Hindawi Publishing Corporation
Scientific Programming
Volume 2015, Article ID 907547, 11 pages
http://dx.doi.org/10.1155/2015/907547

Research Article

The Comparison of Users Activity on the Example of Polish and American Blogosphere

Anna Zygmunt and Bogdan Gliwa

AGH University of Science and Technology, Al. Mickiewicza 30, 30-059 Kraków, Poland

Correspondence should be addressed to Anna Zygmunt; azygmunt@agh.edu.pl

Received 21 March 2014; Accepted 26 November 2014

Academic Editor: Reda Alhajj

Blogs are popular way to express opinions on the Internet. Due to their popularity and their public character blogs attract attention of many researchers. In this paper we compare two national blogospheres (Polish and American) from different angles such as characteristics of messages and interactions, structure of social groups, topics discussed in them, and the influence of real-world events on the behavior of such groups. In our approach we try to combine in advanced manner users activity on both the individual and community level. The comparison reveals some differences and various characters of both portals. Methods for analysis of groups dynamics, users roles, and topics in groups are presented.

1. Introduction

Nowadays a large part of our life has moved to the Internet, particularly to the social media. It is hard to imagine that we stop using them. Willingly or not, we are present in them, even passively searching for sources of information. A large part of the official and unofficial life has moved there. There are various reasons for this situation, but one thing must be said with certainty that this is a process that cannot be stopped. The majority of us are only passively involved in it, treating different types of forms of social media as sources of information, that is, places where one can learn something. But there are also people who participate in social media actively and creatively: expressing their opinions, commenting on others, promoting opinions of others, and so forth. They leave so many "traces" of their activities, which can then be analyzed to find interesting patterns of human life, which can be used in marketing, business, politics, or public security domains.

The social media may take many forms, for example, blogs, forums, media sharing systems, microblogging, social networking, and wikis. Among them, blogs play a special role. The term "blogosphere," first introduced by Brad L. Graham in 1999, should be understood as a term describing all blogs. Observing the development of blogosphere, one can say that they have passed a long way from frivolous diaries to very serious sources of information. Undoubtedly, the reason for this situation has become a development tool for creating blogs, as well as the fact that many important people have discovered that blogs are a very good place to express their opinions and to observe an immediate response to them. It is believed that blogs have become a flywheel for the development of online social networking [1]. Now blogs are used as a communication platform and more and more as of source knowledge. Blogs can be treated as web pages with entries arranged in the reverse order (due to chronology). Such pages can contain text, links, pictures, videos, and so forth.

Blogosphere is an interesting source of data for analysis. It is characterized by (in most cases) high dynamics: posts are often added as well as comments on them; one can analyze the reactions of readers to the posts, both in terms of response speed as well as emotion (sentiment analysis). One can analyze themes of posts and find those that receive the greatest interest (getting the most comments) as well as users who generally write such influential posts. Until recently, the analysis of the processes taking place in blogosphere was the domain of research conducted mainly by psychologists and sociologists. These studies were characterized by carrying out analyses to a limited extent due to problems with data

collection. With the development of technological capabilities allowing for automatic and incremental collection of any amount of data from blogosphere and storing them in huge databases have significantly increased the possible directions of research.

The paper presents a comparison in various aspects of users activity in Polish and American blogosphere.

Generally, to our knowledge, there is no such comprehensive comparative analysis of two blogospheres in such a wide range as we have done. Particular areas of research appear in single studies. In some articles the authors analyze groups in blogosphere (but without taking into account the dynamics of change); others examine influential bloggers or analyze topics of discussion. Our approach assumes broad comparison of two national blogospheres by analyzing the structure of the groups that are formed and continued for a period of time, comparing the roles of users played in both the group and the globe in the whole network, as well as the identification of topics of conversation and the study of reaction time for posts in different blogospheres. Such a global approach to the analysis of the users allows creating much more advanced user profiles, at both the individual and global level, as well as finding user's characteristics that are common to different nationalities, as well as those that differentiate them.

The structure of the paper is as follows. In Section 2, current research directions over blogosphere, as well as a review of research on groups and their dynamics, finding roles, and text analysis are presented. Section 3 contains an overview of our algorithms used for finding stable groups, identifying events, finding roles, and identifying topics of posts and comments. In Section 4 both datasets are described in detail and results are presented and discussed. Section 5 concludes and shows possible directions of future works.

2. Related Work

2.1. Blogosphere: Direction of Research. Blogosphere soon became an interesting research area for psychologists and sociologists. The research methodology was largely based on designing questionnaires and asking questions to a properly selected group of respondents (according to, e.g., demography). The results of the analysis were strictly dependent on the truthfulness of responses and the sample size of blogs, which, due to the need for manual processing, was not big. The most interesting subject of research was to determine why people started a blog and reasons they had for continuing writing. They tried to find differences based on gender and demographics of bloggers.

Initially, these analyses concerned a single nationality. Then blogs belonging to representatives of different nations were analyzed to compare and find out if there were any differences related to cultures diversity. Analyses of individual nationalities concerned tracking changes in the demographics of bloggers or certain groups of bloggers were studied.

The vast majority of authors [2–4] concluded that in general motivations for blogging were the same in all analyzed nationalities (self-expression, social interaction, entertainment, passing time, information, and professional advancement), but they had different priority. In [1] motivations for blogging were linked to identity. In [5], types of characters (extroverts, introverts, etc.), language, and gender were analyzed and their impact on the content and topics discussed on blogs was described.

In [6], the group membership was analyzed, but bloggers indicated which group they belonged to and why. In [1] the authors compared bloggers from different countries and analyzed their habits (e.g., differences in activity depending on time of day). The need for research groups of bloggers and analysis of their dynamics was identified, but no studies were carried out.

Since computer scientists started to be interested in the analysis of blogosphere, research has sped up, because there is a real possibility of automatic data collection from the blogosphere using webcrawlers, saving them to big, effective databases and performing virtually any analysis on such data. So there is no need to develop an experiment, invent questions, and collect responses and analyze only data. Usually all data from the page are collected, such as demographic information, text of posts, comments (as well as information about their authors), links, tags, dates, and all other kinds of available information. Directions of research now are much less related to demography, because such data are usually not available. Because all data are available in database, one can freely invent and change the directions of analysis. Generally, this research can be divided into two directions: structure and content analysis.

One of the directions of the analysis was to use methods of social network analysis [7] to analyze the popularity of bloggers (or posts). In [8, 9] Kleinberg algorithm HITS finding hubs and authorities was used to find top bloggers. A-list blogs of the most read, most quoted, and most number of inbound links from others were used.

