# Analysis of Content of Posts and Comments in Evolving Social Groups

**Bogdan Gliwa, Anna Zygmunt and Piotr Bober**

**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. Analyzing the evolution of communities is useful in many applications such as marketing, politics or public security domains. 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 chapter combine social network analysis and text mining in order to understand groups evolution. Presented approach to the group evolution process takes many aspects of the group analysis into consideration. Due to proposed method the subjects discussed within the groups are known. We noticed that subjects discussed within groups play significant roles in group evolutions.

## 1 Introduction

In today's world it's hard to imagine doing business without the use of virtual reality: many companies have moved there the whole business, others try to combine running traditional business with virtual. However, all companies are trying to analyze the behavior of their customers, watching their activity in social media. Since writing

B. Gliwa · A. Zygmunt (✉) · P. Bober
Department of Computer Science, AGH University of Science and Technology,
30-059 Kraków, Poland
e-mail: azygmunt@agh.edu.pl

B. Gliwa
e-mail: bgliwa@agh.edu.pl

P. Bober
e-mail: bober@student.agh.edu.pl

M. Mach-Król et al. (eds.), *Advances in ICT for Business, Industry and Public Sector*,
Studies in Computational Intelligence 579, DOI 10.1007/978-3-319-11328-9_3

blogs or comments on someone else's posts, participating in discussions on forums, exchanging our opinions on fanpages of for example telecommunications companies and banks whose products we use, we everywhere 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 or influential 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—besides the fact that someone is connected with someone—information that we can extract from the content of opinions or comments that are left behind (e.g. what they are talking about). If the group is talking about the same themes (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. To which of our client should be directed attractive offer? Certainly to the one who has a lot of friends, participates in many long-lasting discussions in different groups of people, and so ensures that our offer reaches a large group of potential customers. Proposed methods and algorithms 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, different blogs, microblogs (e.g., Twitter), opinion websites (e.g. Epinions).

The structure of the rest of chapter is as follows. In Sect. 2 the state of the art concerning the blogosphere and social network analysis, groups and their dynamics and text mining in the context of social network analysis is given. Section 3 describes the main idea and detailed solutions of the process of the group analysis with subjects identification. Dataset, description of experiments and obtained result are shown in the Sect. 4. Finally, the Sect. 5 concludes and presents plans of future works.

# 2 Related Work

## *2.1 Blogosphere and Social Network Analysis*

Internet social media such as online social networking (e.g., Facebook,[1] Myspace,[2]) blogging (e.g., HuffingtonPost,[3]) forums, media sharing systems (e.g., YouTube,[4] Flickr,[5]) microblogging (e.g., Twitter,[6]) wikis, social news (e.g., Digg,[7] Slashdot,[8]) social bookmarking (e.g., Delicious,[9]) Opinion, Review, and Ratings Websites (e.g., Epinions[10]) has revolutionized the Internet and the way of communication between people. Users stopped being only the consumers of information, becoming the creators of information. Among them, blogs play a special role in creating opinions and information propagation. According to [22], blog is a journal-like website for users, to contribute textual and multimedia content, arranged in reverse chronological order.

Blog can be treated as a kind of an Internet diary, where an author gives opinions on some themes or describes interesting events and readers comment on these posts. A typical entry on a blog can consist of text, photos, films and links to other blogs or web pages. Posts can be categorized by tags. A very important element of blogs is the possibility of adding comments, which enables discussions. Access to blogs is generally open, so everybody can read the posts and comments. Nowadays, many blogs are conducted by experts in the field, so they have become a reliable source of information.

Basic interactions between bloggers are writing comments in relation to posts or other comments. Generally, the relationships between bloggers change over time and are temporal: lifetime of a post is very short [1]. Of course, it depends strongly on the domain of discussion: the more controversial discussions, the more intensive but generally shorter.

Based on blogs, posts and comments, we can build a network where nodes are bloggers or posts and there is an edge between two nodes if one user writes the comment to the other user. Such network can be analysed by Social Network Analysis (SNA) methods [8]. 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.

[1] http://www.facebook.com.

[2] http://myspace.com.

[3] http://www.huffingtonpost.com.

[4] http://www.youtube.com.

[5] http://www.flickr.com.

[6] https://twitter.com.

[7] http://Digg.com.

[8] http://slashdot.org.

[9] http://delicious.com.

[10] http://www.epinions.com.

## 2.2 Groups and Their Dynamics

Social networks, built on the basis of the interactions in social media, consist of groups (communities). One can say that the groups are the key structural characteristics of networks [13]. They reflect the structure of relationships in social media: the more frequently users are discussing among themselves on a blog, the higher probability that they will form a common group. It is difficult to find one definition of a group, but in most cases a group is treated as densely interconnected nodes (but loosely to other nodes in network).

Many methods of finding groups have been proposed [12, 29]; discovered groups may be separate (node belong only to one group) or overlapping (node can be a member of many groups). In the case of blogs, more natural approach is using algorithms finding overlapping groups, because given user can participate in many discussions not necessarily with the same people. Popular representative of this class of algorithms is Clique Percolation Method (CPM) [26, 27] which is based on finding in a graph k-cliques, which means a complete, fully connected subgraph of k vertices, where every vertex can be reached directly from all other vertices. Groups are treated as sets of adjacent k-cliques (having k-1 mutual vertices).

Due to the dynamic nature of the 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 [21]. 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 and defined a threshold above which the continuation of the group is assumed. Palla et al. in [25] identified basic events that may occur in the life cycle of the group: growth, merging, birth, construction, splitting and death. Asur in [3] introduced formal definitions of five critical events. Gliwa et al. proposed in [14] two additional events and gave formal definitions. In [17, 18] new tool GEVi for context-based graphical analysis of social group dynamics was proposed.

