g. *Press, internet, blogging* topic), but there are some specific topics that are discussed to a greater extent among members in small group (e.g. topic related with science) or among members in larger ones (e.g. politics).
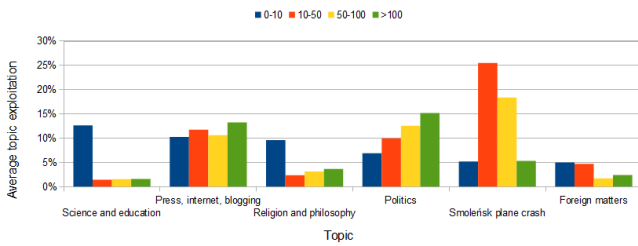
Fig. 6.   Percentage topics exploitation for groups with certain size, k=5

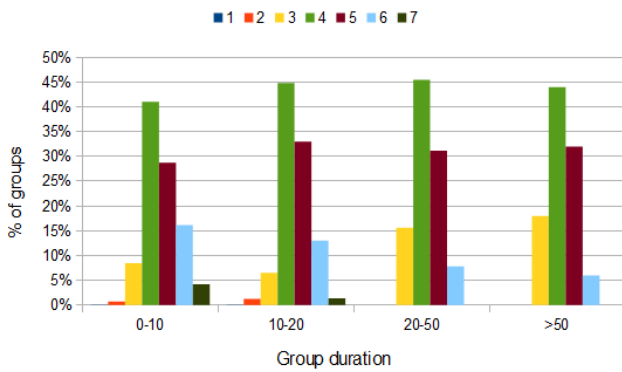*G. Influence of duration time on covered topics*

Fig. 7.   Number of topics covered by groups with duration time, k=5

We can notice some regularities on fig. 7 :

- only very short living groups covers one and two topics,
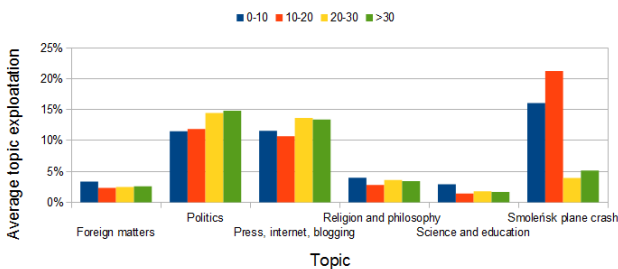- as previously, most groups cover four and five topics.

Fig. 8.   Percentage topics exploitation for groups with certain duration time, k=5

In fig. 8 one can see that most topics achieved similar scores, with an exception - Polish Air Force Tu-154 crash was, to a large degree, discussed by short-term groups.

H. Connection between topics change and groups evolution

This experiment aimed to verify if there is any apparent connection between group evolution type and change in topics coverage for this group. Experiment was based on three measures (described earlier in section III-C):

- average overall change in topics coverage after evolution,

- average maximal positive change of single topic,
- average maximal negative change of single topic.

Evolution events were taken into consideration only if there were at least ten such events in selected time period. Presented results were collected for groups with $k = 5$ and for periods of length 360 days. There are two evolution types that are not present on the chart - *split_merge* that did not occur and *decay* that was omitted. *Decay* event means that group ceased to exist, and therefore we cannot calculate how topics of this group changed, because there is no continuation of the group.

Average overall change in topics. Figure 9 presents some regularities:

- *addition* event has clearly the highest overall topic change,
- *split* and *change_size* are connected with the lowest topic change,
- *merge* and *deletion* are in between,
- quite surprisingly *constancy* seems to vary between periods, even though one could expect that it will be connected with very small topic change.
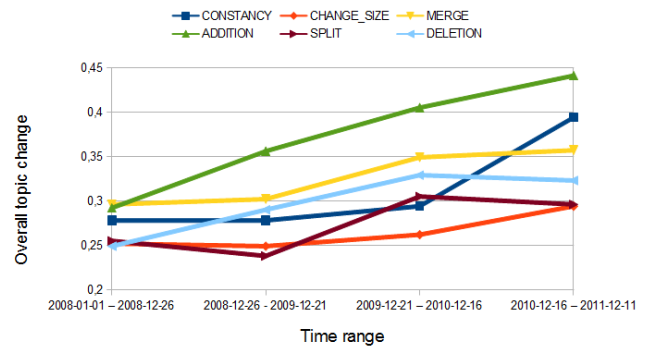

Fig. 9.   Average overall change in topics coverage for every evolution event type - period 360 days

Average maximal positive change of single topic. Figure 10 presents that:

- *addition* has lowest average maximum single topic change. It means that on average after *addition* even topic that gained the most, gained very little.
- *deletion* and *split* caused highest positive change.
- *merge* and change_size were in between.