In other studies [5], the authors attempted to determine what impact, for example, psychological profiles and gender have on the way of writing on blogs. Methods of text processing were used (large blog corpus was collected) in order to extract topics from the text. On the basis of those topics they attempted to create psychological profiles.

The first approach to find clusters in blogospheres and recognize the structure was in [9]. They observed that blogosphere was "selectively interconnected with dense clusters in parts and blogs minimally connected in local neighberhood *[sic]* or flee-floating individually, constituing *[sic]* the majority." In [10] structure A-list was used to find core structures in six national blogospheres. That model was compared with [11]. Differences in cores structures were explained by cultural differences.

In [12, 13] Chinese and German blogospheres were compared to find differences in the structure of pages, length of comments, and time of reactions. It was observed that, in spite of cultural diversity, blogging services worked in a similar way (Chinese bloggers could do more with design of pages).

In [14] data from Polish blogosphere, discussion on MySpace, YouTube comments, and forums BBC were compared according to the length of comments in words and bytes, and it was concluded that overall lengths were similar.

2.2. Groups in Social Networks. Social network is not a homogeneous structure; it rather consists of areas in which vertices communicate to each other more frequently than with vertices outside given area. Such areas are called groups (communities, module, cluster, and subgroups). There are many methods of finding such groups, which can be overlapped (or not) [15, 16]. Finding groups allow simplifying the complex network or analyzing certain processes in micro- and macroscale. Quality of group can be measured by several parameters indicating its size, durability, or importance, for example, density (ratio of the number of links within the group to the maximum possible number of links), cohesion (ratio of the average strength of links between the members to the average strength of their links with people outside the group), or stability between groups (the ratio of the number of people, present in both groups, to the number of all group members). One of the most popular representatives of algorithms finding overlapping groups is CPM (Clique Percolation Method) [17].

2.3. Group Dynamics. Even though most methods have been developed for static environment, many researchers have recognized the need for better reflecting the dynamic nature of the most social networks (especially coming from social media sites) [18, 19]. For dynamic network analysis the common way is to divide given period of time into smaller units called time slots. Then, in each time slot the static network is analyzed and the groups are extracted. Next step is to determine the transitions between groups from neighboring time slots. For this purpose, Greene et al. [16] used the Jaccard index as a measure describing the similarity of groups (the measure is calculated for each pair of groups from neighboring time slots). The value of this measure above arbitrarily defined threshold level means that one group is continuation of another. Some other measures for obtaining transitions between groups have been proposed in literature [20, 21].

Palla et al. in [22] identified basic events (transitions) that may occur in the life cycle of the group: growth, merging, birth, construction, splitting, and death. They did not give any additional conditions. Asur et al. in [18] introduced formal definitions of five critical events. Gliwa et al. proposed in [20] two additional events and gave formal definitions. In [23] new tool GEVi for context-based graphical analysis of social group dynamics was proposed.

2.4. Roles of Users. In social network analysis there are many definitions of role [24–26]. In social media, *role* can be treated as a set of characteristics that describe behavior of individuals and the interactions among them within a social context [27].

Roles in the literature are often discussed in the context of influences [28]. Agarwal et al. in [29] defined influential bloggers and gave their characteristics and described four types of bloggers: active and influential, inactive but influential, active but noninfluential, and inactive and noninfluential.

A lot of studies relate to certain social media and attempt to define their specific roles [30, 31]. For example, an analysis of the basic SNA measures has been used in several studies to define social roles of *starters* and *followers* in blogosphere [32, 33]. *Starters* receive messages mostly from people who are well connected to each other, and therefore they can be identified by low in-degree, high out-degree, and high clustering coefficient in the graph. The distinction between the roles is obtained by combining the difference between the number of in-links and out-links of their blogs.

2.5. Text Mining in Domain of Social Networks. Aggarwal and Wang in [34] provided overview of text mining methods useful for social networks analysis, but in literature text mining combined with SNA is used mostly in some specific cases. Bodendorf and Kaiser in [35] used text mining to extract opinions from texts and then integrated such information with social network analysis approach to find opinion leaders and detect trends in communities. Bartal et al. [36] proposed a method for predicting links in a network based on social network analysis and text data mining approach.

Topic modeling [37] is a statistical technique that uncovers abstract "topics" that can be found in a collection of documents. "Topic" can be defined as a set of words that tend to cooccur in multiple documents, and, therefore, they are expected to have similar semantics. One of the main benefits of this method is that similar texts can be discovered even if they use different vocabulary. One of the most popular methods in topic modeling is Latent Dirichlet Allocation (LDA) [38]. In [39] the authors showed usefulness of topic modeling to analysis of groups dynamics in social networks in blogosphere. Another approach using topic modeling along with social network analysis is presented in [40] where authors track topics in time and automatically assign labels for topics.

3. Methods Used during the Comparison of Different Blogospheres

In this section we describe measures and methods applied to comparison of two blogospheres: American and Polish one. Firstly, we provide definitions of measures utilized to assess different characteristics. Next, we depict methods for analysis of groups dynamics, users roles, and topics in groups.

3.1. Lifetime of a Post. The lifetime lt of a post p can be defined as

$$\mathrm{lt}_p = \max_i \left(t_{c_i}\right) - t_p, \tag{1}$$

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 formalized in the following way (symbols used in the definition were explained above):

$$\mathrm{rt}_p = \min_i \left(t_{c_i}\right) - t_p. \tag{2}$$

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. To analyse groups dynamics, whole range of time was divided into smaller periods of time (called later *time slots*). Next, in each time slot, the static network was analysed and the groups were extracted. To identify events between groups from the neighbouring time slots SGCI method [20, 41] was employed, which consists of the following stages: identification of short-lived groups in each time slot, identification of group continuation, 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).

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.

Using above measures we can define transition $t_{g_{i,k},g_{i+1,l}}$ between group g_k in ith slot and group g_l in $(i+1)$th time slot as

$$t_{g_{i,k},g_{i+1,l}} : \exists g_{i,k} \wedge \exists g_{i+1,l} \wedge MJ(g_{i,k}, g_{i+1,l}) \geq \text{th} \wedge ds(g_{i,k}, g_{i+1,l}) < mh. \quad (5)$$

Now we can label transitions:

(i) addition: when a small group attaches to big one

$$t_{g_{i,k},g_{i+1,l}} : \frac{|g_{i+1,l}|}{|g_{i,k}|} \geq sh, \quad (6)$$

(ii) deletion: when a small group detached from big one

$$t_{g_{i,k},g_{i+1,l}} : \frac{|g_{i,k}|}{|g_{i+1,l}|} \geq sh, \quad (7)$$

(iii) merge: when many groups join together into bigger one

$$t_{g_{i,k},g_{i+1,l}} : ds(g_{i,k}, g_{i+1,l}) < sh \wedge \left[\exists t_{g_{i,m},g_{i+1,l}} : m \neq k \wedge ds(g_{i,m}, g_{i+1,l}) < sh\right] \wedge \left[\nexists t_{g_{i,k},g_{i+1,n}} : n \neq l \wedge ds(g_{i,k}, g_{i+1,n}) < sh\right], \quad (8)$$