For further analysis of the different characteristics describing the communities and their transformation in time [31] 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.

In [28] incremental community mining approach is introduced: instead of extracting the groups at each time slot independently, they introduce L-metric which allows measure the similarity between groups in the subsequent time slots directly.

Analyzing the evolution of communities is useful in many applications such as marketing, politics or public security domains.

### *2.3 Text Mining in Social Networks*

Aggarwal and Wang in [2] analysed a set of text mining methods in terms of usefulness in social networks analysis (SNA). Bartal et al. [4] proposed a method for link prediction in a social network with usage of text mining features. Velardi et al. [30] described content-based model for social network and edges in such a network were created based on similarity of text mining features between nodes.

One of the most important methods of text mining is Topic Modeling [23], a statistical technique that uncovers abstract "topics" in a collection of documents. Notion of "topic" can be defined as a set of words that co-occur in multiple documents, and, therefore, they are expected to have similar semantics. Two main branches of Topic Modeling [9] can be distinguished—algorithms based on Singular Value Decomposition such as Latent Semantic Indexing [23] and algorithms based on probabilistic generative process [5] such as Latent Dirichlet Allocation (LDA) [6, 7]. Topic modeling techniques aim to reduce dimensionality by grouping words with similar semantics together.

In [19, 20] the application of topic modeling to analysis of groups dynamics in social networks is presented and uncovered topics have manually assigned labels. Different approach is described in [24] where authors analyse topics in time and assignment of labels for them is conducted in an automatic way. Diesner et al. [11] used social network analysis methods to identify users in different roles assessing their attitude to change (i.e. roles: change agents and preservation agents) in the innovation diffusion network and topic modeling was utilized to describe their different characteristics. Cuadra et al. [10] presented new method of community detection in social networks based on similarity of users messages using topic modeling method.

## 3 General Concept

In this section we provide the concept of our method of analysis topics of groups and their impact on group behaviour. Analysis consists of the several steps. The introductory phase is gathering data from social media services (e.g. blogosphere) into repository (e.g. database). Collected data is divided into series of time slots, so each time slot contains static snapshot of the state of network from defined period of time. In each such slot the static groups are generated and their dynamics in time is inspected. So the further analysis comprises the following steps:

1. generating transitions between groups from neighbouring time slots (also finding stable groups) and identification of their type (group dynamics),
2. for each group assigning the set of topics discussed by group members (based on the content of messages exchanged between members of a group),
3. analyzing changes of topics during group transitions.

Measures that show very well the global user activity in blogosphere are: *lifetime of a post* and *reaction time for a post*. They can be used both in characterizing blogosphere and in determining the length of time slot and the length of overlapping.

### 3.1 Lifetime of a Post

The lifetime $lt$ of a post $p$ can be depicted in the following way

$$lt_p = \underset{i}{max}(t_{c_i}) - t_p \tag{1}$$

where $t_p$ is the date when post $p$ was published and $t_{c_i}$ are dates of comments in the thread of post $p$.

In other words, lifetime of a post is the range of time between writing the post and the last comment for that post.

### 3.2 Reaction Time for a Post

The reaction time $rt$ for a post $p$ can be formulated as (symbols used below in the definition were explained above)

$$rt_p = \underset{i}{min}(t_{c_i}) - t_p \tag{2}$$

Reaction time for a post is the range of time between writing the post and the first comment for that post.

### 3.3 Groups Dynamics

In order to analyse groups dynamics, overall time range was divided into smaller periods of time (called *time slots*). Next step is to build static networks in each time slot and discover groups in them. To identify events between groups from the neighbouring time slots SGCI method [16, 17] was utilized, which comprise the following stages: identification of short-lived groups in each time slot, identification of group continuations (groups transitions), separation of the stable groups (lasting for at least certain time interval) and the identification of types for group changes (labels for transitions between the states of the stable group).

Identification of continuation between groups $A$ and $B$ (from neighbouring time slots) is performed using $MJ$ measure

$$MJ(A, B) = \begin{cases} 0, & \text{if} A = \emptyset \vee B = \emptyset, \\ max\left(\frac{|A \cap B|}{|A|}, \frac{|A \cap B|}{|B|}\right), & \text{otherwise.} \end{cases} \quad (3)$$

and if the calculated value is above predefined threshold $th$ (in experiments we set $th = 0.5$) and the ratio of groups size

$$ds(A, B) = max\left(\frac{|A|}{|B|}, \frac{|B|}{|A|}\right) \quad (4)$$

is below predefined threshold $mh$ (in tests $mh = 50$), then we assumed that group $B$ is a continuation of group $A$.

Let us define $G_i$ as a set of groups in a time slot $i$. Then we can formally define transition $t_{m,n}$ between group $g_m$ from slot $k$ and group $g_n$ from slot $k + 1$ as

$$t_{g_m,g_n} : \exists g_m \in G_k \wedge \exists g_n \in G_{k+1} \wedge MJ(g_m, g_n) \geq th \wedge ds(g_m, g_n) < mh$$

The SGCI method identifies such event types as:

- **split**, takes place 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 (difference of sizes should be at least 10 times),
- **merge**, when several groups in the previous time slot join together and create larger group,
- **addition**, similar to merge, but it describes situation when small group attaches to significantly bigger group (difference of sizes should be at least 10 times),
- **split_merge**, when for the predecessor group the event is split and for the successor group of given transition the event is merge in the same time,
- **decay**, when analysed group does not exist in the next time slot,
- **constancy** means simple transition without significant change of the group size (in tests changes of group size should be smaller than 5 %),
- **change_size**—simple transition with the change of the group size.

