Agenda Diversity in Social Media Discourse: A Study of the 2012 Korean General Election

Souneil Park[1], Minsam Ko[2], Jaeung Lee[3], Junehwa Song[3,4]

[1]School of Information
University of Michigan
bspark@umich.edu

[2]Dept. of Knowledge Service Engineering
KAIST
msko@nclab.kaist.ac.kr

[3]Division of Web Science and Technology
KAIST
leejai@nclab.kaist.ac.kr

[4]Dept. of Computer Science
KAIST
junesong@nclab.kaist.ac.kr

Abstract

The diversity of agendas in the media plays a critical role in shaping the scope of public discourse. Due to the increasing impact of social media, it is important and necessary to examine if it is circulating a wide range of issue and extending the realm of public discussion. In this paper, we put forward agenda diversity as a core value for evaluation of social media discourse and examine the current diversity status. Our study provides a firm basis for understanding the topics, quantity, diversity, and relative salience of the agendas offered by the social media. We further look into the information diffusion process from the diversity viewpoint, and explore the potential to increase diversity.

Introduction

Addressing diverse agendas in public discussion is an important element to the success of a democratic community. Exposure to diverse agendas expands individuals' consciousness of the community's problems and supports informed decision-making. It also provides the opportunity for constituents to understand other's different ideas and communicate them to the community. The constituents can further have a sense of inclusion when they experience that the community is considering the agendas of their interest, possibly encouraging them to be more open and reflect others' perspectives (Garrett 2005).

In this paper, we propose *agenda diversity* as a core aspect comprising the value of social media and investigate the current diversity status. The intensity and frequency of its use makes the diversity in social media play a critical role in shaping the scope of users' world view, telling them "what to think about" (McCombs and Shaw 1972). The extensive scale of adoption and the density of connection broaden the impact of the diversity in the discourse. Considering the increasing influence as a news delivery channel (Pew 2012), it is necessary to question if social media is covering a variety of agendas and extending the realm of public discussion. Yet, related exploration is limited to the agenda-setting research of communication studies that focuses on traditional mass media. The nature of social media is fundamentally different from the traditional media in all aspects, from agenda evaluation to selection and distribution, and requires new investigations.

We conduct a study of agenda diversity in Twitter, targeting the 2012 Korean general election, a contentious period when many groups competed to set the agendas and when the attention to public issues was high. The study provides a firm basis for understanding the topics, quantity, diversity, and relative salience of the agendas offered by the social media. The key observation of the study is agenda concentration; the main agendas in social media were limited both nominally and topically, hence likely to provide users with fewer number of and topically homogeneous issues than major news outlets. Nearly 70% of the tweets were concentrated to a few contentious issues. Social media also amplifies the gap of attention to different agendas more than the mass media and rarely allows less popular issues to reach a substantial number of audiences. In contrast to our expectation about the numerous different users who may have different issues of their interest, they collectively skew the circulation to a few items.

We further investigate the information diffusion process from the diversity perspective. Although many works have analyzed information diffusion in social media and made progress in understanding diffusion models and relevant factors, the meaning of the diffusion in terms of diversity is an important aspect that remains less explored. Our analysis comprehensively covers the diversity aspect. First, we track the intensification of agenda concentration and

show how diversity gets lost along the diffusion process. We then look into a major underlying cause of concentration, i.e., tendency toward popular issues, and examine its strength in shaping the overall information space. Lastly, we explore the potential to increase agenda diversity and provide implications for diversification strategies.

Background and Related Work

With the growing impact of social media on mass communication, HCI/CSCW researchers are increasingly considering the values of public communication, such as information credibility (Morris et al. 2012), deliberation (Kriplean et al. 2012), usefulness (André et al. 2012), and quality (Diakopoulos et al. 2011). They use these values as important means to understand the discourse of the media and also as design goals in building new communication interfaces.

In addition to these values, our work introduces 'agenda diversity' as a new perspective to the discourse, which is connected with the core feature of social media. A major characteristic of social media is in connecting people of different ideas, interest, and perspectives in an unprecedented scale and enabling them to reach a wide range of audiences. We attempt to understand how successfully social media is incorporating different voices and extending the realm of public discussion by observing the diversity in the discussed issues.

Agenda setting and moderation is an essential role in nearly all kinds of discussion and decision-making process. At a very high-level, there are three classes of approaches to the role: organizing delegates for the role (e.g., independent external experts), fully opening the role to any participant of the discussion (e.g., town hall meetings, referendum), or combining the two former approaches. Agenda setting and moderation becomes more challenging as the scale of a discussion gets larger and more diverse people participate. As for national level debates or decision making, the role is mainly played by the mainstream media, which acts as professional delegates. However, the wide adoption of social media is enabling the fully opened models for large-scale discussions: any person can prioritize, select and disseminate issues he or she thinks important and wants to discuss. The resulting set of agendas in social media can be different from the mainstream media as users have different characteristics or considerations: many of them are not trained journalists, and may care less about the scale of potential audience and free from market related problems.

The importance of social media as a space for news distribution and public discussion has already been discussed in many works. Kwak et al. (Kwak et al. 2010) conducted a study with an early stage of Twitter and observed that it shows characteristics as an information dissemination medium than a social networking service. Lerman (Lerman 2010a) focused on this information dissemination function and observed that the spread of news is very similar to social news aggregators. In addition, a recent survey from Pew Research (2012) shows that the number of people who see news on social media increased from 9% at 2010 to 19% at 2012, and who regularly get news through social media from 7% to 20%. A number of works used social media to capture the public response to newsworthy events (Starbird et al. 2012; Shamma et al. 2011).