(iv) split: when group divides into 2 or more groups in the next time slot

$$t_{g_{i,k},g_{i+1,l}} : ds(g_{i,k}, g_{i+1,l}) < sh \wedge \left[\exists t_{g_{i,k},g_{i+1,n}} : n \neq l \wedge ds(g_{i,k}, g_{i+1,n}) < sh\right] \wedge \left[\nexists t_{g_{i,m},g_{i+1,l}} : m \neq k \wedge ds(g_{i,m}, g_{i+1,l}) < sh\right], \quad (9)$$

(v) split_merge: combination of event *merge* and *split* for the same transition

$$t_{g_{i,k},g_{i+1,l}} : ds(g_{i,k}, g_{i+1,l}) < sh \wedge \left[\exists t_{g_{i,m},g_{i+1,l}} : m \neq k \wedge ds(g_{i,m}, g_{i+1,l}) < sh\right] \wedge \left[\exists t_{g_{i,k},g_{i+1,n}} : n \neq l \wedge ds(g_{i,k}, g_{i+1,n}) < sh\right], \quad (10)$$

(vi) constancy: simple continuation of a group without significant change of size

$$t_{g_{i,k},g_{i+1,l}} : \frac{\text{abs}(|g_{i,k}| - |g_{i+1,l}|)}{|g_{i,k}|} \leq dh \wedge \left[\nexists t_{g_{i,m},g_{i+1,l}} : m \neq k \wedge ds(g_{i,m}, g_{i+1,l}) < sh\right] \wedge \left[\nexists t_{g_{i,k},g_{i+1,n}} : n \neq l \wedge ds(g_{i,k}, g_{i+1,n}) < sh\right], \quad (11)$$

(vii) change_size: simple continuation of a group with significant change of size

$$t_{g_{i,k},g_{i+1,l}} : \frac{\text{abs}(|g_{i,k}| - |g_{i+1,l}|)}{|g_{i,k}|} > dh \wedge \left[\nexists t_{g_{i,m},g_{i+1,l}} : m \neq k \wedge ds(g_{i,m}, g_{i+1,l}) < sh\right] \wedge \left[\nexists t_{g_{i,k},g_{i+1,n}} : n \neq l \wedge ds(g_{i,k}, g_{i+1,n}) < sh\right], \quad (12)$$

(viii) decay: when a group disappear in the next time slot

$$\nexists t_{g_{i,k},g_{i+1,l}}. \quad (13)$$

In above definitions we used function abs which means absolute value function and some parameters: sh, threshold for ratio of groups size and dh, threshold for groups size differences. In experiments we set value of sh to 10 and value of dh to 0.05.

3.4. Roles of Users. Users can play different roles on a global level and different ones in each of the groups they belong to (local level of roles). The set of roles we use for analysis in this paper was proposed by us in [42].

The presented roles take into consideration responses from other users on the content the user writes (in both the form of posts and comments). To meet such assumptions, we defined *Post* and *Comment Influence*.

Post Influence for author a has the following form (in this definition we use the notation $c(X, \text{cond})$ that means

the number of elements in X that every element of X fulfills condition cond):

$$\begin{aligned}\text{PostInf}_a = {} & 4 \cdot c\left(p_a, pr \geq A_1\right) + 2 \cdot c\left(p_a, pr \geq A_2\right) \\ & + c\left(p_a, pr \geq A_3\right) - c\left(p_a, pr < A_4\right) \\ & - 2 \cdot c\left(p_a, pr < A_5\right) - 4 \cdot c\left(p_a, pr < A_6\right),\end{aligned} \tag{14}$$

where p_a is the posts of author a; pr is the number of comments for a given post excluding the author's comments in his own thread; for global roles we set the following values: $A_1 = 50$, $A_2 = 25$, $A_3 = 10$, $A_4 = 2$, $A_5 = 1$, and $A_6 = 0$; for local roles we set the following values: $A_1 = 10 \cdot B$, $A_2 = 0.25 \cdot A_1$, $A_3 = 0.25 \cdot A_2$, $A_4 = A_5 = 0$, $A_6 = 1$, and B = group Density · group Size.

Comment Influence for author a is calculated in the following way (in this definition we use the notation w(cond) that returns 1 when the condition cond is satisfied, otherwise—0):

$$\begin{aligned}\text{ComInf}_a = {} & 4 \cdot w\left(r_a \geq 1.25\right) + 2 \cdot w\left(r_a \geq 1\right) \\ & + w\left(r_a \geq 0.75\right) - w\left(cr_a < C_1\right) - 2w\left(cr_a < C_2\right) \\ & - 4 \cdot w\left(cr_a < C_3\right),\end{aligned} \tag{15}$$

where r is the number of received comments from other users divided by the number of written comments by given authors; cr is the number of received comments from other users; for global roles we set the following values: $C_1 = 50$, $C_2 = 20$, and $C_3 = 10$; for local roles we set the following values: $C_1 = 0.5 \cdot B$, $C_2 = 0.25 \cdot C_1$, $C_3 = 0.25 \cdot C_2$, and B = group Size · group Density.

Using the above definitions we can describe the set of roles:

(1) Influential User (infUser): PostInf > 2 and ComInf > 0,

(2) Influential Blogger (infBlog): PostInf > 2 and ComInf ≤ 0,

(3) Influential Commentator (infComm): ComInf > 0 and PostInf ≤ 2,

(4) Standard Commentator (comm): c(comments) ≥ 20 and c(posts) ≤ 2,

(5) Not Active (notActive): c(posts) < 1 and c(comments) < 2,

(6) Standard Blogger (stdBlog): user that does not match any from above roles.

3.5. Topics in Groups. Topics for groups were assigned based on clusters uncovered by LDA method. The method for analysis topics in groups was used by us in [23, 39].

Whole method can be described as a set of the following steps. Firstly, we used LDA method provided by mallet tool (http://mallet.cs.umass.edu/) for all posts and the method discovered 350 clusters of words. Next, we manually annotated each cluster by set of topics and joined similar clusters into bigger ones. After that operation, we infer in every comment a set of topics that are referenced by this comment (the network is being built based on writing comments in response to other messages—precise way of building network for each dataset is described in Section 4.1). We consider 2 variants of the method (in results referred to as *method 1* and *method 2*) which differ only in a way of assigning a topic for a comment when LDA could not find any matching topics. The first variant (*method 1*) does not assign any topic if it could not be inferred for given comment, but the second variant (*method 2*) in such case uses topics assigned for the parent comment (if the analysed comment has the parent one and the parent comment has any assigned topics) or the post in the thread where the comment was written. Next step is to assign for the group a set of topics discussed by members of this group (we required that topic should be present in at least 5% of all interactions inside a group to assign such topic for the group).