## 3.4 *Topics in Groups*

Topics for groups were examined based on clusters found by LDA method. Presented method for analysis topics in groups was used by us in [17, 19, 20].

Firstly, we applied LDA method provided by mallet tool[11] for all posts and, as a result, the method identified 350 clusters of words. Next, we manually annotated each cluster by set of topics and joined similar clusters into bigger ones. The next step was inferring in every comment a set of topics that are referenced by analysed comment (the network is being built based on writing comments in response to other

[11] http://mallet.cs.umass.edu/.

message). Finally, we assigned for the group a set of topics discussed by members of this group (the following condition should be met in order to assign such topic for the group: a topic in a group should be present in at least 5 % of all interactions inside such group).

To describe the activity of topics in groups we formulated *topic exploitation* for given topic and group as a ratio between number of group messages on certain topic and all messages for this group:

$$topicExploitation_k = \frac{|T_k|}{\sum\limits_{i=1}^{n} |T_i|} \tag{5}$$

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

## 3.5 Topics Changes in Groups

Topics changes during transition between groups are assessed using introduced following metrics:

- *Overall change in topic exploitation* for transition $t_{m,n}$ from $m$-th to $n$-th group is calculated as:
$$c_{m,n} = \sum_i |g_{m,i} - g_{n,i}| \tag{6}$$
where: $i$ is a number of topic and $g_{m,i}$ is the topic exploitation of $i$-th topic for $m$-th group (the same denotations are used in definitions below).
- *Maximal positive change of single topic* (how much a topic gained) for transition $t_{m,n}$ from $m$-th to $n$-th group is calculated as:
$$mpc_{m,n} = \max\{g_{n,i} - g_{m,i}\} \tag{7}$$
- *Maximal negative change of single topic* (how much a topic lost) for transition $t_{m,n}$ from $m$-th to $n$-th group is calculated as:
$$mnc_{m,n} = \max\{g_{m,i} - g_{n,i}\} \tag{8}$$

Using above metrics we can analyse effect 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.

## 3.6 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 aggregated topics in model, $t_{group}$ 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 and weights for topics are determined based on topic exploitation measure for a group or an user.

The 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 carried out experiments focused on 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*. We assumed that candidates for joining are all users that were active in the 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* = $m$.

During calculations of 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 a small group.

# 4 Results

Experiments are conducted on datasets obtained from Polish blogosphere. Blogosphere is analysed from different points of view: especially in terms of users activity, groups formation and dynamics, topics discussed by users in groups.

## 4.1 Data Sets Description and Characteristics

The dataset contains data from the portal *Salon24* [12] (Polish blogosphere). This portal comprises blogs from different subjects, but political ones constitute the largest part of them. The data from this dataset is from time range 1.01.2008–6.07.2013.

Detailed characteristics of the dataset is presented in Table 1. We can see that on average posts are much longer than comments (almost 10 times). Moreover, commenting others is highly popular which can be perceived by the average number of comments per person (193.09). We can also observe that in general threads are quite long (on average 18.65 comments per post).

The whole period of time is divided into overlapping time slots, each lasting 7 days with overlap equals 4 days. After this operation dataset contains 504 slot.

In every time slot a static network is built [15], where 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.

## 4.2 Lifetime of Posts

Figure 1 shows *life time* of posts, i.e. time which elapsed from publish date and the moment when last comment to given post is written. As it can be seen life time of posts in Salon24 is very short—most of posts live up to one day.

## 4.3 Reaction Time for Posts

Figure 2 shows *reaction time*–time elapsed from post publish to the first comment. For Salon24 there is a tendency to write comments as soon as the post is published—most posts have comments within first hour. Both lifetime of a post and reaction time for posts show that this dataset is very dynamic and real world event have fast reflection in blogosphere.

[12] www.salon24.pl.

**Table 1** Data description

| | Salon24 |
|---|---|
| Nr of posts | 380,700 |
| Nr of comments | 5,703,140 |
| Nr of tags | 176,777 |
| Avg nr of posts per author | 37.58 |
| Avg nr of comments per person | 193.09 |
| Avg nr of comments to a post | 18.65 |
| Avg post length (characters) | 3,165.3 |
| Avg comment length (characters) | 350.72 |

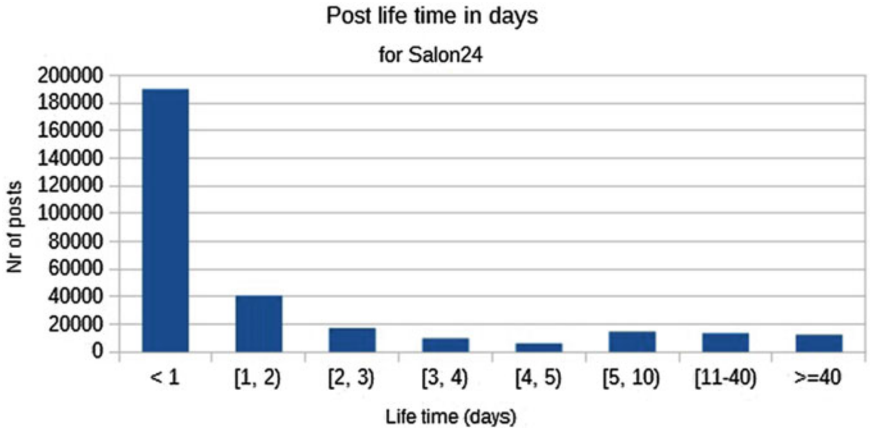

**Fig. 1** Post life time

## *4.4 Groups and Their Dynamics*

For group extraction we used CPM method (CPMd version which is designed to discover groups in directed networks) from CFinder[13] tool for k equals 3. Transitions between groups were assigned using our method SGCI described earlier.