Average maximal negative change of single topic. Figure 11 shows that:

- *addition* is connected with highest drop for a single topic. It means that after *addition* there is a topic that significantly loses popularity.
- *change_size*, *split*, *deletion* and *merge* has very low average drop in topic popularity - even topics that lose popularity after such events lose very little.

Summary for different types of topic change. *Addition* is connected with: highest overall change in topics, highest negative change and lowest positive change. Therefore, we can deduce that when multiple small groups are forming a single
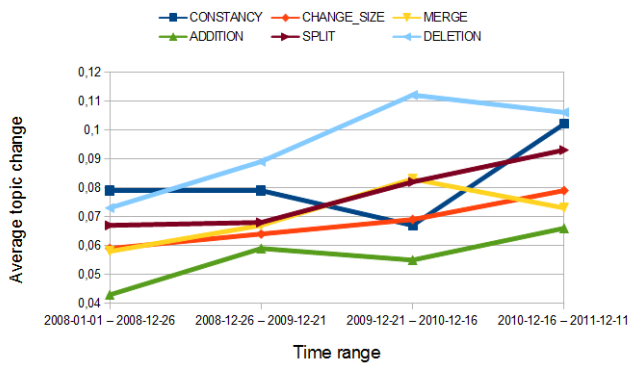
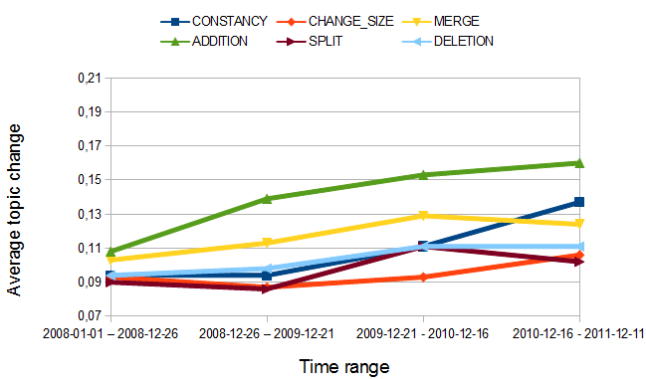Fig. 10. Average maximal positive change of single topic - period 360 days

Fig. 11. Average maximal negative change of single topic - period 360 days

large group it is usually connected with significant drop of popularity of the main topic of each of the groups, and small rise in popularity of different topics - presumably of main topics of other groups.

Deletion and *split* cause small overall topic change, small negative change and large positive change. It means that splitting a group causes a rise of popularity of a single topic at the expense of all the others.

Change_size has very small changes in topics. We could expect that *constancy* behave the same way, however it does not.

Merge event causes medium rise of single topic popularity at the expense of all the others, meaning that joining groups are very similar and after join the leading topic emerges.

I. Migrations between groups

This experiment was conducted to check if migration of users between groups could be predicted using information about topic preferences of users and groups.

Important: currently analysed user does not have to be a member of the group. In fact, when calculating probability of joining a group we consider only users that are not yet members.

As a measure connected with users and groups topics we used *topic divergence* between user and group (details in section III-D).

Based on this measure, we tested their influence on:
- probability of leaving the group (candidates for leaving group are all members of group),
- probability of joining the group (candidates for joining are all users that were active in the previous time slot).

While calculating joiners and leavers sets we considered all group continuations to be a single group. The reason for that is to prevent *deletion* event to distort results - if a group splits into multiple small groups and we are assuming that anyone from the group can leave, then we will get very high accuracy from each event when huge group changes into many small groups.