We can formalize it in the following way. Let us define T as a set of topics (after operation of annotating and joining similar clusters from LDA):

$$T = \{t_1 \cdots t_k\}, \tag{16}$$

members of a group G

$$\text{members}\,(G) = \{a_1 \cdots a_n\}, \tag{17}$$

edges in a group G

$$\text{edges}\,(G) = \left\{e_{xy} : x \in \text{members}\,(G) \wedge y \in \text{members}\,(G)\right\}, \tag{18}$$

topics for edge e_{xy}

$$\text{topics}\left(e_{xy}\right) = \{t_k\} \wedge \text{topics}\left(e_{xy}\right) \subset T. \tag{19}$$

Using above notation we can define topics for a group G

$$\begin{aligned}&\text{topics}\,(G) \\ &= \left\{t_k : \forall_k \exists_x \exists_y \left[e_{xy} \in \text{edges}\,(G) \wedge t_k \in \text{topics}\left(e_{xy}\right)\right]\right. \\ &\quad \left. \wedge \forall_k \frac{t_k}{\sum t_k} \geq h\right\},\end{aligned} \tag{20}$$

where h is a threshold and we used $h = 0.05$.

4. Results

In this section we compare Polish and American blogosphere from different points of view, especially in terms of users activity, groups formation, and topics discussed by users in groups. For this purpose, we chose one dataset as a representative for Polish blogosphere and one for American one.

Table 1: Comparison of data quantity in both datasets.

Measure	*Salon24*	*Huffington Post*
Number of posts	380 700	414 225
Number of posts without comments	74 979 (19.7%)	45 604 (11%)
Average number of comments in one post	18.65	48.28
Number of comments	5 703 140	17 796 819
Number of comments to posts	2 781 303 (48.77%)	6 961 369 (39.12%)
Number of comments to other comments	2 921 837 (51.23%)	10 753 162 (60.88%)
Number of authors	31 750	680 341
Number of authors of posts	10 131 (31.91%)	1 027 (0.15%)
Number of authors of comments	29 536 (93.03%)	661 676 (97.26%)

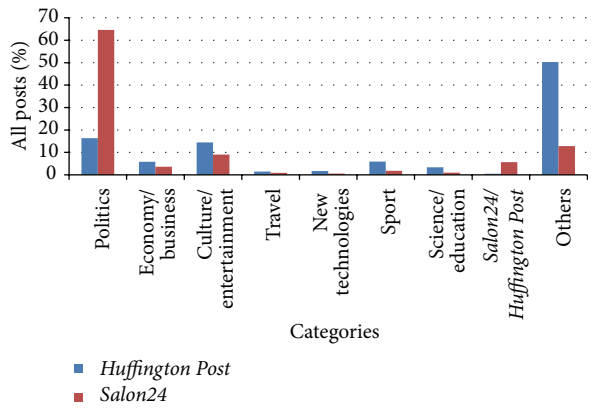

Figure 1: Categories of posts.

4.1. Datasets Description. The first dataset contains data from the portal *Salon24* (http://www.salon24.pl/) (Polish blogosphere). This portal comprises blogs from different subjects, but political ones constitute the largest part of them (as you can see in Figure 1). The data from this dataset is from time range 1.01.2008–6.07.2013. Whole period of time was divided into overlapping time slots, each lasting 7 days and the neighbouring slots overlap each other by 4 days. After this operation the dataset contains 504 slots. In every time slot a static network is built according to *comments model* introduced in [43]; that is, 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.

The second dataset is the *Huffington Post* dataset (http://www.huffingtonpost.com/) (American blogosphere) which contains news and blogs from various subjects (we can see in Figure 1 that political topics constitute significant part of all posts, but this topic does not outnumber other ones as it was in the case of *Salon24*). This dataset contains data from period 1.01.2010–14.11.2013. Similarly as for Salon24 dataset, the whole period of time was divided into overlapping time slots, each lasting 7 days with overlap equal to 4 days, which produced 442 slots (but for the analysis we used slots in this dataset starting from 97 because in previous one there were some slots where groups were not found). In *Huffington* dataset networks in time slots are built in similar way as for *Salon24* dataset (edges between an author of given comment and an author of a comment the response is addressed for, or, if a comment is not an answer for another comment, between an author of given comment and an author of a post), but in this case the explicit references between comments exist (hierarchical structure of comments).

Moreover, due to the performance issues of group extraction method in order to detect communities, we eliminated the edges with weight equal to one in each time slot. But for other types of analyses (such as role finding) we conducted them on full graphs without any edge removal.

4.2. Basic Statistics. As we can observe in Table 1 the *Huffington Post* dataset is bigger than *Salon24* one. Threads in *Huffington Post* are also longer; that is, on average posts have more comments in *Huffington Post* than in *Salon24*. We can see that in both datasets the responses to other comments represent a substantial part of all comments. Another interesting fact is that authors of posts in *Huffington Post* constitute much smaller fraction of all authors (less than 1%) as compared with *Salon24* (almost 32%). This means that character of both portals is quite different. In *Salon24* a significant number of users have a contribution to creating posts and informing about new events from the world, but in *Huffington Post* the users are oriented towards commenting on posts and this portal plays a role more similar to an Internet newspaper.

4.3. Lifetime of Posts. Figure 2 presents lifetime of posts (it is a cumulative chart so it depicts percentage of all posts that have lifetime equal or less than specified value). We can see that almost 90% of posts in *Salon24* have their lifetime up to 1 week, but similar lifetime in *Huffington Post* is achieved after 2 months (8 weeks). This means that in *Salon24* posts older than 1-2 weeks are rarely commented on and the attention of users is brought mostly by new posts, which is a bit different than in *Huffington Post* where significant part of users comments also on older posts than 1 week. Such a difference in lifetime of posts between these 2 datasets also emphasizes higher dynamics in *Salon24*.