Figure 3 presents how number of stable groups is influenced by group size. The general characteristics is the smaller group the bigger number of those groups.

Histogram of events taking places is shown in Fig. 4. Two very popular events are: change size, which seems natural because few users may join the group and some may leave, and decay—it is also expected as each group must finish at some point. Quite big values of addition and deletion may suggest that users of Salon24 are dynamic—they join discussions but also decide to leave threads.

[13] www.cfinder.org.

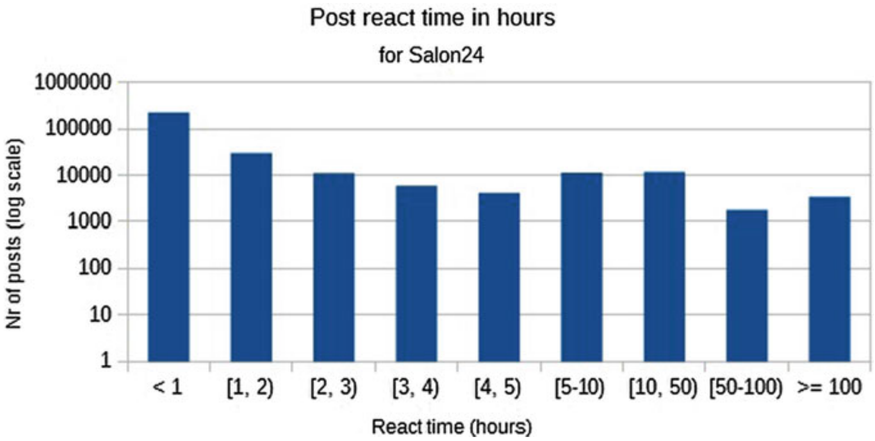

**Fig. 2** Post reaction time

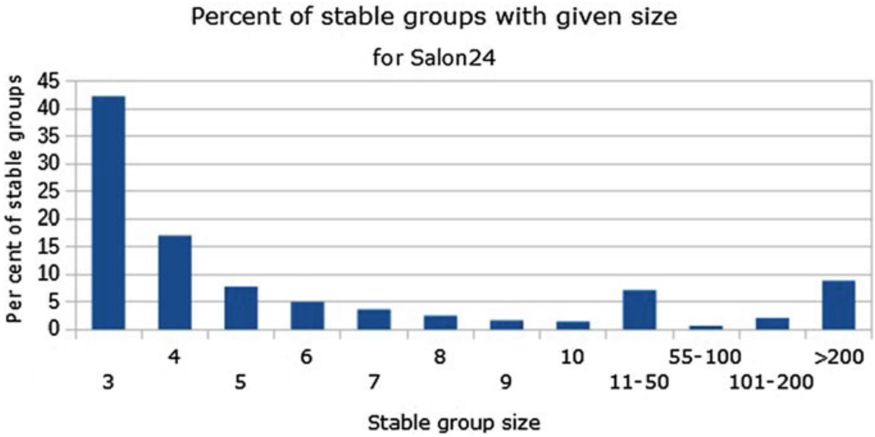

**Fig. 3** Number of stable groups

## 4.5 Topics in Groups

Figure 5 shows categories of topics groups are talking about. Each group may discuss different topics belonging to different categories. As it can be seen, the most popular categories are *economics* and *politics*. Next are *other*—collection of topics which are difficult to put in one category. High value for *disaster* is most probably connected with Tu-154 crash in Smolensk—this crash is still widely commented in Polish media.

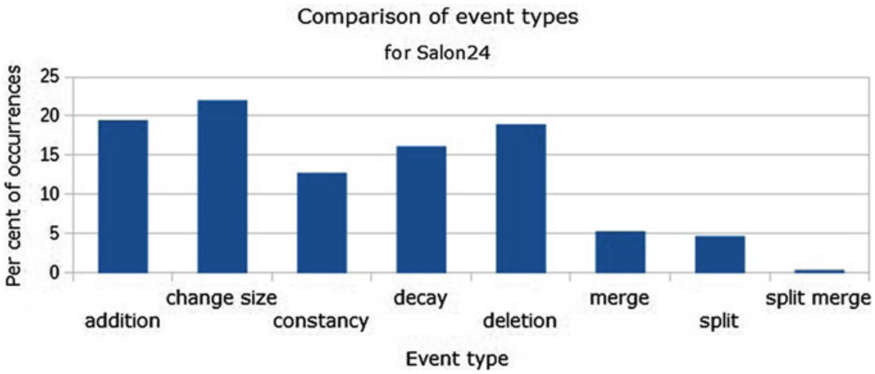

**Fig. 4** Number of events found

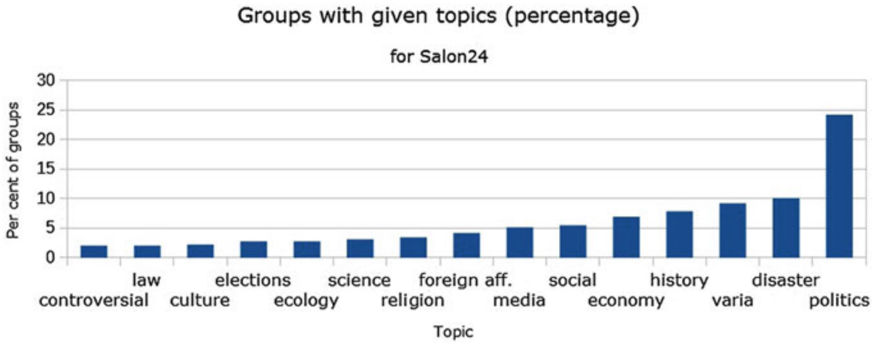

**Fig. 5** Percent of groups talking about particular topics

Number of topics discussed by groups depending on their size is presented in Fig. 6. Generally, in Salon24 groups have quite a few topics. A tendency that bigger group means lower number of discussed topics can be explained by 5 % threshold the topic must achieve to be counted. In big groups there are more topics but with lower weight which leads to only few ones that exceed the limit.

