Diversity-seeking users and their influence on social news sites

Jooyeon Kim
Department of Computer Science
KAIST
Daejeon, South Korea
jooyeon@weblab.t.u-tokyo.ac.jp

Joon Hee Kim, Dongwoo Kim, Alice Oh
Department of Computer Science
KAIST
Daejeon, South Korea
{joon.kim, dw.kim}@kaist.ac.kr,
alice.oh@kaist.edu

ABSTRACT

Social news sites where users actively engage in reading, discussing, and sharing news with their network can serve as a rich dataset for observing and analyzing the behavior of online social news consumption. In this paper, we combine machine learning and network analysis of users' textual contents and network characteristics to propose metric that measures user's degree of seeking diversity in a social new site. Our results reveal that the proposed metric serve to identify influential users who span structural holes and promote to create smaller information network. We discuss this result using a dataset of Huffington Post articles from the Politics section containing over 43,000 articles and activities of over 35,000 users.

1. INTRODUCTION

Tracking activities of users in a social news site is essential for news publishers to understand how the contents they made are spread and affect people. Among those users, some are more influential then others, and therefore, finding these users can ease the burden of analyzing the whole network to get feedback for the news contents.

In this paper, we propose a metric to measure users' pursuit of diversity, and investigate how the metric can be used to identify influential users in the network.

Specifically, we use topic model to infer users' topical interests and calculate their topical similarity with friends to define their degree of seeking diversity. Then, users with high degree of seeking diversity are defined as diversity-seeking users.

Then, we investigate how the metric can be used to identify influential users who span structural holes [3] in social networks who bridge dense clusters of strong connections and attain the benefits of accessing less-redundant information that pass through them. Also, we look at how diversity-seeking users benefit social news sites by increasing its connectivity; they tend to bridge alien articles to prevent the information network from being fragmentized.

NewsKDD '14 New York, NY USA

Table 1: Summary statistics of the users in our Huffington Post dataset. We filtered out users with 30 or fewer comments.

	Min.	Max.	Avg.	Med.
Before filtering				
# friends	0	7816	12.87	0
# comments	0	38108	92.67	3
After filtering				
# friends	0	7816	109.53	31
# comments	31	38108	578.82	202

2. DATA AND METHODOLOGY

We use data from the popular online news aggregator, the Huffington Post [1] where users can read and comment on news articles. They can also share articles via other social networking sites such as Facebook, and they can follow other users as on Twitter. When two users mutually follow each other, they become friends. So the social network on Huffington Post has both unidirectional followers as well as bidirectional friends. In this paper, we use only the bidirectional friendships in reconstructing and analyzing the social network for users' network and behavioral characteristics. Similar to the Twitter or Facebook news feeds, Huffington Post provides users with their friends' reading and commenting activities, so we can assume that friends influence one another's reading and commenting behaviors.

Our dataset consists of 43,957 articles in the politics section and activities of 338,897 users from May 10, 2005 to March 10, 2012. From that dataset, we chose only the users with 30 or more comments because users with fewer comments would result in less accurate topic analysis which we describe in the next section. Also, users with fewer than thirty comments are less likely to be sufficiently experienced on Huffington Post to behave in accordance to the general patterns of adding friends or making comments with meaningful content. Keeping only those active users, the filtered dataset contains 35,044 users, and we use this dataset for all of the analyses in the paper. Table 1 shows the detailed statistics about the users in our dataset. For the 43,957 articles, the average number of comments is 724.27, with the median of 117.

To measure topic similarity between users and figure out what topic distribution users have, we used LDA [1], which is widely used for topic modeling in the field of machine learning. First, we

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

Table 2: Examples of discovered topics from our Huffington Post dataset. There are 50 topics in total, and we show the top five words for each topic.

Topic #	Keywords
0	iraq, baghdad , iran, almaliki, alqaida
7	parenthood, abort, clinic, fetus
9	romney, herman, komen, mitt, cain
19	gay, samesex, marriage, lesbian
31	superdelegation, hillary, delegation
46	debt, ceiling, boehner, medicare

extracted top 10 keywords from each article, using the conventional TF-IDF:

$$w_{i,j} = tf_{i,j} \times \log \frac{N}{df_i} \quad (1)$$

for a term *i* in document *j*, where $tf_{i,j}$ is the number of occurrences of *i* in *j*, df_i is the number documents containing *i*, and *N* is the total number of documents. When selecting the number of keywords of each article, we also also tried using 5, 20 keywords for each article but showed no significant difference in the result. Based on the articles that users have commented on, we made a list of top 300 keywords for each user, by aggregating the keywords for each article and selecting top 300 ones sorted by TF-IDF value. If a keyword is overlapped among documents, we summed up the value for the keyword, and accordingly, no user had less than 300 keywords since we filtered out users with 30 or less comments. Then, considering each set of keywords as a single document for user with each keyword's term frequency being 1, we ran LDA with variational inference to infer a total of 50 topics, and defining each user's topic distribution θ_u. We set the number of keywords equal regardless of number of comments users have wrote, in order to suppress bias; the topic distribution of each user should have an information entropy which is irrelevant to how long the user have been involved in the community, and only dependent on the user's skewness of interest. As a result, 50 topics are extracted from LDA. Examples of extracted topics are listed in Table 2.

DEFINITION 1. ***Topic similarity (TS)*** *between two users i and j is measured by weighted cosine similarity [5]*

$$TS(i,j) \equiv \frac{\theta_i^t \Delta \theta_j}{\sqrt{\theta_i^t \Delta \theta_i}\sqrt{\theta_j^t \Delta \theta_j}}, \quad (2)$$

where θ_i, θ_j is user i, j's topic distribution and Δ is cosine similarity matric.

We used weighted cosine similarity in order to consider similarities between topics. In addition to weighted cosine similarity, we tried cosine similarity, and pearson correlation and we could draw similar results.

Users with high *social diversity (SD)* are the ones who do not have high degree of homophilic relationships, and have a balanced spectrum of friends in terms of similarity. We made two following standards for the diversity seekers to qualify: *a*) avoid many strong-similarity relationships; *b*) have various kinds of relationships in terms of similarity. For condition a), we calculate mean similarity with friends and subtracted the value by 1, and for condition b), we measure the standard deviation for similarities with friends. Social diversity is measured by multiplying the two values acquired by the two conditions, calculated as below.

DEFINITION 2. ***Social diversity (SD)*** *of user i: The difference between the average topic similarity with friends and the average topic similarity with random users as follows:*

$$SD(i) \equiv (1 - \overline{TS(i)}) \times \sqrt{\frac{1}{|f(i)|}\sum \left(TS(i,j) - \overline{TS(i)}\right)^2}, \quad (3)$$

where $f(i)$ is a set of friends of user i, and $\overline{TS(i)}$ is user i's mean TS with friends.

As a result, minimum, maximum, average, and median for all users' SD were 0.0052, 0.85, 0.56, and 0.59 respectively.

3. ANALYSIS

We assess various traits that diversity-seekers have on a social news site. Fragmentation in online social network is caused by natural human tendency for homophily. In a network with topical fragmentation, most neighbors share same interests, and provide little advantage in terms of information gathering to one another. Diversity-seeking users, however, would have more friends of different taste and will loosen the fragmentation in the network. In particular, we examine two major features of diversity-seeking users on the online social network. First, we measure the influence level of diversity-seekers inside and outside of their social group. Then, we show how diversity-seeking users promote to create smaller information networks, an equivalent role as weak-ties in social network.

3.1 Finding structural hole spanners

We investigate noticeable trait of diversity-seeking users. We can confirm that SD can be a useful measurement for identifying users whose influence span large range of domains, or structural hole spanners.

Experimental setup We quantify user i's influence Inf_i based on their commenting behaviors, calculated as below:

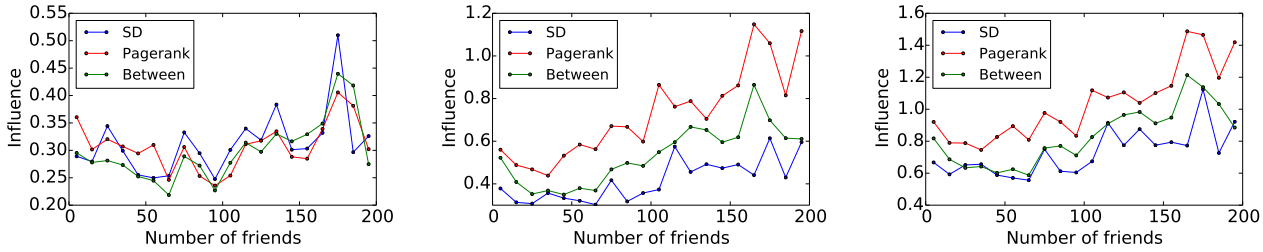
DEFINITION 3. ***Influence (Inf)*** *of user i:*

$$Inf_i = \frac{1}{K_i}\sum_{k}^{K_i}\sum_{j}^{Na_{ik}} \frac{1}{Nb_{jk}}, \quad (4)$$

where Na_{ik} is user i's number of friends who wrote comments after the user on article k, Nb_{jk} is the number of friends who wrote comments before the user on article k, and K_i is the number of articles user i wrote comments on.