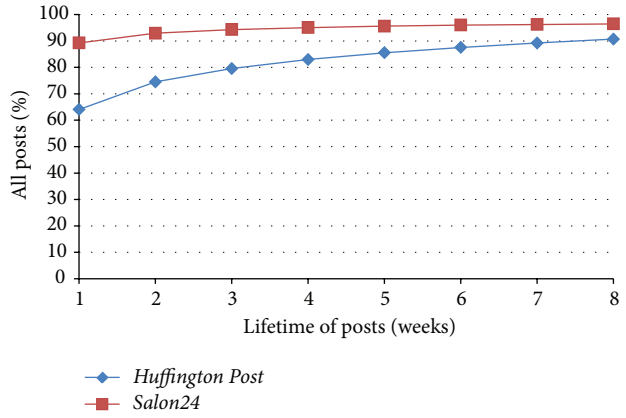

Figure 2: Lifetime for posts.

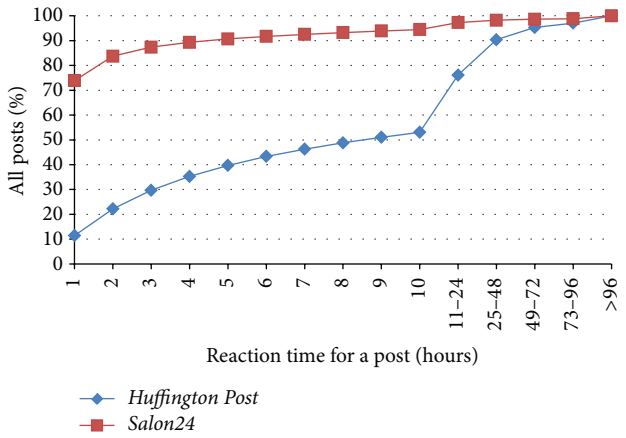

Figure 3: Reaction time for a post.

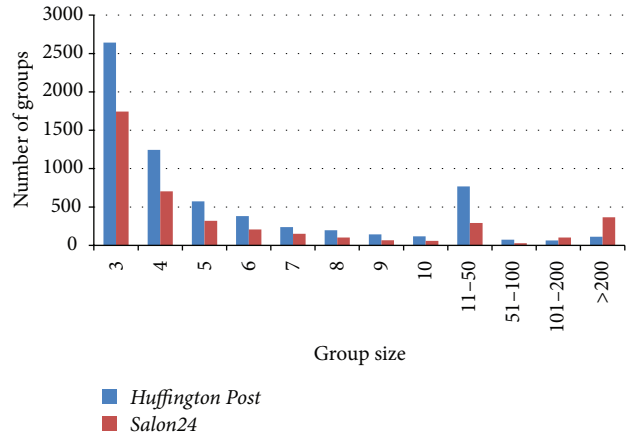

Figure 4: Number of stable groups at given size.

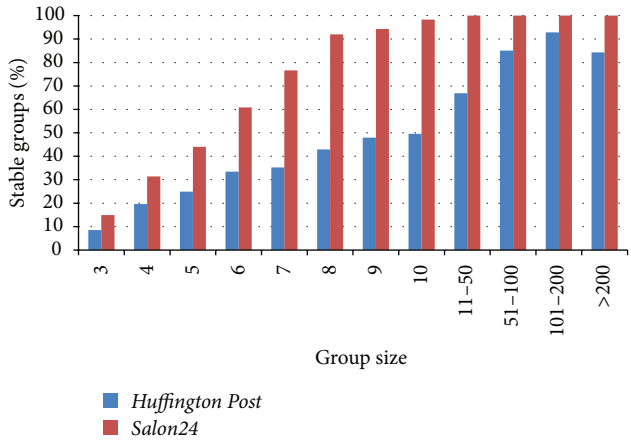

Figure 5: Percentage of stable groups in relation to all groups.

4.4. Reaction Time for Posts. Figure 3 depicts reaction times for a post in both datasets. One can notice a big difference in dynamics between *Huffington Post* and *Salon24*—in the first hour after publishing a post in *Salon24* 73.9% of all posts received at least one comment, but in *Huffington Post* only 11.4% of all posts. After 2 days after writing a post, in both blog portals more than 90% of posts were commented on. We also investigated the amount of time needed for a half of all posts to get the first comment. For *Huffington Post* we need about 8 hours, but in *Salon24* it is sufficient to wait only 32 minutes after writing a post to receive a comment.

4.5. Groups and Their Dynamics. For group extraction we used CPM method (CPMd version which is designed to discover groups in directed networks) from CFinder (http://www.cfinder.org/) tool for k equals 3.

Figure 4 presents number of stable groups with their size in both datasets. One can notice that *Huffington Post* contains more groups overall. Moreover, the mentioned dataset comprises more small and medium size groups, but *Salon24* has more big groups.

In Figure 5 we can see the fraction of stable groups in relation to all groups. Stable groups have additional restriction that they have to be present in at least given number (in experiments we used value 3 due to the fact that the presence in 2 time slots is not hard to achieve because slots are overlapping) of time slots. One can observe that the lowest fraction of groups is stable for groups with small size and increases with group size. Furthermore, we can notice that *Salon24* has higher fraction of stable groups than *Huffington Post*.

Figure 6 depicts number of evolution events in both datasets. *Huffington Post* includes a large amount of medium size groups, so there are more events related to joining and dividing groups with similar size (i.e., *merge*, *split* events). Conversely, *Salon24* contains a relatively large number of huge groups, so in this dataset the events related to joining and dividing groups with substantial difference of size, that is, *addition*, *deletion* events, dominate over ones with similar size.

4.6. Reaction for Real-World Events. Figures 7 and 8 present number of groups and evolution events for *Huffington Post* and *Salon24* with marking key events from real world. One can notice some correlation between peaks on these charts

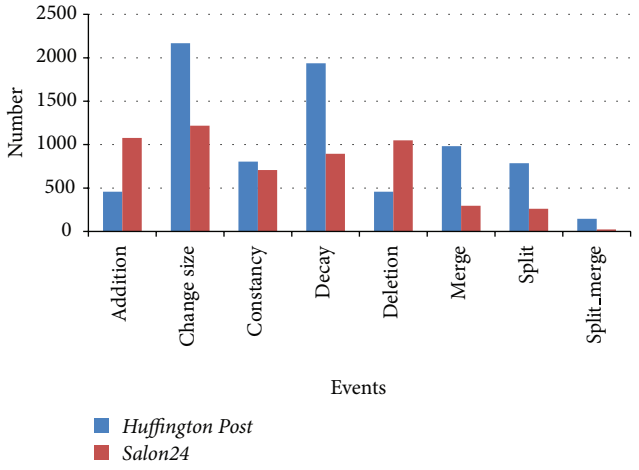

FIGURE 6: Number of events.