Information diffusion in social media has been actively researched in many works. They conducted large-scale measurement and developed diffusion models (Wu et al. 2007), prediction methods (Bandari et al. 2012), and identified related factors such as network structure (Lerman et al. 2010a), user influence (Lerman et al. 2010b), and novelty of information (Wu et al. 2007). Asur et al. (2011) studied top trends in social media and observed they are strongly influenced by news media, and social media contributes to the selection of important news. Our work views the diffusion specifically from the diversity perspective and extends the understanding made in these prior works.

Our work is closely related to the agenda diversity research in communication literature which concerns how many issues are considered salient by society, and how diverse those issues are (Allen and Izcaray. 1988). The primary interest of the agenda diversity research is in the study of media effects on supporting a community to consider a greater number of social problems as important. The agenda diversity is often measured by asking the Most Important Problem (MIP) question, which asks, "What do you think are the most important problems facing this country?" and repeatedly asks "Any other important problems?" The research examines the effect of relevant media factors such as different media environment (e.g., number of subscribing media outlets), media usage pattern of individuals (e.g., frequency of exposure to the media). As the research mainly dealt with the traditional news media, the advent of social media raises new questions and opens interdisciplinary research opportunities.

One line of relevant research is the studies of political or ideological diversity of the web. Investigations included the analysis of social network between people with opposing ideological views (Conover et al. 2011), identification of opposing views from web contents (Zhou et al. 2011; Park et al. 2011), development of interfaces for exposure to diversity (Faridani et al. 2010; Park et al. 2009), and the response to different perspectives (Munson and Resnick 2010). These works generally interpreted diversity based on opposing political views or pro vs. con

(Kriplean et al. 2012). We interpret diversity from a different perspective and add a different dimension to diversity analysis. We focus on the range of issues discussed in public debates rather than on the different viewpoints to a discussed issue. For example, a discussion may expose different viewpoints to an issue such as health care but may not address environmental issues.

A few works explored topic diversity in social websites. Choudhury et al. (2011) studied diversity in a real-time search context and discussed implications of including diverse items in search results. Sideline (Munson et al. 2009) addresses a closely related topic, that is, to increase diversity in social recommenders. The work interprets diversity as considering the different preferences (votes) of participants. It proposes diversity metrics and an algorithm for creating a top-k list considering the proportion of people whose preferred item is included in the list. The diversity in different preferences can be potentially related to agenda diversity depending on the voting pattern of individuals. For example, if there are many different groups of users who vote-for or re-tweet different issues, reflecting the different interests of the groups will lead to an inclusion of diverse issues; however, if the votes or re-tweets are concentrated to a small number of issues, inclusion of only a few agendas will cover the interests of a large proportion. Exploring the diversity of social media users and applying the idea of Sideline can be an interesting future work.

Agenda Diversity in Social Media

In this section, we describe our study of the diversity in social media discourse. With the term 'agenda', we indicate the topics of public issues, following the definition of agenda-setting research (McCombs and Shaw 1972). An agenda can be a story related to specific issues, public figures or organization. Similar to agenda research of communication research, we are mainly interested in the salient topics of the media, not the entire body of available topics. In order to understand the media influence on large-scale public discourse, it is important to focus on the topics that are frequently and prominently covered, which are the ones subscribers are likely to discover. It is difficult to assume that the topics of the tail reach many people and make big impact. We thus measure the diversity in the salient topics, not that of the entire topics.

The diversity is observed from multiple perspectives considering the electoral context. We first observe the diversity in the discussed public issues in general, and then specifically in the coverage of the election and the candidates. All the analyses are also conducted with the mainstream news media to better understand the diversity status of social media through comparison.

2012 Korea General Election

We first briefly describe the Korean general election held on April 11 2012 to provide the context of our study. The election organized the 19th national assembly, which consists of 300 seats, 246 directly elected seats and 54 nationwide proportional representation seats. The voters cast two votes, one for the candidate of their district and one for the party they support. The directly elected seats are filled with the winners of the 246 single-member districts and 1,114 candidates competed for the seats. 54 seats are allocated to parties according to the proportion of votes received and the parties listed 188 candidates. The official campaign period was from March 29 to April 10 and four parties won seats in the election.

Election periods provide a useful environment for agenda diversity research (McCombs and Shaw 1972; McCombs and Zhu 1995). During the period, policy debates are made over public issues of various areas and are extensively covered in the media. Many agenda-setting research works also considered this point and conducted studies during the election periods.

Data Set Development Method

Our analysis requires a dataset that captures the discussion of public issues in Twitter. For dataset development, we used two approaches, one with news articles and one with candidate names. The first approach considers the characteristic of Twitter as a news dissemination medium. It identifies the discussed issues by observing Twitter users' response to the news articles published during the election period. This approach treats the news articles as a knowledge base of public issues and collects the tweets that include links to the articles in order to identify what issues are picked and discussed in Twitter. Similar approaches were effectively used in other Twitter analysis works (Bandari et al. 2012).

The second approach collects the tweets that include candidate names. The candidates are the core and symbolic elements of the election, and their name can serve as a space for observing the distribution of public attention. This approach also partly addresses the limitation of the first approach, which is that it cannot capture the discussion of the issues that are not linked to news articles. We describe the datasets later in more detail together with the analysis results.