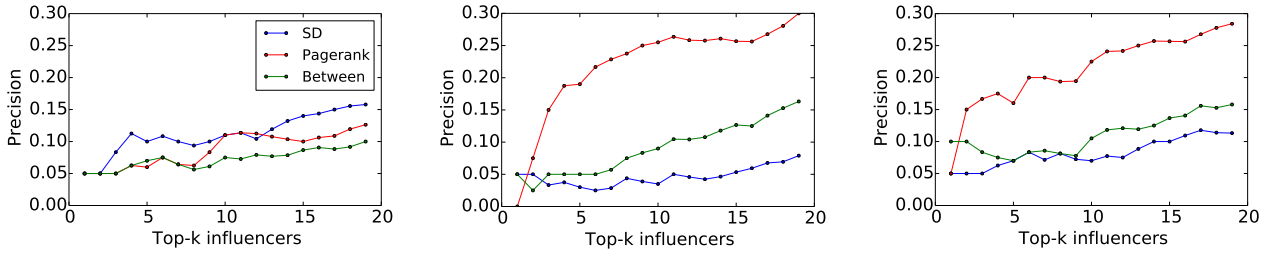
When a user writes a comment on an article, and if the user's friends write comments below him/her, we assume that the friends are influenced by the user who first wrote the comment. If there are multiple n users who are friends and have written comments above the targeted user, we assign each user with the influence score of $1/n$. We sum up the score for each article and return the mean value over the articles to come up with user's overall influence.

We adopted previously proposed method to identify influential users who span structural holes; we divided the entire social network into several domains, and for each user, define inner and outer domains. Users who have high influence on outer domains rather than on inner domain are defined as structural hole spanners [8]. In order to divide the network into multiple domains, we ran a community detection algorithm that maximizes modularity of the network [6] and use Louvain method [2] for optimizing the algorithm. As a result, a total of 14 groups were detected; 5 major



(a) Outer domain (b) Inner domain (c) Entire network

Figure 1: Users' quantified influence regarding (a) outer-domain, (b) inner-domain, (c) entire network. Points are the mean influence values of top 10% SD (blue), pagerank (red), and betweenness centrality (green), within each set divided by user's degree centrality.



(a) Outer domain (b) Inner domain (c) Entire network

Figure 2: Precision for identifying top-k users within each set using top-k SD (blue), pagerank (red), and betweenss centrality (green).

groups with over 1,000 members, and 9 minor groups with less than 10 members. Table 3 shows the general statistics of user influence regarding different domains. We can observe that user's inner-domain influence is larger than that of outer-domain, regardless of the fact that inner domains have smaller number of users than outer-domains.

Compared to SD, user's influence, pagerank, and betweenness centralities are strongly correlated with degree centrality. Therefore, in order to suppress the effect of degree centrality, we used users whose number of friends ranging from 0 to 200 to make 20 evenly spaced sets within which elements' number of friends differ less than 10. We did not sample users who have more than 200 friends, because the number of samples were small to make statistically significant correlations with user's influence. As a result, all 20 sets have at least 100 users, and elements' varying numbers of friends within the sets have little impact on user influence.

Comparison Methods We compare (a) social diversity (b) betweenness centrality and (c) pagerank [7] for influence propagation patterns. We have also looked for other measurements such as topic diversity, eigenvector, closeness centralities, fraction of triangles, and 2-Step Connectivity [8] but will skip mentioning the results since these have not shown competitive outcomes.

Results Figure 1 shows the mean quantified influence of top 10% users of social diversity, betweenness centrality and pagerank within each set, regarding different domain types. For entire network and in inner-domain, we can see that users with top pagerank scores have the highest influence, followed by top betweenness centrality users and top social diversity users. In outer-domain however, we can observe that for most sets influence of diversity seekers with top social diversity surpass that of top betweenness and top pagerank scorers.

In Figure 2, we conducted a task of identifying users with top-k influencers within each set, using top-k users with SD (blue), pagerank (red), and betweenness centrality (green). Points are the mean precision over all sets. As can be seen in the Figure, SD clearly outperforms pagerank and betweenness centrality for identifying top outer-domain influencers; in average, 30.7% better than pagerank and 52.4% better than betweenness centrality. In contrary, pagerank is much more effective than other two measure for finding top-k inner-domain and entire network influencers.

Table 3: Summary statistics of user influence regarding inner, outer-domains and whole network.