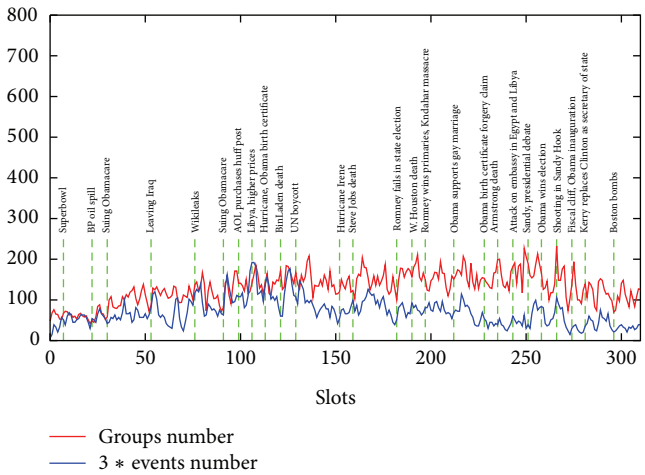

FIGURE 7: Number of groups and evolution events in time and correlation with real-world events for *Huffington Post*.

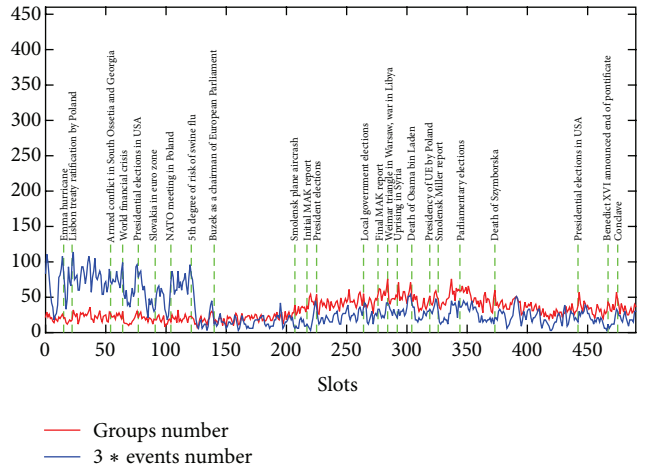

FIGURE 8: Number of groups and evolution events in time and correlation with real-world events for *Salon24*.

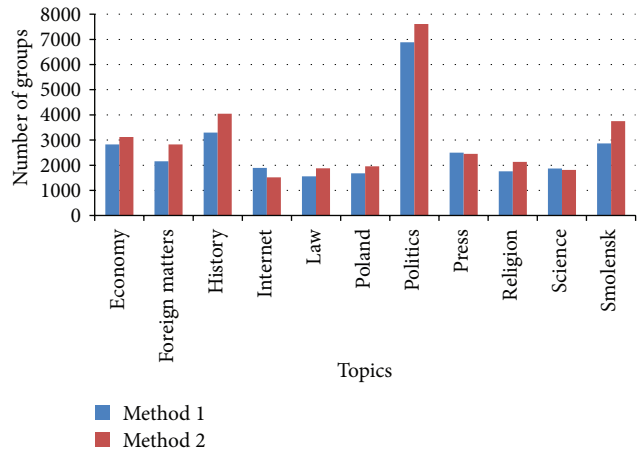

FIGURE 9: Topics discussed in at least 10% of all groups in *Salon24*.

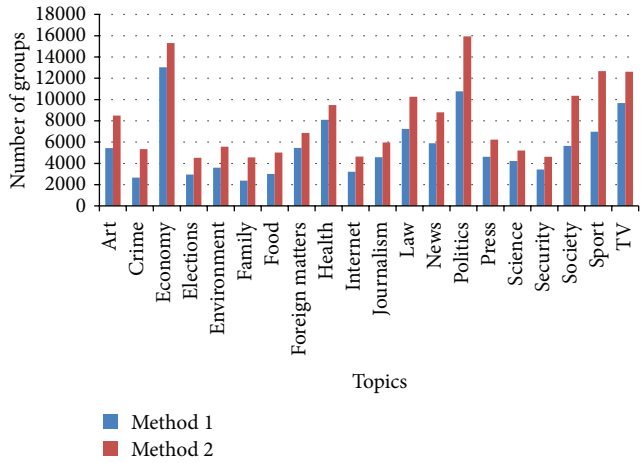

FIGURE 10: Topics discussed in at least 10% of all groups in *Huffington Post*.

and mentioned events. It means that blog portals are a kind of mirror that reflects actual events from real world and such events influence on groups in blogosphere to a large degree.

4.7. Topics in Groups. Figures 9 and 10 describe most popular topics discussed in groups in both blogospheres. Each chart presents topics being present in at least 10% of all groups. We used 2 methods to assess topics in groups, both described in Section 3.5. The motivation for introducing the second method was to determine topics for larger number of groups (e.g., in *Huffington Post* using the first method we assigned topics for about 75% of all groups and using the second method we assigned topics for about 93% of all groups).

One can notice that *Huffington Post* contains more different topics, but in *Salon24* one can observe that topics related to *politics* are dominating. Another interesting thing is the topic of *Smolensk* which appears frequent in groups in *Salon24* and it concerns Polish President airplane crash in Smolensk (10.04.2010) and other events related to investigation of this catastrophe.

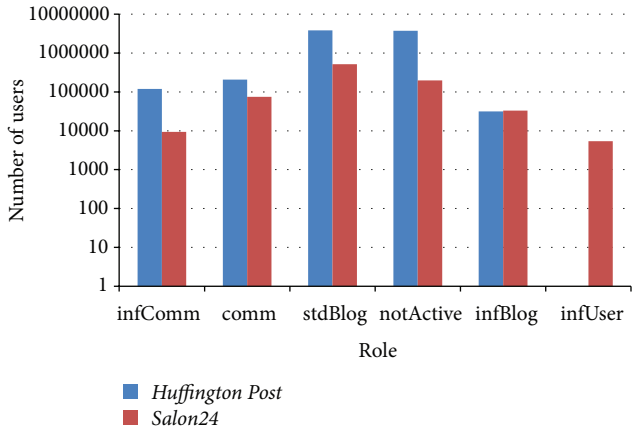

Figure 11: Global roles of users.

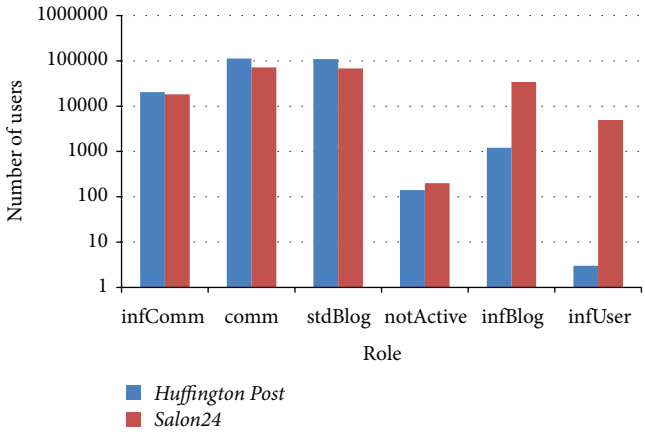

Figure 12: Local roles of users.