We also considered several other approaches. Keyword cue-based collection (e.g., hashtags or trending terms) is often used (Conover et al. 2011; Starbird and Palen 2012) to identify popular issues and collect related tweets; however, our interest was to have an overview of the discourse rather than to focus on a number of selected popular issues. In addition, the effectiveness of hashtags was limited for our analysis. As reported in Hong et al's

work (2011) hashtags are less popular in Korean compared to English (the use was even less in our own measurement).

Mining topics or issues from a large collection of tweets is another promising approach and is being actively researched in the data mining and information retrieval area (Bernstein et al. 2010; Ramage et al. 2010). Despite recent advances, we did not adopt it for our study since topic identification or topic-based clustering of tweets is in the experimental stage, especially for our target language, Korean. We describe the data set in detail together with the analysis result below.

Diversity in the Coverage of General Public Issues

We first take a comprehensive view and observe the attention to issues regardless of their topic category. Issues of various areas other than politics and elections, such as education or environment, can also have impact on people's view of the world and be considered in their decision-making process.

We analyze the diversity using the two metrics, that is, nominal diversity and thematic diversity. These two metrics are adopted from the agenda diversity research of the communication studies (Allen and Izcaray 1988). Nominal diversity is defined as the number of issues considered salient. Thematic diversity concerns the semantic variety of salient issues, and it is measured by counting the number of thematic categories to which the issues belong. We applied the issue category commonly used in agenda diversity research, which is composed of the following categories, "jobs/unemployment, welfare, money, public spending, law and order, government/political decision making, social relations, environment/food, technology/research, N. Korea-related problems (adapted from 'EU-related problems'), foreign policy, miscellaneous" (McCombs and Zhu 1995).

- **Data set 1. Tweets to General Coverage (TG set):** This data set is made with the news articles of a selected source. The selected source is Hankyoreh, which is the most commonly cited news media in Korean tweets. The scope of the collection was confined to the articles of this source in order to conduct a qualitative coding, i.e., manual identification of the covered issue and its thematic category. The coding was conducted by two researchers (We also checked the reliability of coding: the inter-rater agreement was measured with 100 sample articles and the kappa value was over 0.88.).

  We collected 1,105 articles published by the source during the election period. Although the selected source cannot capture the all topics of Twitter, the popularity of it enables a good sampling of important public issues, and it is possible to observe the distribution of attention to them. More concrete analysis would be possible if we can expand the coverage to many news sources; however, it requires a scalable solution to annotate the articles with the covered issue and their thematic category.

  The tweets linking to these articles were collected via Tweetmix (http://tweetmix.net), which aggregates Korean tweets based on the address of the included link (Shortened URL is also addressed by restoring the original address). We crawled all tweets linking to the collected news articles, their posting time, user's ID and profile information. We collected 55,292 tweets from 16,659 users.

We first identified the salient issues for each day of the election period. For the identification, we selected the articles that received the most tweets and identified the issues covered by them. The articles were selected until the tweets made to them exceed 70% of the tweets made to the entire articles of that day. Eight articles were selected per day, on average, and all the selected articles were those which received more than a hundred tweets. We then aggregated the tweets made to the articles of the same issue, and calculated the number of tweets made to each issue.

Concentration to Narrowed Range of Agendas

Figure 1 shows the proportion of tweets made to the issues identified. Each segment in a bar represents one issue and the segments of a bar are ordered by the number of tweets received. The figure shows that near half of the tweets are centralized to the top two issues throughout the election period, and that the proportion of tweets made to other issues rapidly decreases. The number of the most salient issues receiving over 70% of the tweets was 3.3 per day on average. The salient issues were mostly the issues highly disputed by the political parties, directly related to the election or the president. These issues include the illegal surveillance incident, past raunchy remarks of the candidate Kim[1], polling reports, and coverage of party leaders.

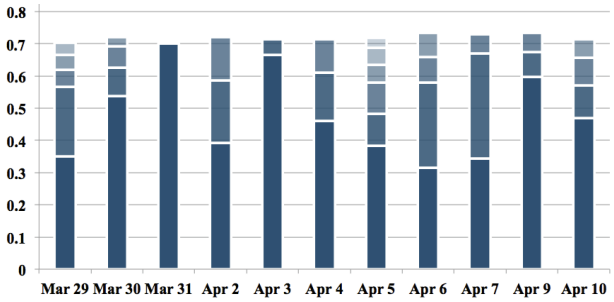

Figure 1. Proportion of tweets to salient issues

Similar to the nominal diversity, the salient issues are concentrated to a few thematic categories (See Figure 2). Each segment in a bar represents one thematic category. As tweets are centralized to a few political issues that are

[1] http://www.koreatimes.co.kr/www/news/nation/2012/04/116_108429.html

directly related to the election, the two most popular thematic categories are 'government/political decision making', which includes the articles covering the election, and 'law and order', which includes those of the illegal surveillance incident.

