An Exploratory Analysis of Browsing Behavior of Web News on Twitter

Hao Han∗, Hidekazu Nakawatase∗ and Keizo Oyama†
†The Graduate University for Advanced Studies (SOKENDAI), Japan
∗National Institute of Informatics, Japan
{han,nakawtse,oyama}@nii.ac.jp

Abstract—Twitter, a most popular micro-blog, serves more as a news medium than a social networking service. It attracts and expands the news oriented researches. News oriented in-formation and personalized community recommendation such as their main news portals for quick acquisition of information and knowledge, and even some specialists said Twitter would attract their curiosities and attentions, and share their concerned news on Twitter. So, it is natural to find out users’ concerned topics by analyzing their shared news. These facts and assumption are reflected in many recommendation approaches widely employed by news sites, blog sites, electronic commerce sites and some social sites.

However, an important difference between Twitter and other services, especially traditional blogs, is ignored. As a quick and convenient way of receiving updates about news, most tweets of shared news contain few or even no original user opinions and comments. Users copy and paste news titles and page URLs into tweets or click Tweet Buttons to share news. A tweet can be posted within an extremely short period of time such as 5 seconds, which is far inadequate for posting a traditional blog article. This quick pace of Web life brings many uncertain elements in browsing behaviors. For example, some users may share highlighting news articles for attracting more retweets/replies/followers on Twitter without serious considering whether these news articles bring concerns or interests to themselves since news sharing costs few time and brings nothing harmful to their Web profiles.

Here, we confirm whether abovementioned assumption is consistent with current actual situations. In this paper, we analyze the Web news oriented browsing behavior on Twitter and usual Web including news access, news sharing, and news related query by using net view data, Twitter data, and Web news articles as data sets. We also compare the analysis results to present the different trends in various news topics and categories, and discuss the possible influences to users’ information acquisition and news recommendation. Our analysis emphasis is laid on Web news provided by external news sites instead of posted photos or chatters.

The organization of the rest of this paper is as follows. Section II gives research motivation and an overview of the related work. In Section III, we present the data set in detail. Then, we analyze and compare the different browsing behaviors, and discuss the possible influences brought by different news media in Section IV, V and VI. Finally, the conclusion and the future work are given in Section VII.

II. RESEARCH MOTIVATION AND RELATED WORK

Twitter has been breaking news before the mainstream media since it first sputtered into life [1], which stimulates the

References

1. http://www.twitter.com/
2. http://dev.twitter.com/docs(tweet-button)
news oriented researches on Twitter. Hu et al. present a study of the role of micro-blogs in breaking and spreading news by analyzing how Osama Bin Laden’s death leaked through Twitter. An in-depth analysis of how the news broke and spread on Twitter was provided, and Twitter’s rising potential in news reporting was confirmed [7]. Wu et al. investigated a classic problem in media communications research, “who says what to whom?” in the context of Twitter. Users were classified into “elite” and “ordinary” and different categories of users emphasized different types of content [18]. Kwak et al. crawled the entire Twitter site and analyzed the text, topic and users subscription of Twitter fully. The study of topological characteristics of Twitter gave the conclusion that Twitter serves more as a news medium [9]. Therefore, detection of news from the timely tweets and recommendation of news on the analysis of tweets become the popular research topics.

- Event/news detection: Sakaki et al. investigated the real-time interaction of events such as earthquakes in Twitter and proposed an algorithm to monitor tweets and to detect a target event. Each Twitter user was considered as a sensor, and location estimation methods were used to estimate the locations of events. An earthquake reporting system was developed to notify people promptly of an earthquake event [12]. Twitcident is a system that allows users to explore, search and analyze information about incidents or crises available on the social Web. Given an incident, it automatically starts tracking and filtering information that is relevant for the incident from social Web streams and Twitter particularly. Cao et al. present a framework to automatically collect relevant micro-blogs from microblogging websites to generate comments for popular news on news websites [3].

- News/Twitterer recommendation: TwitterRank [17] is an extension of PageRank algorithm, and measures the influence taking both the topical similarity between users and the link structure into account. TwitterStand [13] is a news processing system used to capture tweets that correspond to late breaking news by determining tweet clusters. Buzzer [11] mines the real-time information from Twitter to provide a basis for news matching and recommendation. However, these analysis are limited into the tweet texts and ignores the external data shared in tweets. Since the text of each tweet is restricted to 140 characters, the URLs are often used to link to the external Web pages by Twitter users. Abel et al. developed a user modeling framework for Twitter and investigated how the different design alternatives influence the characteristics of the generated user profiles. Profiles changes over time were observed and temporal patterns such as characteristic differences between weekend and weekday profiles were also discovered [2]. Morales et al. leveraged information found in real-time Web to make personalized news recommendations. Data collected from twitter and Yahoo! news was used to build user profiles that model user interests, content found in the social neighborhood

of users, and topic popularity [10]. However, the news shared in tweets are used to analyze for news recommendation directly without considering the difference between the news reading and sharing.