Joining groups. Figure 12 shows that there is high probability of joining for users that are high convergent with the group, however, when we look at figure 13, we can notice that some of them (most convergent users with groups) are marginal cases (very few migrations). Moreover, we can notice that probability of joining groups is rather constant regardless the value of *topic divergence*, except the smallest values of this measure.
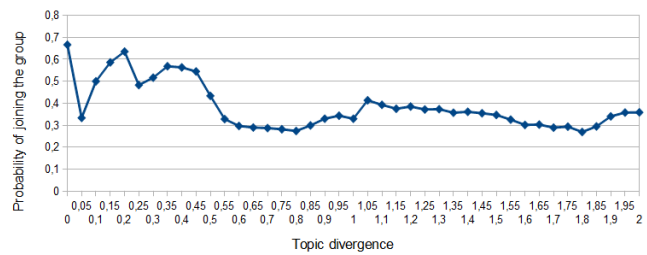

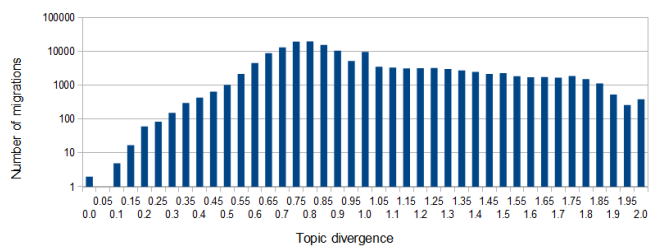Fig. 12. Probability of user joining a group based on topic divergence.

Fig. 13. Number of users that did join the group based on topic divergence.

Most migrations occur between 50% and 100% divergence and we can also see a rise in probability there, due to more cases. There is an interesting local extreme in figure 13 around 100% divergence. It would seem that there is a large portion of joiners that in previous time slot wrote all their posts and comments on one topic, or more likely wrote only a single post or comment, and then joined the group where this topic was not relevant. It is worth noting that our candidates were only a fraction of real joiners. It seems that about half joiners were people that were inactive in previous time slot (151 819 joiners were inactive user and 154 977 real joiners were from the candidate set).

Leaving groups. As can be deduced from figure 14, for low divergence values probability of leaving is low and is rising along with divergence up until 50% divergence and, further, the probability of leaving is rather constant. Alhough, it drops down for very high divergence, however, figure 15 tells us that it is a marginal case.
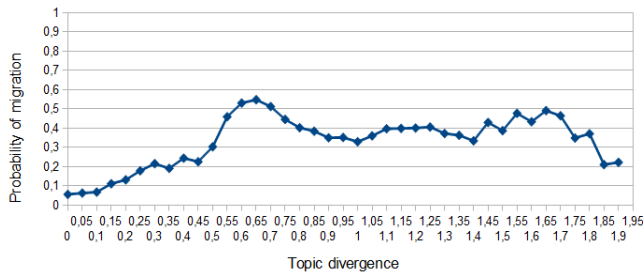

Fig. 14.   Probability of user leaving a group based on topic divergence.
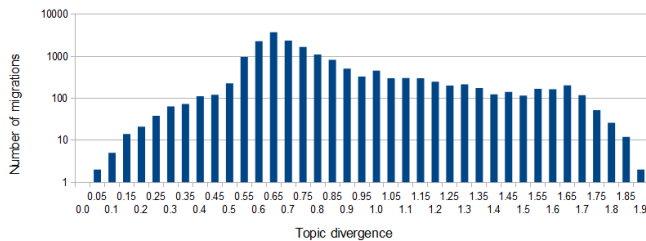

Fig. 15.   Number of users that did leave the group based on topic divergence.

V. Conclusion

In this paper the analysis of topics for communities detected in real-world data from blogosphere is presented. We conducted experiments concerning relations between discussed topics by members of groups and some aspects of groups such as their duration time or their size. Furthermore, we also analysed influence of group evolution events on changes of topics and investigated the impact of topic divergence on users behaviour such as joining or leaving group.

Presented results seem promising and they reveal new insights into behaviour of groups and individuals. Analysis of topics discussed inside communities can be useful tool enabling better understanding of processes inside social network.

In future we are planning to use information about topics to improve our method of prediction of group behaviour [20]. Moreover, we intend to carry out similar experiments (related to analysis of topics in communities) on other datasets - we want to use also blogs in English language and datasets from different kinds of social media e.g. microblogs. Another interesting direction of further research is the analysis of key persons in groups in terms of topics they discuss and such analysis could lead to enhance our method of defining user roles and finding most influential people [21].