Topic convergence between users and groups they belong to is depicted in Fig. 7. Convergence is calculated in cosine measure where 0 means total divergence and 1 total convergence. As it could be expected the lowest value is for low convergence. Definitely in more than 50 % cases the convergence is higher than 0.5. Lower value for convergence between 0.9 and 1 and extremely low for exactly 1 (total convergence) is obvious—users creating a particular group put their topics into the group, but do not share all of them.

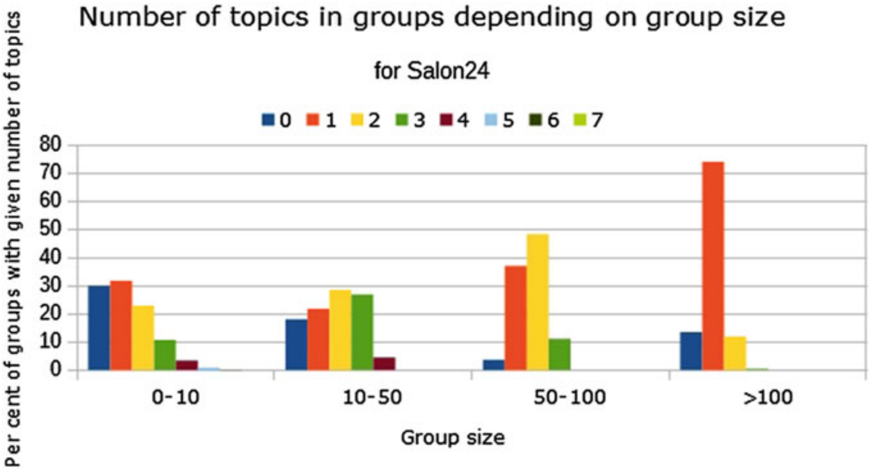

**Fig. 6** Number of topics depending on group size for Salon24

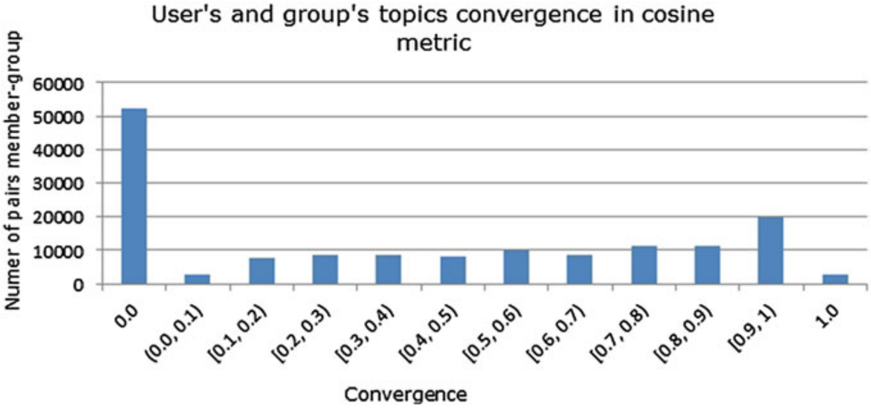

**Fig. 7** Convergence between user's and group's topics for Salon24

## 4.6 Topic Changes

Figure 8 shows average topic change depending on event type in half-year periods. As expected, *constancy* and *change size* have similar topic change. *Addition* and *deletion* also have similar values of change and are the highest ones. *Merge* and *split* cause medium topic change. As we can see, the biggest changes of topics happen during interaction groups with big difference in size (events *deletion* and *addition*), so it may suggest that interests of people are quite stable and changes are mostly related with exchange of members in groups.

Figure 9 shows maximal positive topic change. One can observe that biggest topic gain is achieved during event *deletion*, and the smallest one-event *constancy*.