To address the issue of difference between news reading and sharing, we realize a comprehensive analysis and comparison of Web news browsing behavior on Twitter and other news media. Compared with these developed work, our analysis achieves the following contributions.

- The different sources of data are integrated and the analysis range is extended to the general Web users (not limited to Twitter users), which realizes the effective comparison of the difference between the news reading, sharing and news related query. Furthermore, we give the analysis of browsing trends of different users by using the topics and classifications of news article content.

- In order to give an analysis with higher accuracy and less subjectivity, user profile generated based on net view data and metrics considering precision of news categorization and dwell time of page access are employed to better evaluate the news browsing behaviors.

- The possible influences to users’ information/knowledge acquisition and news recommendation mechanism are discussed based on analysis results of trends of browsing behaviors and users subscription on Twitter.

III. DATA SET

In order to analyze the news browsing behavior on Twitter and other news media, we use the following data sets to complete our data statistics. All the provided data sets are collected from June 18 to July 15 in 2010 (news articles of the Japan earthquake, tsunami and nuclear power plant accident account for a very heavy proportion of news of 2011, and data of 2010 is more typical and suitable for general analysis).

- Net View Data: Net view data is a log of webpage access, collected and provided by Nielsen Online (Nielsen Online owns copyright of this data). It contains 80,912,228 records of webpages access made by 36,704 panel users. Each record contains panel user ID, access time, dwell time (seconds), target URL and referer URL. The personal information, such as the user name or ID parameters in URL, is deleted beforehand in order to protect the personal information of panel users.

- Twitter Data: 14,286,223 pieces of tweets/retweets/replies are collected by Twitter APIs in the status of public_timeline. Each tweet/retweet contains the Twitter user ID, tweet ID, screen name, user name, client name, text and time (GMT). Additionally, reply contains the reply ID.

- Web News Articles: Yahoo! Japan news are selected as our news data set because the Yahoo! Japan news are the most highly accessed (over 60% of accessed news are Yahoo! Japan news) and shared news in the net view data and Twitter data. 1,671 pieces of news articles

4http://twitcident.com/
5http://www.netratings.co.jp/
6http://dev.twitter.com/doc
are extracted from Yahoo! News Topics\textsuperscript{7} and Yahoo! Top Backnumber\textsuperscript{8}. Each news article contains the top page publication period (Japan Standard Time), topic, title, headline, headline page URL, full content, and full content page URL. Here, the full content page URL is an unique and unchangeable URL for each news article. The headline page shows the first paragraph of news article only during the top page publication period, when the news appears in the top page of Yahoo!.

IV. OVERALL ANALYSIS OF NEWS ACCESS AND SHARING

Access represents “this news (title or content) attracts my concern and I want to know the details” and sharing represents “I want to let others know this news attracts my concern and want to discuss on it with others”. By a simple observation, we more or less find that highly accessed news are not always shared on Twitter. For example, many users accessed scandal news about immorality, murder case and peeping photo, which are not mostly highly shared on Twitter. For further precise analysis of news oriented browsing behaviors, the abovementioned data sets are matched and integrated. Different rankings are given to compare news access and sharing, and fluctuation of each week shows variation of user browsing behaviors.

A. Data Matching and Integration

The original data sets described in Section III are noisy and independent to each other. Before taking statistics, we integrate the net view data and news data for the analysis of news browsing behavior on the general overall news media. We also integrate the Twitter data and news data for the analysis of news browsing behavior on Twitter.

1) Net View Data and Web News: Example of the collected news articles is shown in Table I, which is used to integrate with net view data to make the statistics of news access of panel users shown in Table II by the following steps.

1) The URLs found in net view data and news data are converted into the same format.

2) For a panel user \( M \) and a news \( N \), \( NewsAccess_{M,N} \) is calculated as follows:

\[
NewsAccess_{M,N} = \begin{cases} 1 & \text{(NewsURL}\_N \in \text{TargetURL}_M) \\ 1 & ((\text{HeadlineURL}\_N \in \text{TargetURL}_M) \\ \text{and}(\text{AccessTime}_{M,N} \in \text{Period}_N)) \\ 0 & \text{(otherwise)} \end{cases}
\]

(1)

where \( NewsURL\_N \) is the URL of full content page of news \( N \), \( TargetURL\_M \) is the list of target URLs accessed by user \( M \), \( HeadlineURL\_N \) is the URL of headline page of news \( N \), \( AccessTime_{M,N} \) is the access time of headline page of news \( N \) by user \( M \), and \( Period_N \) is the top page publication period of news \( N \).

3) For each news \( N \), \( Count_N \) is calculated as follows:

\[
NewsAccessCount_N = \sum_{M \in U} NewsAccess_{M,N}
\]

(2)

where \( NewsAccessCount_N \) is the sum of access count of news \( N \). \( U \) represents all users (panel users).