	Min.	Max	Avg.	Med.
Inner-domain	0.0047	13.60	0.43	0.26
Outer-domain	0.0025	11.06	0.33	0.20
Whole network	0.0170	20.60	0.76	0.50

Based on the results, we can sum up the experiment by stating that measuring social diversity for individuals does not help to target users with higher influence in their belonging communities and in overall networks; conventional measurements such as pagerank and betweenness centrality are more competitive. However, it can be a useful measurement for detecting structural hole spanners, or users with high outer-domain influences which cannot be detected by other measurements such as pagerank and betweenness centrality.

3.2 Connecting the articles

Weak ties promote to create social network more efficient in information propagation by serving as bridge between distant social

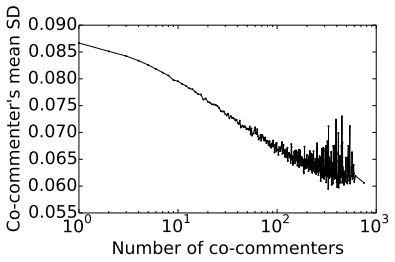
Table 4: Correlations between edge-betweenness centrality(EB), tie strength(TS), and SD of users consisting the tie.

	EB	TS	SD
EB	1.0	-0.050	0.031
TS	-0.050	1.0	-0.14
SD	0.031	-0.14	1.0

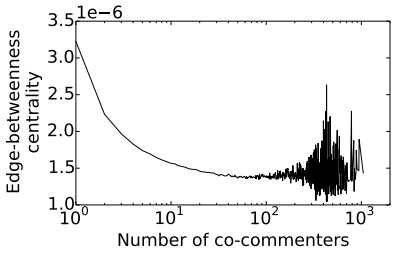
groups [4]. Diversity-seekers play similar role as they connect to diverse items, and help reduce gap among the interests of distinct social groups. We evaluate this roles of diversity-seekers as weak tie. We create a weighted information network *G(N, E)* where $N = \{n_1, ..., n_k\}$ is a set of news articles on Huffington Post, and we generate edge $e_{ij} \in E$ if a user comments on both articles. The weight of edge e_{ij} is equal to the number of users who commented on both articles.

First we check if diversity-seekers are found more frequently in weak ties. We randomly sample 2,000 nodes (articles) with 767,406 edges (existence of co-commenters). Each edge has a weight equal to the number of co-commenters. For each tie, we calculate average social diversity score of its co-commenters. Also we calculate edge-betweenness centrality of each tie. Then we compare these scores with the weight of edges. Figure 3 shows that users of high diversity-seeking behavior have a tendency to form weak ties in the information network. Also Figure 3 (b) shows that weak tie have higher edge-betweenness centrality, meaning that they are more important in information propagation.

Second we look at how the tie strength, user's topic and social diversity correlates with edge betweenness centrality in graph G. In Table 4, we can observe the negative correlations for edge-betweenness centrality with tie strength, and positive correlations with topic, social diversity of users consisting the edge. All the correlations are statistically significant, with p-value < 0.001. Therefore, it can be assumed that lack of diversity seekers in the community might lead the social network to be fragmentized, with a number of closed communities lacking interactions.



(a) SD



(b) Edge betweenness

Figure 3: X-axis is the number of users who wrote comments on both sampled pair of articles and y-axis indicates the mean diversity measurements of the co-commenters.

4. CONCLUSION

In this paper, we proposed a metric that measures user's degree of seeking diversity. This metric is successful in identifying structural hole spanners, and this result allows us to assume that diversity-seeking users have high influence that covers wide range of social network.

We combined user's content information as well as the egocentric network properties to come up with the metric that measures degree of seeking diversity. In terms of user's content information, we analyzed the articles he/she wrote comments on, but have not used the actual contents of the comments. Therefore, in future work, we would like to use user's comments and perhaps analyze the sentiments in the comments to understand how he/she thinks about the issues in the article. Another potential issue is designing our own probabilistic graphical model to generate user's topic distribution and topic diversity that purely depends on user's spectrum of interest.

5. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[3] R. S. Burt. The social structure of competition. *Networks and organizations: Structure, form, and action*, pages 57–91, 1992.

[4] M. S. Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

[5] B. Li and L. Han. Distance weighted cosine similarity measure for text classification. In *Intelligent Data Engineering and Automated Learning–IDEAL 2013*, pages 611–618. Springer, 2013.

[6] M. E. Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133, 2004.

[7] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[8] J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogenous networks. In *Proceedings of the fifth ACM international conference on Web search and data mining*, 2012.