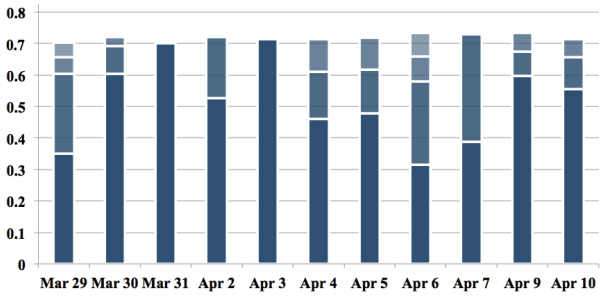

Figure 2. Proportion of tweets to thematic categories

Comparison of Diversity to Mainstream News Outlet

Though it is difficult to conduct an objective comparison between the two media with different characteristics, we attempt to study the difference of diversity in the social media's pick of the issues and that of the mainstream news media (the Hankyoreh). For the salient issues of the news outlet, we take advantage of the 'integrated section.' The mainstream news outlets in Korea generally include the integrated section every day and use the space for the issues of the day. It occupies about the first 30% of the pages though it slightly varies depending on the news items. The news producers put the articles they think the most newsworthy without distinction of topic category. Thus, we consider the issues covered in the articles of this section as the news producers' selection of the 'issues of the day.' They usually cover a number of issues in detail with many articles throughout multiple pages. There are cases when the whole section is filled with the articles of the same issue when there is a very important event.

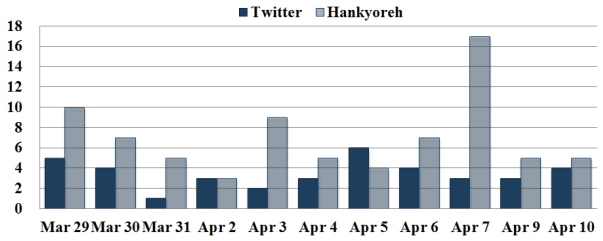

Figure 3. Number of salient issues of the two media

Figure 3 compares the number of salient issues of the two media. The news outlet generally covers more number of issues, four to six issues for a day. The issues range from the specialized coverage of parties' policies, environmental issues, international trade, and diplomatic issues. Although these issues were saliently presented in the news, they were not circulated much in social media.

The only exceptions are observed on Apr. 2nd and 5th. On both days, the news media focused on a few specific issues and produced many articles on them throughout the integrated section, hence the nominal diversity dropped. The reversal on Apr. 5 is partly because the news media focused on the illegal surveillance incident issue; however, two columns and one editorial that covered a different issue from the articles of the integrated section received more than a hundred tweets and were selected as salient articles of the day. While these articles increased the nominal diversity for the social media, they did not contribute to that of the news media as columns and editorials are located at the end of the newspaper.

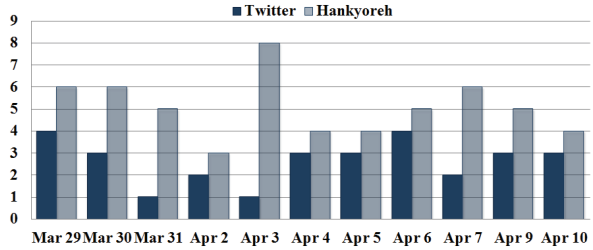

Figure 4. Number of thematic category of the two media

In terms of thematic diversity, the news media consistently covers a broader range of thematic categories (See Figure 4). The news media additionally covered the categories, 'environment/food', 'foreign policy', 'welfare', to name a few. It is also important that although the nominal diversity of the news media is lower on Apr. 5, the range of the covered thematic categories is broader.

We do not claim that the above comparison is the correct and fair way of comparing the two different media; however, we believe our comparison provides a feasible option since it reflects the media's judgment of relative importance of the issues. In addition, the comparison is somewhat generous to social media since its salient issues are selected from the 70% of the tweets whereas those of the news media are selected from the first 30% of the pages.

Diversity in the Coverage of Politics and Election

We narrow down to the articles on politics considering the electoral context of the study. Political discourse is an integral part of public communication during the election period as the media and the public pay much attention to it. Although we narrow down to the political discourse, there is much space to explore the diversity as the discourse can involve many different issues such as different candidates, agendas and policies of the parties, and important agendas during the past term of the administration or the parliament.

- **Data set 2. Tweets with Candidate Names (TC set):** We collected the tweets that include a name of a candidate regardless of the inclusion of links to news articles. Using the names of all 1,114 candidates as queries, we collected the tweet search results via Topsy (http://topsy.com). We collected 495,653 tweets returned by the search engine.
- **Data set 3. Tweets to Politics-Election Coverage (TP set):** We first collected the news articles on politics or the election through a news search engine. Using very general keywords (e.g., "candidate", "president office", "election", names of the parties, and so on) that frequently appear in such articles as queries, we crawled all the news articles returned by the news search engine. We collected 42,638 articles published by 134 sources during the election period. The tweets linking to these articles were collected in the same manner as in the TG set. We collected 165,678 tweets from 30,214 users.

Dominance of Major Candidates and Hot Issues

First, we observed the discourse through the candidates of the election. As mentioned, the candidates can serve as a space for observing the distribution of public attention. Since many issues are related to the candidates during the election period, observing the distribution of attention over the candidates also helps understand the amount of attention given to the issues. We measured the frequency of each candidate's name in each media. As for the social media, we measured it from the TC set. As for the news media, we measured it from the 42,638 news articles on politics (those used to create the TP set).

We observed that both types of media tend to extremely concentrate on a few nationally recognized candidates although there are more than a thousand candidates. More importantly, the degree of concentration was higher in social media. Figure 5 illustrates that the frequency of the candidates' name follows a power law distribution in both media, and the slope is steeper for social media. For example, while the top-twenty candidates of the news media appeared in nearly one third of the coverage, the top-five candidates of the social media appeared in one third of the tweet. These top candidates are usually the leaders of the parties, potential presidential candidates, and their competitors.