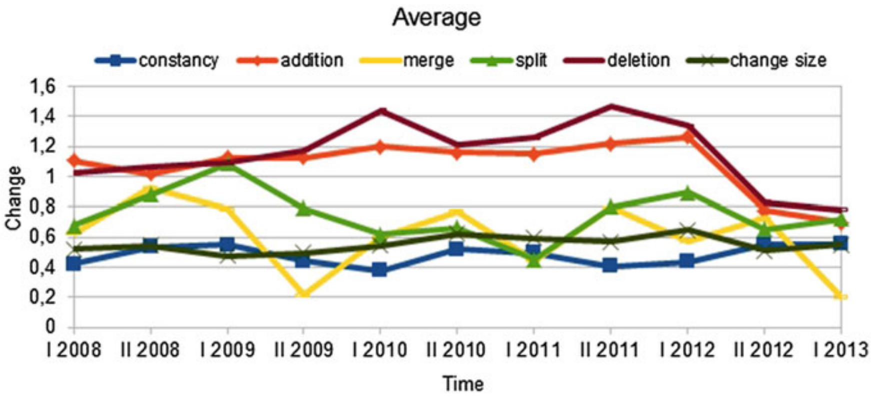

**Fig. 8** Average topic change in half-year slots for Salon24

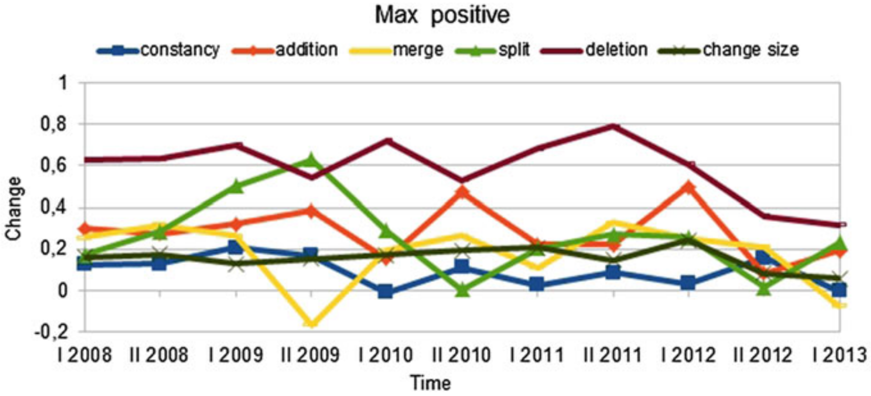

**Fig. 9** Maximal positive topic change in half-year slots for Salon24

Figure 10 depicts maximal negative topic change. In this case the event that mostly changes topics is *addition*. Furthermore, changes caused by the events *merge* and *split* are quite similar (even more similar than these events for the maximal positive topic change).

Generally, *addition* is connected with: highest overall change in topics and highest negative change. Therefore, we can deduce that when a small group attaches to a single large group it is usually connected with significant drop of popularity of the main topic of the group, and small rise in popularity of different topics—presumably of main topic of the other group.

*Deletion* cause large overall topic 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* and *constancy* have the smallest effect on topics changes.

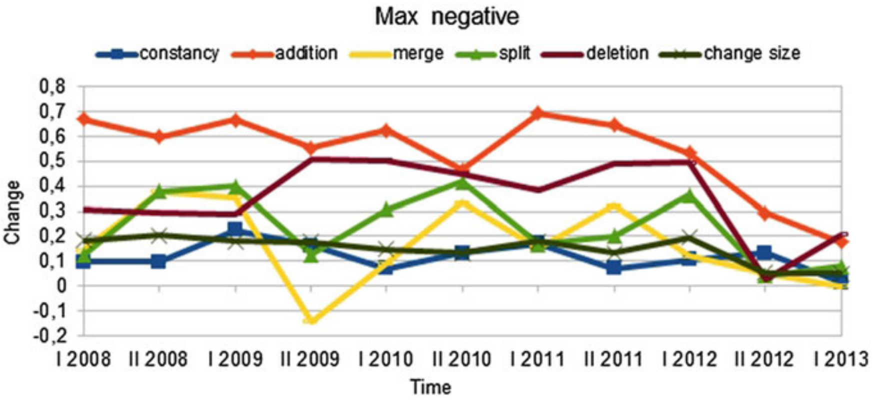

**Fig. 10** Maximal negative topic change in half-year slots for Salon24

*Split* has higher values for average and maximal negative change than *merge*. Moreover, when *split* achieves high values in terms of maximal positive topic change then it has rather small values in terms of maximal negative topic change. Different behaviour is for *merge* event—in that situation high values for maximal positive topic change corresponds with high values of maximal negative change. It may suggest that *merge* has averaging behaviour—after merge one topic gain more and other one decrease more.

## *4.7 Users Migrations*

Figure 11 presents probability of joining to a group and leaving it based on topic divergence and Fig. 12 shows numbers of people who actually did it. The value equals 0 for divergence means that a user talks about exactly the same topics as a group does and the value equals 2 means completely different topics. It is surprising that the highest convergence of topics does not mean high probability of joining [absolute number of such events is outstanding for total convergence (0)], but there are also some points (such as 0.05 and 0.4) where high convergence corresponds with higher probability of joining group, however, numbers of people with such values of convergence are rather small. We can observe different behaviour for leaving group by its members. Leaving the group is more probable when divergence is low and slightly increases for bigger divergence (however, for small values of topic divergence there are few members with such values of this measure, except the point 0 which means total convergence).