References

[1] N. Agarwal and H. Liu, *Modeling and Data Mining in Blogosphere*. Moegan & Claypool Publishers, 2009.
[2] P. Carrington, J. Scott, and S. Wasserman, *Models and Methods in Social Network Analysis*. Cambridge University Press, 2005.
[3] S. Fortunato, "Community detection in graphs," in *Phys. Rep.*, 2010, ch. 486.
[4] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, pp. 814–818, 2005.
[5] D. Greene, D. Doyle, and P. Cunningham, "Tracking the evolution of communities in dynamic social networks," in *Proc. International Conference on Advances in Social Networks Analysis and Mining (ASONAM'10)*. IEEE, 2010.
[6] G. Palla, A. lászló Barabási, T. Vicsek, and B. Hungary, "Quantifying social group evolution," *Nature*, vol. 446, p. 2007, 2007.
[7] J. Xu, B. Marshall, S. Kaza, and H. Chen, "Analyzing and visualizing criminal network dynamics: A case study," in *IEEE Conference on Intelligence and Security Informatics*, Tuczon, 2004.
[8] C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in *Mining Text Data*, C. Aggarwal and C. Zhai, Eds. Springer, 2012, pp. 163–222.
[9] H. L. X. Hu, "Text analytics in social media," in *Mining Text Data*, C. C. Aggarwal and C. Zhai, Eds. Springer, 2012, pp. 385–414.
[10] S. Crain, K. Zhou, S. Yang, and H. Zha, "Dimensionality reduction and topic modelling: from latent semantic indexing to latent dirichlet allocation and beyond," in *Mining Text Data*, C. Aggarwal and C. Zhai, Eds. Springer, 2012, pp. 129–162.
[11] Y. Huang, "Support vector machines for text categorization based on latent semantic indexing," Electrical and Computer Engineering Department, The Johns Hopkins University, Tech. Rep., 2003.
[12] Y. Hassan-Montero and V. Herrero-Solana, "Improving tag-clouds as visual information retrieval interfaces," in *Proceedings of the InScit2006, International Conference on Multidisciplinary Information Sciences and Technologies*, 2006.
[13] C. M. R. McCreadie and I. Ounis, "Crowdsourcing a news query classification dataset," in *Proceedings of CSE'10*, Geneva, Switzerland, 2010.
[14] H. W. C. Aggarwal, "Text mining in social networks," in *Social Network Data Analytics*, C. Aggarwal, Ed. Springer, 2011, pp. 353–378.
[15] V. Tuulos and H. Tirri, "Combining topic models and social networks for chat data mining," in *WI '04 Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, Washington DC, USA, September 2004, pp. 206–213.
[16] R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu, "Mining newsgroups using networks arising from social behavior," in *Proceedings of the 12th international conference on World Wide Web*, Budapest, Hungary, May 2003, pp. 529–535.
[17] J. Caverlee and S. Webb, "A large-scale study of myspace: Observations and implications for online social networks," in *Proceedings of the 2nd International Conference on Weblogs and Social Media (AAAI)*, Seattle, USA, March 2008.
[18] B. Gliwa, S. Saganowski, A. Zygmunt, P. Bródka, P. Kazienko, and J. Kozlak, "Identification of group changes in blogosphere," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012 Istanbul, Turkey, 26-29 August 2012*. IEEE Computer Society, 2012.
[19] B. Gliwa, J. Kozlak, A. Zygmunt, and K. Cetnarowicz, "Models of social groups in blogosphere based on information about comment addressees and sentiments," in *Social Informatics - 4th Int. Conf., SocInfo, Lausanne, Switzerland*, ser. Lecture Notes in Computer Science, vol. 7710. Springer, 2012, pp. 475–488.
[20] B. Gliwa, P. Bródka, A. Zygmunt, S. Saganowski, P. Kazienko, and J. Kozlak, "Different approaches to community evolution prediction in blogosphere," in *ASONAM 2013: IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining: Niagara Falls, Turkey*, 2013, accepted for printing.
[21] B. Gliwa, A. Zygmunt, and J. Kozlak, "Analysis of roles and groups in blogosphere," in *CORES - 8th International Conference on Computer Recognition Systems*, ser. Advances in Intelligent and Soft Computing, vol. 226. Springer, 2013, pp. 299–308.