The ranking of the candidates was also highly similar between the two media. For each media, we ranked the candidates based on their frequency, and compared the ranking between the two media using the Kendall τ rank correlation coefficient. The measure ranges between −1 and 1, where the value 1 represents that the two rankings are completely identical and -1 represents a complete reversal. If there is no correlation between the rankings, the value becomes 0 (Lapata 2006). The ranking was considerably similar and the Kendall's τ value was 0.576 ($p < 0.01$).

Though it is not enough to make causation claim, the result provides some evidence for the hypothesis of agenda-setting research (McCombs and Shaw 1972) about the influence of the news media to the salience of topics on public agenda. It is expected that the news media is likely to focus on popular candidates because it has to reach out as many audiences as possible. However, social media users are under different conditions from the press in terms of readership or market concerns. The similarity in the coverage in spite of this different media characteristic implies the influence of news media.

Another related observation is that the candidates who did not raise much attention in the news media are usually not popular in the social media as well. Except a few politicians who have a large number of followers and have been very popular in Twitter for a long time, the candidates who are not covered in the news media are not highly ranked in the social media. An alternative interpretation would be that the social media influences the news media and leads the coverage of the candidates; however, we could not observe a clear evidence for this interpretation. It was difficult to find a case where the social media give much attention to a candidate ahead of the news media.

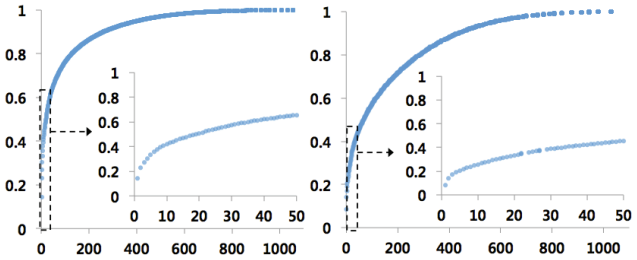

Figure 5. Cumulative distribution of candidate mentions (left: Twitter, right: news outlets)

Secondly, we observed the discourse through popular issues of the election, specifically, the degree of concentration of the media to these issues. We manually compiled a list of issue keywords and observed their frequency in each media. As it is difficult to cover all political issues, the issue keywords were made for top twenty popular issues, including the illegal surveillance incident, past raunchy remarks of the candidate Kim, North Korea's rocket experiment, and so on. As for the news media, we measured the number of news articles that include the issue keyword; as for the social media, we counted the tweets made to the articles including the issue keyword.

Similar to the candidate coverage, both media concentrated on a few issues and the concentration was higher in social media. Specifically, the most popular issue

covered 12% of the articles of the news media and 39% of the tweets. The top five issues covered 21% of the articles and 61% of the tweets. The top five issues of each media largely overlapped (four out of five). The illegal surveillance incident was the most popular issue in both media, followed by the issues about the candidate Kim's past remarks, Korea-U.S. free trade agreement, and the four major rivers project. The ranking of all twenty issues was also similar, resulting in the Kendall's τ value 0.568 ($p < 0.01$).

Agenda Concentration: The Diversity Aspect of Information Diffusion

Our previous analyses show that social media amplifies the gap of attention to different agendas more than the news media, and results in offering only a few narrowed range of issues to users. In this section, we extend the analyses to obtain a deeper understanding of the information diffusion process in terms of diversity.

We traced the diffusion of the articles on politics with the following features.

- *Import count:* The number of users who brought an article into the Twitter network.
- *Retweet rate:* The ratio of users who retweet the article among the total number of users the article reached. It is calculated by dividing the number of retweets of an article by the sum of the followers of those who retweeted it.
- *Total tweet count:* The ultimate number of tweets an article receives.

Filtering from the Import Stage

We observed that the skewed attention of social media originates from the start of the propagation, i.e., certain articles are much more frequently imported than others. A correlation analysis of the import count and total tweet count showed that the two features are significantly correlated (0.871, $p < 0.01$). We also analyzed the similarity of two article rankings, a ranking based on the import count and that based on the total tweet count. The Kendall's τ value was 0.817 ($p < 0.01$), which indicates that the articles that are more frequently imported are tweeted more ultimately.

The measurement of the retweet rate also supports that the number of initial propagation paths (import count) is critical for wide dissemination. The retweet rate is generally lower than 0.1% (0.03% on average) and its correlation with the total tweet count is low. Low retweet rate indicates that the propagation path of an article does not diverge much after being imported to the network. In addition, the time is limited for an article's propagation path to extend and diverge. On average, 53.4% of the total tweets to an article are made within one hour after publication, and 71% of them are made within five hours. Considering these constraints, the dissemination of an article is likely to be determined by the number of importers, who open initial propagation paths.

Intensifying Concentration through Retweets

After the import stage, the gap of circulation between popular items and unpopular items grows further and, consequently, salient issues narrow down to a few.

An important cause of the this concentration is the strong, general retweet tendency toward popular items. In order to observe the strength of this tendency, we analyzed the tweet history of 30,214 accounts that tweeted at least one article on politics during the election period. We see how many users mainly tweet popular stories.