When we look into results of both methods to associate topics for groups, we can spot that they are quite similar (in terms of proportions for different topics).

4.8. Global and Local Roles of Users. Figures 11 and 12 show number of users with global and local roles (roles on the level of a group), respectively. For global roles, we can notice that in *Huffington Post* users with a role of *Influential User* (users with this role write influential posts and influential comments) almost do not exist (there is only one person with such a role), which is very different from *Salon24*. This difference can be explained by various types of nature of these portals—in American portal there is very small fraction of authors of posts and they rarely write any comments.

As far as local roles are concerned, one can notice a few interesting observations. Firstly, the number of inactive users is much lower than in previous case—this means that most inactive users (actually, the conditions in experiments let them write no more than one comment) are outside groups which is understandable. Moreover, the number of *Influential Bloggers* and *Influential Users* is smaller in American portal than in Polish one. The difference has its roots in different nature of portals (as we explained above) and the fact that in *Huffington Post* the responses to a post constitute a smaller fraction of all responses in comparison with *Salon24* (which can be seen in Table 1).

5. Conclusion

In the paper, a comparative analysis of two different blogospheres, Polish and American, is presented. This approach is based on a comprehensive analysis of the structure and content of blogosphere.

The preliminary analysis of the structure of both blogospheres shows that discussions conducted in *Salon24* are much more intense: generally the first comment appears much more quickly, but the lifetime of the post is much shorter than in *Huffington Post*. Discussions in *Huffington Post* are much more stable. The structure of groups is different: in *Huffington Post* there are smaller groups of comparable size, which is the reason why there are more events *split* and *merge* (characteristic of groups of similar size). In *Salon24* there is a greater variation in the group size and thus different events dominate (*deletion*, *addition*).

Differences in the number of these groups are significant: in *Huffington Post* there are three times more groups than in *Salon24*. A probable reason for this is a considerable difference in the ratio of the number of posts to the number of comments: in *Huffington Post* most people write comments, but very few write posts (for *Salon24* the situation is different).

In turn, events have a big impact on the dynamics of both blogospheres. Due to the different nature of *Huffington Post*, where few people write posts and most comments, some roles, which are in *Salon24*, in *Huffington Post* are not present. As far as topics discussed in groups are considered, *Salon24* is more oriented on topics related to politics, but *Huffington Post* is more diverse.

So, the comparison of two blogospheres gave interesting results: in some aspects nationality does not matter but sometimes has a big impact on user behavior. One can see differences in the characteristics of people from different countries in the context of their activity in the social media (taking into account their dynamic nature), for example, categories of interesting topics, speed of reaction to novelty, and way of reaction according to the categories of the world events. Presenting approach may have many practical applications. It can, for example, support sociologists and psychologists in their research on behavioral analysis in different national communities (e.g., among emigrants). The results of our experiments show that, for example, in marketing, making user profiles, one should take into account nationality, and therefore product marketing campaigns should be differentiated depending on countries (e.g., global advertising campaign). Similarly, to predict customer behavior, one should take into account the context of nationalities. These observations can be used in the development of election campaigns.

Research can be continued in several ways. One of them is analyzing and comparing differences in sentiment, for example, which nation is more optimistic? Another direction of research could be comparing the ability to predict the future

of groups in both blogospheres. Furthermore, extension of comparison to other national blogospheres possibly could reveal some characteristics related to their nationality.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The research reported in the paper was partially supported by the Grants nos. INNOTECH-K2/IN2/89/182461/NCBR/13 and 008/R/ID1/2011/01 from the Polish National Centre for Research and Development.

References

[1] M. Kobayashi, "Blogging around the globe: mitivations, privacy concerns, and social networking," in *Computational Social Networks: Security and Privacy*, A. Abraham, Ed., chapter 3, pp. 55–86, Springer, London, UK, 2012.

[2] S. Penderson, *Why Blog?: Motivations for Blogging*, Woodhead, Cambridge, UK, 2010.

[3] B. A. Nardi, D. J. Schiano, M. Gumbrecht, and L. Swartz, "Why we blog," *Communications of the ACM*, vol. 47, no. 12, pp. 41–46, 2004.

[4] K. D. Trammell, A. Tarkowski, J. Hofmokl, and A. M. Sapp, "Rzeczpospolita blogów [Republic of Blog]: examining polish bloggers through content analysis," *Journal of Computer-Mediated Communication*, vol. 11, no. 3, pp. 702–722, 2006.

[5] A. J. Gill, S. Nowson, and J. Oberlander, "What are they blogging about? Personality, topic and motivation in blogs," in *Proceedings of the 3rd AAAI International ICWSM Conference*, E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov, and B. L. Tseng, Eds., The AAAI Press, San Jose, Calif, USA, May 2009.

[6] M. Taki, *Bloggers and the blogosphere in lebanon & syria meanings and activities [Ph.D. thesis]*, University of Westminster, London, UK, 2010.

[7] P. J. Carrington, J. Scott, and S. Wasserman, Eds., *Models and Methods in Social Network Analysis*, Cambridge University Press, Cambridge, UK, 2005.

[8] D. Obradovic and S. Baumann, "Identifying and analysing Germany's top blogs," in *KI 2008: Advances in Artificial Intelligence*, A. Dengel, K. Berns, T. M. Breuel, F. Bomarius, and T. Roth-Berghofer, Eds., vol. 5243 of *Lecture Notes in Computer Science*, pp. 111–118, Springer, Berlin, Germany, 2008.

[9] S. C. Herring, I. Kouper, J. C. Paolillo et al., "Conversations in the blogosphere: an analysis 'from the bottom up,'" in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS '05)*, vol. 4, p. 107.2, IEEE Computer Society, Washington, DC, USA, January 2005.

[10] D. Obradovic and S. Baumann, "A journey to the core of the blogosphere," in *International Conference on Advances in Social Network Analysis and Mining (ASONAM '09)*, N. Memon and R. Alhajj, Eds., pp. 1–6, IEEE Computer Society, 2009.

[11] S. P. Borgatti and M. G. Everett, "Models of core/periphery structures," *Social Networks*, vol. 21, no. 4, pp. 375–395, 2000.

[12] H. Yilin, F. Caroli, and T. Mandl, "The Chinese and the German blogosphere: an empirical and comparative analysis," in *Mensch & Computer*, T. Gross, Ed., pp. 149–158, Oldenbourg, 2007.