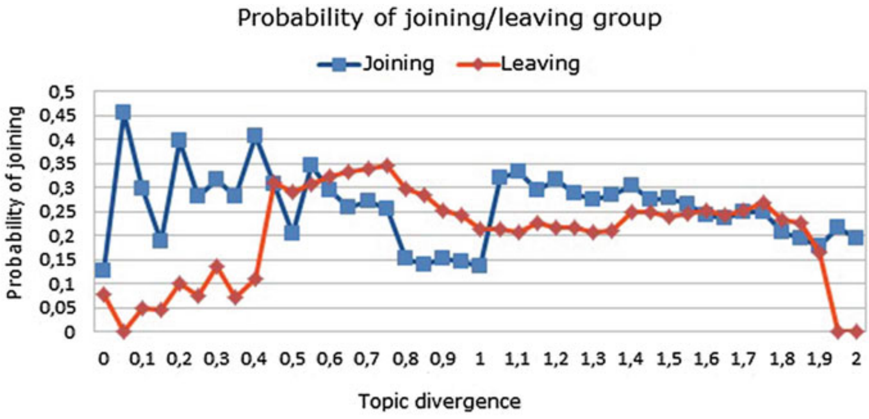

**Fig. 11** Migrations of users depending on topic similarity between user and group—probability

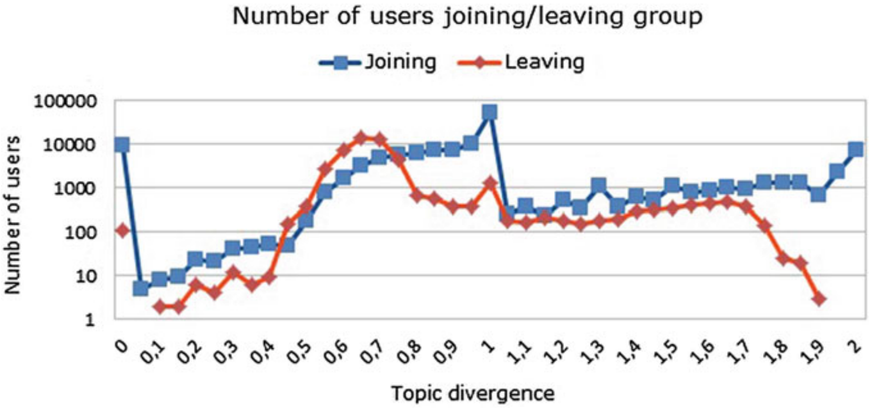

**Fig. 12** Migrations of users depending on topic similarity between user and group—numbers

### *4.8 Sample Lifecycle of Groups*

Figure 13 (screenshot from GEVi tool [18]—tool for visualisation of groups evolution) and Fig. 14 present sample group and its evolution in time. The notation in each boxes includes:

- slot (time period) and group name,
- 6 most important singular topics,
- 4 most important categories (aggregated and labeled topics),
- users belonging to the group.

Red colour indicates element that appears in next slot, green—element that does not go to the next period. Underline is used for elements that are new in current

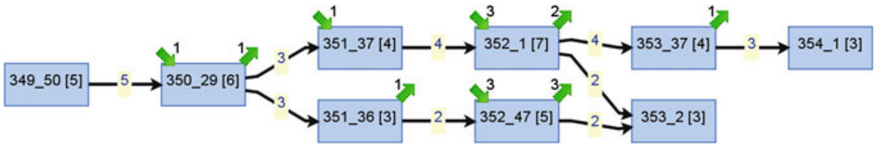

**Fig. 13** Sample group transitions for Salon24 in GEVi

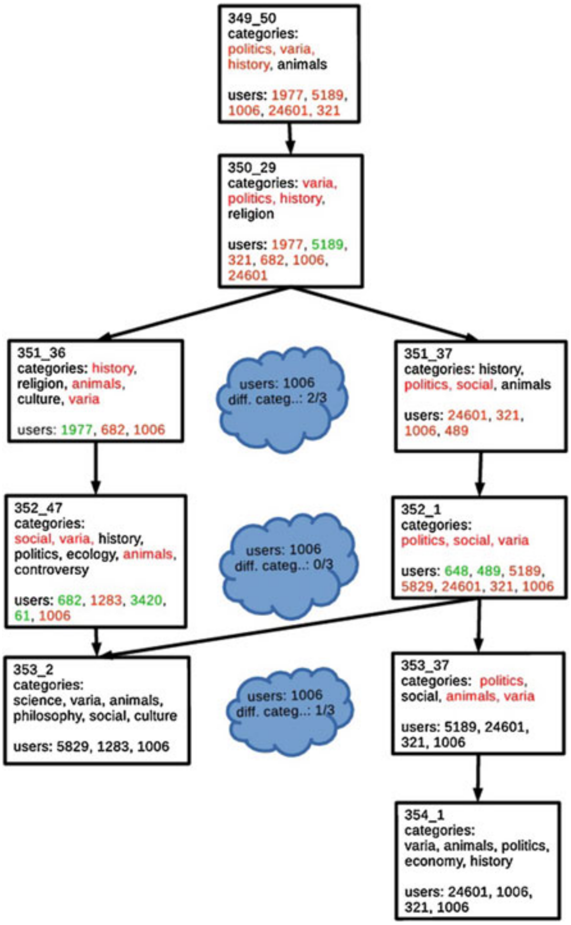

**Fig. 14** Sample group transitions for Salon24

group (not used for topics and categories, since only most important are listed). Cloud presents: ids of common users and number of extra categories and topics each group has (e.g. 6/5 means that upper group has 6 more topics than common ones, and lower has 5 more).

There is one group of 5 users that split in two in a second slot. One user stays in both, split groups. There are no clear topic transitions from one slot to other, however categories suggest some consequence in transitions. At the beginning group discuss philosophy, media and politics. Next, controversy and sport become most important, but politics appears as well. After split, upper group's first category is history, whereas in lower is sport. It is important to notice, that history also appears but with less importance. In next slot there are more differences between those groups: upper begins talking about animals, sport, law and religion, and continue those categories in next slots (animals appears in next slot, religion in last). Lower group talks about economy, history and food, and mainly those categories are in its next slot. It can be seen that one member from upper group and two from lower merges in slot 353.

## 5 Conclusion

This paper presents analysis of communities and their topics for real-world data from blogosphere. We carried out experiments regarding relations between topics discussed by members of groups and the behaviour of groups. Most important findings in this work are some patterns of topics changes during different group evolution events. Moreover, we assessed convergence topic interests of users and groups they belong to. Furthermore, we examined influence of topic differences between an user and a group on users migrations. Example of scenario for analysis topics during groups evolution was also presented.

Future work can be performed in a few different directions. Firstly, we are planning to conducts experiments on different datasets (e.g. in English language) and compare results, especially in terms of the patterns of topics changes during groups evolution. The second direction of further research is the analysis of key persons in terms of discussed topics and their influence on changes of topics during discussions. The third direction is to incorporate knowledge about topics interest of user to our method of predicting group behavior [14].