For each account, we calculate the ratio of tweets made to articles of different popularity. The articles are classified into three classes according to their popularity: popular (P), moderately popular (MP), and unpopular (UP). Articles belong to the P class if the z score of the number of received tweets is greater than $z_{0.25}$ (which was 77), to the MP class if the z score is between $z_{0.25}$ and $z_{0.5}$ (which was 19), and to the UP class otherwise. The accounts are represented in a three dimensional vector according to their tweet ratio of the three classes. For example, if a user only tweeted popular stories, the vector representation is (P: 1, MP: 0, UP: 0).

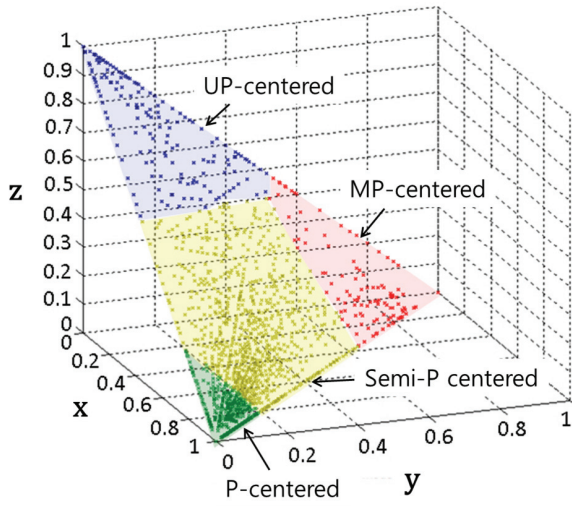

Figure 6. Distribution of accounts (x: popular, y: moderately popular, and z: unpopular)

Figure 6 shows the distribution of accounts in the three-dimensional vector space. It shows that a large proportion of the accounts are located near (P:1, MP:0, UP:0), demonstrating the strength of the tendency of tweeting popular stories. If we perform K-means clustering of the accounts in this vector space with k set to 4, the cluster which includes the point (P:1, MP: 0, UP:0) forms the majority, including 63% of the accounts (the green

segment in Figure 6). We refer to this cluster as the P-centered group. On average, 98% of the articles tweeted by an account of this group were popular stories. The cluster next to the P-centered group (which we call semi-P-centered) is less concentrated to popular articles; however, the figure shows that many of its member accounts are located near to the P-centered group. The group includes 18.9% of the accounts, and 59% of the articles tweeted by these accounts are popular stories.

Though the size of the group is not large, there are two groups that show different characteristic, the UP-centered group and the MP-centered group. We manually inspected a number of accounts of both groups and found that these groups commonly include unpopular candidates who disseminate news about themselves, unpopular news outlets, and some accounts that repeatedly post articles on a specific candidate or party (the candidates and the news outlets of the MP-centered group were relatively more popular ones that those of the UP-centered group). The UP-centered group includes 8.41% of the accounts, and the MP-centered group includes 9.64%. These accounts rarely disseminate popular stories. About 3.9% of the articles tweeted by these groups were popular stories.

Impact of P-centered Groups on the Information Space

We observed that the users of the P-centered and the semi-P-centered group have collective impact in deciding the salient issues of social media not only because they have better sense of potential popular items but also maintain relationships through retweeting others. Having active interaction with others was the key difference of the P-centered and semi-P-centered groups from the MP-centered and UP-centered groups. The MP-centered and the UP-centered groups have weak impact as they put less effort in building relationships and care less about what people may be interested.

The following features were used to identify the difference of the P-centered groups from the others.

- *# of followers*: number of followers of an account.
- *# of tweets*: number of tweets of an account
- *# of references*: number of references in others' tweets.
- *# of imports*: number of importing a story from outside.
- *Import ratio*: dividing the *# of imports* by the *# of tweets*.
- *# of spreading*: the number retweeting an article inside the Twitter network.
- *RT impact*: dividing the *# of references* by the *# of tweets*.

A linear discriminant analysis was conducted to identify the features that effectively distinguish the P-centered groups. The import ratio feature showed the highest correlation with the discriminant function whereas the other features showed less correlation. For example, the classifier distinguished the P-centered group from the MP-centered group with 70% accuracy, and the import ratio feature (# of imports / total tweets) showed the highest correlation with the discrimination function (corr. = 0.809). It indicates that the users of the P-centered group frequently retweet others' tweets, and those of the MP-centered group focus on importing stories even if their stories are not picked and disseminated by others. The result is similar for distinguishing the P-centered from the UP-centered, the semi-P-centered from the MP-centered, and the semi-P-centered from the UP-centered.

Possibility of Diversification

Lastly, we attempt to examine the potential for diversification of agendas in social media. We explore whether there are people who pursue diversity by tweeting articles of various topics. For each account, we measured the diversity based on the articles the user tweeted.

We defined the diversity measure as the average of topic differences between the articles tweeted by an account. The measure increases if an account (re)tweets articles with different topics and decreases if the articles cover similar topics. We adapted the calculation process used in (Livne et al. 2011). We first construct a word-based language model (LM) for every article tweeted by the account. The difference of topic between two articles is calculated by applying the Kullback-Leibler (KL) divergence on the LM of the articles. The topical difference is calculated for all pairs of articles, and the average is taken as the diversity measure of that account.