[13] T. Mandl, "Comparing chinese and german blogs," in *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia (HT '09)*, pp. 299–308, ACM, New York, NY, USA, July 2009.

[14] P. Sobkowicz, M. Thelwall, K. Buckley, G. Paltoglou, and A. Sobkowicz, "Lognormal distributions of user post lengths in Internet discussions—a consequence of the Weber-Fechner law?" *EPJ Data Science*, vol. 2, no. 1, article 2, 2013.

[15] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3–5, pp. 75–174, 2010.

[16] D. Greene, D. Doyle, and P. Cunningham, "Tracking the evolution of communities in dynamic social networks," in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM '10)*, pp. 176–183, IEEE Computer Society, Washington, DC, USA, 2010.

[17] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[18] S. Asur, S. Parthasarathy, and D. Ucar, "An event-based framework for characterizing the evolutionary behavior of interaction graphs," *ACM Transactions on Knowledge Discovery from Data*, vol. 3, no. 4, article 16, 2009.

[19] M. Spiliopoulou, "Evolution in social networks: a survey," in *Social Network Data Analytics*, C. C. Aggarwal, Ed., pp. 149–175, Springer, New York, NY, USA, 2011.

[20] B. Gliwa, S. Saganowski, A. Zygmunt, P. Bródka, P. Kazienko, and J. Koźlak, "Identification of group changes in blogosphere," in *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '12)*, pp. 1201–1206, Istanbul, Turkey, August 2012.

[21] P. Bródka, S. Saganowski, and P. Kazienko, "GED: the method for group evolution discovery in social networks," *Social Network Analysis and Mining*, vol. 3, no. 1, pp. 1–14, 2013.

[22] G. Palla, A.-L. Barabási, and T. Vicsek, "Quantifying social group evolution," *Nature*, vol. 446, no. 7136, pp. 664–667, 2007.

[23] B. Gliwa, A. Zygmunt, and A. Byrski, "Graphical analysis of social group dynamics," in *Proceedings of the 4th International Conference on Computational Aspects of Social Networks (CASoN '12)*, pp. 41–46, IEEE, November 2012.

[24] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge,UK, 1994.

[25] E. Gleave, H. T. Welser, T. M. Lento, and M. A. Smith, "A conceptual and operational definition of 'social role' in online community," in *Proceedings of the 42nd Hawaii International Conference on System Sciences (HICSS '09)*, pp. 1–11, IEEE Computer Society, January 2009.

[26] H. T. Welser, D. Cosley, G. Kossinets et al., "Finding social roles in wikipedia," in *Proceedings of the iConference (iConference '11)*, pp. 122–129, ACM, New York, NY, USA, 2011.

[27] V. Junquero-Trabado and D. Dominguez-Sal, "Building a role search engine for social media," in *Proceedings of the 21st International Conference Companion on World Wide Web (WWW '12)*, A. Mille, F. L. Gandon, J. Misselis, M. Rabinovich, and S. Staab, Eds., pp. 1051–1060, ACM, 2012.

[28] E. Keller and J. Berry, *One American in Ten Tells the Other Nine How to Vote, Where to Eat and, What to Buy*, The Free Press, New York, NY, USA, 2003.

[29] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Modeling blogger influence in a community," *Social Network Analysis and Mining*, vol. 2, no. 2, pp. 139–162, 2012.

[30] R. D. Nolker and L. Zhou, "Social computing and weighting to identify member roles in online communities," in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 87–93, September 2005.

[31] A. Zygmunt, "Role identification of social networkers," in *Encyclopediaof Social Network Analysis and Mining*, R. Alhajj and J. Rokne, Eds., pp. 1598–1606, Springer, New York, NY, USA, 2014.

[32] D. L. Hansen, B. Shneiderman, and M. A. Smith, "Visualizing threaded conversation networks: mining message boards and email lists for actionable insights," in *Active Media Technology*, A. An, P. Lingras, S. Petty, and R. Huang, Eds., vol. 6335 of *Lecture Notes in Computer Science*, pp. 47–62, Springer, Berlin, Germany, 2010.

[33] M. Mathioudakis and N. Koudas, "Efficient identification of starters and followers in social media," in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology (EDBT '09)*, pp. 708–719, ACM, Saint-Petersburg, Russia, March 2009.

[34] C. C. Aggarwal and H. Wang, "Text mining in social networks," in *Social Network Data Analytics*, C. C. Aggarwal, Ed., pp. 353–378, Springer, New York, NY, USA, 2011.

[35] F. Bodendorf and C. Kaiser, "Detecting opinion leaders and trends in online communities," in *Proceedings of the 4th International Conference on Digital Society (ICDS '10)*, L. Berntzen, F. Bodendorf, E. Lawrence, M. Perry, and S. Smedberg, Eds., pp. 124–129, February 2010.

[36] A. Bartal, E. Sasson, and G. Ravid, "Predicting links in social networks using text mining and SNA," in *Proceedings of the International Conference on Advances in Social Network Analysis and Mining (ASONAM '09)*, pp. 131–136, IEEE Computer Society, Washington, DC, USA, July 2009.

[37] Y. Huang, "Support vector machines for text categorization based on latent semantic indexing," Tech. Rep., Electrical and Computer Engineering Department, The Johns Hopkins University, 2003.

[38] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, 2003.

[39] B. Gliwa, A. Zygmunt, and S. Podgorski, "Incorporating text analysis into evolution of social groups in blogosphere," in *Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS '13)*, pp. 931–938, Krakow, Poland, September 2013.

[40] M. Nguyen, T. Ho, and P. Do, "Social networks analysis based on topic modeling," in *Proceedings of the IEEE RIVF International Conference on Computing and Communication Technologies: Research, Innovation, and Vision for Future (RIVF '13)*, pp. 119–122, November 2013.

[41] A. Zygmunt, P. Bródka, P. Kazienko, and J. Koźlak, "Key person analysis in social communities within the blogosphere," *Journal of Universal Computer Science*, vol. 18, no. 4, pp. 577–597, 2012.

[42] B. Gliwa, A. Zygmunt, and J. Koźlak, "Analysis of roles and groups in blogosphere," in *Proceedings of the 8th International Conference on Computer Recognition Systems (CORES '13)*, vol. 226 of *Advances in Intelligent Systems and Computing*, pp. 299–308, Springer, Cham, Switzerland, 2013.

[43] B. Gliwa, J. Koźlak, A. Zygmunt, and K. Cetnarowicz, "Models of social groupsin blogosphere based on information about comment addressees and sentiments," in *Proceedings of the 4th International Conference on Social Informatics*, vol. 7710 of *Lecture Notes in Computer Science*, pp. 475–488, Springer, Lausanne, Switzerland, 2012.