## References

1. Agarwal, N., Liu, H.: Modeling and Data Mining in Blogosphere. Moegan & Claypool Publishers, US (2009)
2. Aggarwal, C., Wang, H.: Social network data analytics. In: Aggarwal, C. (ed.) Text Mining in Social Networks, pp. 353–378. Springer, New York (2011)
3. Asur, S., Parthasarathy, S., Ucar, D.: An event-based framework for characterizing the evolutionary behavior of interaction graphs. ACM Trans. Knowl. Discov. Data **3**(4) (2009)
4. Bartal, A., Sasson, E., Ravid, G.: Predicting links in social networks using text mining and sna. In: Social Network Analysis and Mining, 2009. ASONAM '09. International Conference on Advances in, pp. 131–136 (2009). doi:10.1109/ASONAM.2009.12
5. Blei, D.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)

6. Blei, D., Lafferty, J.: Dynamic topic models. In: Proceedings of the 23rd international conference on machine learning, p. 113120 (2006)
7. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 9931022 (2003)
8. Carrington, P., Scott, J., Wasserman, S.: Models and Methods in Social Network Analysis. Cambridge University Press, Cambridge (2005)
9. Crain, S., Zhou, K., Yang, S., Zha, H.: Mining Text Data. In: Aggarwal, C., Zhai, C. (eds.) Dimensionality reduction and topic modelling: from latent semantic indexing to latent dirichlet allocation and beyond, pp. 129–162. Springer, New York (2012)
10. Cuadra, L., Rios, S., L'Huillier, G.: Enhancing community discovery and characterization in vcop using topic models. In: 2011 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 3, pp. 326–329 (2011). doi:10.1109/WI-IAT.2011.97
11. Diesner, J., Carley, K.: A methodology for integrating network theory and topic modeling and its application to innovation diffusion. In: 2010 IEEE Second International Conference on Social Computing (SocialCom), pp. 687–692 (2010). doi:10.1109/SocialCom.106
12. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**, 75–174 (2010)
13. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. **99**(12), 7821–7826 (2002)
14. Gliwa, B., Bródka, P., Zygmunt, A., Saganowski, S., Kazienko, P., Kozlak, J.: Different approaches to community evolution prediction in blogosphere. In: ASONAM 2013: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining: Niagara Falls, Turkey (2013). Accepted for printing
15. Gliwa, B., Kozlak, J., Zygmunt, A., Cetnarowicz, K.: Models of social groups in blogosphere based on information about comment addressees and sentiments. In: Social Informatics—4th International Conference, Social Informatics, Lausanne, Switzerland, Lecture Notes in Computer Science, vol. 7710, pp. 475–488. Springer (2012)
16. Gliwa, B., Saganowski, S., Zygmunt, A., Bródka, P., Kazienko, P., Kozlak, J.: Identification of group changes in blogosphere. In: ASONAM 2012: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, Turkey (2012)
17. Gliwa, B., Zygmunt, A.: Gevi: context-based graphical analysis of social group dynamics. Soc. Netw. Anal. Min. **4**(1), 1–15 (2014)
18. Gliwa, B., Zygmunt, A., Byrski, A.: Graphical analysis of social group dynamics. In: CASoN, pp. 41–46. IEEE (2012)
19. Gliwa, B., Zygmunt, A., Koźlak, J., Cetnarowicz, K.: Application of text mining to analysis of social groups in blogosphere. In: 5th Workshop on Complex Networks, CompleNet 2014, Bologna, Italy, 12–14 March 2014
20. Gliwa, B., Zygmunt, A., Podgórski, S.: Incorporating text analysis into evolution of social groups in blogosphere. In: Federated Conference on Computer Science and Information Systems, FedCSIS 2013, Krakow, Poland, 8–11 September 2013
21. Greene, D., Doyle, D., Cunningham, P.: Tracking the evolution of communities in dynamic social networks. In: Proceedings of International Conference on Advances in Social Networks Analysis and Mining (ASONAM'10). IEEE (2010)
22. Gundecha, P., Liu, H.: Mining social media: A brief introduction. Tutorials in Operations Research 1,4, Informs. Arizona State University, US (2012)
23. Huang, Y.: Support vector machines for text categorization based on latent semantic indexing. Electrical and Computer Engineering Department, The Johns Hopkins University, Technical report (2003)
24. Nguyen, M., Ho, T., Do, P.: Social networks analysis based on topic modeling. In: IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), pp. 119–122 (2013). doi:10.1109/RIVF.2013.6719878
25. Palla, G., Barabsi, I.A., Vicsek, T., Hungary, B.: Quantifying social group evolution. Nature **446**, 664–667 (2007)

26. Palla, G., bel, D., Farkas, I.J., Pollner, P., Dernyi, I., Vicsek, T.: Handbook of large-scale random networks. In: Bollobs, B., Kozma, R., Mikls, D. (eds.) k-clique Percolation and Clustering. Springer, New York (2009)
27. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**, 814–818 (2005)
28. Takaffoli, M., Rabbany, R., Zaiane, O.R.: Incremental local community identification in dynamic social networks. In: J.G. Rokne, C. Faloutsos (eds.) ASONAM, pp. 90–94. ACM (2013)
29. Tang, L., Liu, H.: Community Detection and Mining in Social Media. Morgan & Claypool, US (2010)
30. Velardi, P., Navigli, R., Cucchiarelli, A., D'Antonio, F.: A new content-based model for social network analysis. In: IEEE International Conference, Semantic Computing, pp. 18–25 (2008). doi:10.1109/ICSC.2008.30
31. Xu, J., Marshall, B., Kaza, S., Chen, H.: Analyzing and visualizing criminal network dynamics: A case study. In: IEEE Conference on Intelligence and Security Informatics. Tuczon (2004)