The LM of an article is constructed as follows. Given an article a, the distribution $P(t|a)$ is made by computing the weights for all words t in the vocabulary, V, of the corpus.

$$P(t|a) = (1 - \lambda)\, tfidf^N(t,a) + \lambda P^N(t|D)$$

The term $tfidf^N(t,a)$ is the normalized tf-idf value of t in a. The term $P^N(t|D)$ is used for smoothing the weight of the words that do not appear in a. It is the normalized value of the marginal probability of the term in the collection D, i.e.

$$P^N(t|D) = \frac{P(t|D)}{\sum_{t \in V} P(t|D)}, \quad where\ P(t|D) = \overline{tf}(t,D) \cdot df(t,D).$$

The term $\overline{tf}(t,D)$ is the average frequency of t in the collection D, and $df(t,D)$ is the number of the articles including the t. The smoothing factor λ is set to 0.001.

We use $P^N(t|a)$, the normalization of $P(t|a)$, as the LM of a.

$$P^N(t|a) = \frac{P(t|a)}{\sum_{t \in V} P(t|a)}$$

Given two articles a_1 and a_2, the KL divergence measures the difference in the distribution of their LM.

$$D_{KL}(a_1, a_2) = \sum_{t \in V} \left[P^N(t|a_1) \log \frac{P^N(t|a_1)}{P^N(t|a_2)} + P^N(t|a_2) \log \frac{P^N(t|a_2)}{P^N(t|a_1)} \right]$$

We observed the divergence measure for a number of sample article pairs; as for the article pair which covers the same topic, the value was usually less than 20, ranging between 10 and 20; for those covering different topics, the value was greater than 20, ranging between 20 and 30.

We observed that there are a small segment of people (near 4%) who circulate many articles of different topics. The spike in Figure 7 shows that the diversity measure of the accounts that actively disseminate many articles ranges between 25 and 30. Among 1,272 accounts that circulated more than 25 articles, 1,168 of them showed the diversity measure higher than 25. Their diversity measure indicates that they (re)tweet articles of different topics on average.

We also observed that most of these accounts (92%) belong to the P-centered or semi-P-centered group (See Figure 7). The result indicates that they have a sense of popular topics even though they circulate a broad range of topics. Most of the articles tweeted by them received tweets from a few tens of people. On the other hand, people who pursue diversity of topics are rarely found from the UP-centered and MP-centered group. This can be also seen from the cumulative distribution of the accounts' diversity measure. The cumulative distribution of the UP-centered and MP-centered group more rapidly increases as fewer accounts in the groups pursue diversity. We argue how these accounts can be used for diversification in the discussion section.

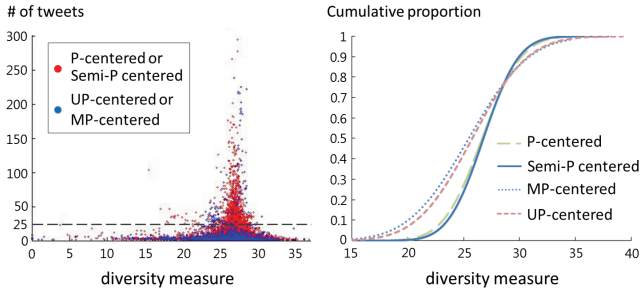

Figure 7. Diversity measurement of accounts (left: in relation to # of tweets, right: cumulative distribution)

Discussion and Future Work

Though we observed that the salient issues of Twitter are concentrated and strongly influenced by the mainstream news media, it is unclear how much the observation can be generalized due to the demographics of the Korean Twitter. There are surveys reporting the demographic difference between Korean Twitter users and the general public in terms of age group and gender (KISA 2011). The agenda diversity may not be similar for different user groups as they have different social and cultural experience. The different intensity in the usage of social media may also lead to different result. A direct extension would be to explore different social media user groups and evaluate the generalizability of our observations.

A related direction is to conduct the study for different language or country. Our result may have deep relation with the socio-cultural context of Korea (e.g., scale, cultural and ethnical homogeneity, strong centralization around the capital city). Countries with different political system (e.g., federation of relatively autonomous regions) and ethnic diversity, or those with strong local media may show different result.

Our study results lead to questions of whether it is possible to diversify the agendas and how. What we observed is that different issues exist in the Twitter network but many of them are not widely distributed. We thus believe that it is possible to increase diversity by encouraging and nudging social media users to discover and circulate more diverse issues.

Considering the strong influence of mainstream news outlets, a possible direction is to investigate whether more diversity in mainstream news media increases diversity in the social media. On the other hand, our observation of the collective power of users in determining the salient issues suggests that a more direct approach is to develop strategies within social media for promoting the dissemination of issues that deserve more attention.

We conjecture that one practical strategy for agenda diversification is to promote potentially popular items of different topics. It will be possible to increase the probability for a user to discover and distribute them by explicitly recommending such tweets or give higher ranks to them at the tweet timeline. Our diversity analysis of individual accounts revealed that a certain portion of users in the P-centered and semi-P-centered group disseminate many articles of different issues which receive considerable number of tweets. The behavioral feature analysis also shows that such users can be identified automatically (e.g., through a linear discriminant analysis); they leave many tweets with news articles and frequently retweet others. Utilizing these users can provide a solution for successful recommendation as they cover a wide range of issues, maintain relationship with others, and care about other's response and have sense of popular issues.

It is also important that the number of people who import a story is critical in the dissemination of an article. Designing incentive mechanisms to encourage users to import more diverse stories or interface components for feedback of diversity status can be another solution.

Conclusion

In this paper, we introduced agenda diversity as a new perspective to social media discourse and investigated the current diversity status of Twitter. The diversity in the

discourse is important since the media has strong influence in shaping the scope of individuals' world view and the topics of public agendas. Our study examined the agenda diversity during the 2012 Korean general election period. Despite the variety of available issues and candidates, we observed agenda concentration, a convergence of attention to a few issues directly related to the election and several nationally recognized candidates. Social media is likely to provide users with fewer number of and topically homogeneous issues than major news outlets. We elaborated on the diversity aspect of information diffusion and discussed potential diversification strategies.

Acknowledgement

This work was supported by the National Research Foundation of Korea grant funded by the Korea government (MEST) (No. 2012-0005733) and by the WCU (World Class University) program of the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology of Korea. (Project No: R31-30007)

References

Allen, R. L., and Izcaray, F. 1988. Nominal agenda diversity in a media-rich, less-developed society. *Communication Research* 15 (1): 29-50.

André, P.; Bernstein, M, S.; and Luther, K. 2012. Who gives a tweet?: evaluating microblog content value. *In Proc. CSCW*.

Asur, S.; Huberman, A.; Szabo, G.; and Wang, C. 2011. Trends in social media: Persistence and decay. In Proc. of *ICWSM.*

Bandari, R.; Asur, S.; and Huberman. 2012. The pulse of news in social media: Forecasting popularity. In Proc. of *ICWSM*

Bernstein, M. S.; Suh, B.; Hong, L.; Chen, J.; Kairam, S.; and Chi, E. H. 2010. Eddi: interactive topic-based browsing of social status streams. *In Proc. UIST*.

Choudhury, M. D.; Counts, S.; and Czerwinski, M. 2011. Identifying relevant social media content: leveraging information diversity and user cognition. *In Proc. of the Hypertext*.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Goncalves, B.; Menczer, F.; and Flammini. 2011. A. Political polarization on Twitter. *In Proc. ICWSM.*

Diakopoulos, N., and Naaman, M. 2012. Towards quality discourse in online news comments. *In Proc. CSCW*.

Faridani, S.; Bitton, E.; Ryokai, K.; and Goldberg, K. 2010. Opinion space: a scalable tool for browsing online comments. *In Proc. CHI.*

Garrett, R. K. 2005. Exposure to controversy in an information society. *Ph.D. diss*., University of Michigan.

Hong, L.; Convertino, G.; and Chi, E. D. 2011. Language matters in twitter: A large-scale study. *In Proc. of ICWSM.*

KISA(Korea Internet and Security Agency). 2011. Internet Usage Survey. Available at *http://isis.kisa.or.kr/board/index.jsp?pageId=060200&itemId=793*

Kriplean, T.; Morgan, J.; Freelon, D.; Borning, A.; and Bennett, L. 2012. Supporting reflective public thought with considerIt. *In Proc. CSCW*.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is Twitter, a social network or a news media?. *In Proc.of WWW*.

Lapata, M. 2006. Automatic Evaluation of Information Ordering: Kendall's Tau. *Computational Linguistics* 32 (4): 471-484.

Lerman, K. and Ghosh, R. 2010a. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. *In Proc. of ICWSM.*

Lerman, K. and Hogg, T. 2010b. Using a model of social dynamics to predict popularity of news. *In Proc. of WWW*.

Livne, A.; Simmons, M. P.; Adar, E.; and Adamic, L, A. 2011. The party is over here: Structure and content in the 2010 election, *In Proc. ICWSM.*

McCombs, M. and Shaw, D. L. 1972. The agenda-setting function of mass media. *Public Opinion Quarterly* 36 (2): 176-187.

McCombs, M. and Zhu, J. H. 1995. Capacity, diversity, and volatility of the public agenda. *Public Opinion Quarterly* 59 (4): 495-525.

Morris, M.R.; Counts, S.; Hoff, A.; Roseway, A.; and Schwarz, J. 2012. Tweeting is believing?: understanding microblog credibility perceptions. *In Proc. CSCW*.

Munson, S. A. and Resnick, P. 2010. Presenting diverse political opinions: how and how much. *In Proc. CHI.*

Munson, S. A.; Zhou, D. X.; and Resnick, P. 2009. Sidelines: An algorithm for increasing diversity in news and opinion aggregators. *In Proc. ICWSM.*

Park, S.; Ko, M.; Kim, J.; Liu, Y.; and Song, J. 2011. The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns. *In Proc. CSCW*.

Park, S.; Kang, S.; Chung, S.; and Song, J. 2009. NewsCube: delivering multiple aspects of news to mitigate media bias. *In Proc. CHI.*

Pew Research Center. 2012. In Changing News Landscape, Even Television is Vulnerable. Available at *http://www.people-press.org/2012/09/27/in-changing-news-landscape-even-televisionis-vulnerable/*

Ramage, D.; Dumais, S.; and Liebling, D. 2010. Characterizing microblogs with topic models. *In Proc. ICWSM.*

Shamma, D. A.; Kennedy, L.; and Churchill, E. F. 2011. Peaks and persistence: modeling the shape of microblog conversations. *In Proc. CSCW*.

Starbird, K. and Palen, L. 2012. (How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising. *In Proc. CSCW*.

Wu, F. and Huberman, A. 2007. Novelty and collective attention. *Proceedings of the National Academy of Sciences,* 104 (45): 17599-17601.

Zhou, D.X.; Resnick, P.; and Mei, Q. 2011. Classifying the political leaning of news articles and users from user votes. *In Proc. ICWSM.*