2) Twitter Data and Web News: The collected Twitter data is noisier than the net view data because of the Twitter bot (robot), shorten URL and others. We clean the Twitter data and select the useful information from them as follows:

1) Tweets/retweets/replies generated by the following Twitter bots are deleted for the further human being oriented behavior analysis.

- Official bot named “YahooNewsTopics”
- Unofficial bots or possible fake accounts such as “YahooJTopicsFan”, “yahooyj” and “yahoonews_top”
- Possible bots whose names contain the keywords like “Yahoo”, “news”, and “bot”

2) URLs embedded in tweets/retweets/replies are accessed to get the original ones because the shorten URLs are often used in Twitter.

3) The GMT is converted into corresponding Japan Standard Time.

We integrate the news shown in Table I with the Twitter data to make the statistics of news sharing (\( NewsShare_{M,N} \) and \( NewsShareCount_N \)) on Twitter as shown in Table III by the similar steps described in Section IV-A1.

B. News Access and Sharing

We analyze the trends of news access and sharing by using the integrated data sets, which are divided into four parts by weeks (June 18 - June 24, June 25 - July 1, July 2 - July 8, and July 9 - July 15) to display the fluctuation. For each week, the same calculation methods are used, and the analysis results of the first week are presented as the examples. First, two simple and standard metrics are used to analyze based on topics (provided by Yahoo! News) and word frequency. Then, since each user’s concern/endorsement for a piece of news is not equal in reality, dwell time of viewing webpages and precision of categorization are considered in a further category-based analysis.
1) **Topic and Word Frequency:** Topic is used to classify the news articles by Yahoo!, and can be used to reflect the users’ trends of access and sharing more clearly in our analysis. The rankings of news topic access and sharing show the difference between the news browsing behavior of the panel users and Twitter users most simply. We use the following formula to calculate the access count $NewsAccessCount_T$ of topic $T$:

$$NewsAccessCount_T = \sum_{N \in T} NewsAccessCount_N$$

(3)

where $NewsAccessCount_N$ is used to describe the sum of news access (See Section IV-A1). The same formula is used to calculate $NewsShareCount_T$, the sharing count of each topic on Twitter. Here, we find the 255 accessed topics and 84 shared topics. Table IV and Table V show the detailed topic-based rankings (top 10) including the rank, topic, count and the rank in the each other ranking:

<table>
<thead>
<tr>
<th>Topic</th>
<th>Count</th>
<th>Rank@Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>wrestling and gambling</td>
<td>2869</td>
</tr>
<tr>
<td>2</td>
<td>soccer Japan</td>
<td>2750</td>
</tr>
<tr>
<td>3</td>
<td>2010FIFA</td>
<td>2282</td>
</tr>
<tr>
<td>4</td>
<td>comedy</td>
<td>1884</td>
</tr>
<tr>
<td>5</td>
<td>immorality</td>
<td>1658</td>
</tr>
<tr>
<td>6</td>
<td>TV program</td>
<td>1535</td>
</tr>
<tr>
<td>7</td>
<td>2010WorldCup Team A</td>
<td>1531</td>
</tr>
<tr>
<td>8</td>
<td>2010WorldCup Team C</td>
<td>1364</td>
</tr>
<tr>
<td>9</td>
<td>murder case</td>
<td>1340</td>
</tr>
<tr>
<td>10</td>
<td>hayabusa spacecraft</td>
<td>1269</td>
</tr>
</tbody>
</table>

The rankings show that the sports-based topics, especially the World Cup, are most accessed because many news articles are continually generated for the live broadcasts and easily appear in the top page of news site as highlighting news during the games. Besides, topics of scandals like gambling, immorality and murder case are highly accessed, at the same time, topics of IT, fashion and animation are highly shared on Twitter. Topic is given manually in Yahoo! News without strict standards. There are hundreds of topics used in Yahoo! News and some of them are temporary or partially replicated to each other (e.g. “2010 FIFA World Cup”, “2010 World Cup Group A” and “2010 World Cup Japanese Team”). For further analysis of browsing behaviors, we use Sen’s, a morphological analysis tool, to parse the news title and the headline to calculate the word occurrence frequency $WordAccessCount_W$ and $WordShareCount_W$ as follows. Table VI shows the ranking of noun occurrence frequency after we manually remove the meaningless words (e.g. today, Mr., there) and World Cup oriented words, which are of temporary occurrence.

$$WordAccessCount_W = \sum_{W \in N} NewsAccessCount_N$$

(4)

$$WordShareCount_W = \sum_{W \in N} NewsShareCount_N$$

(5)

<table>
<thead>
<tr>
<th>Topic</th>
<th>Count</th>
<th>Rank@Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>suspect</td>
<td>21478</td>
</tr>
<tr>
<td>2</td>
<td>league</td>
<td>19531</td>
</tr>
<tr>
<td>3</td>
<td>arrest</td>
<td>18592</td>
</tr>
<tr>
<td>4</td>
<td>gambling</td>
<td>14217</td>
</tr>
<tr>
<td>5</td>
<td>baseball</td>
<td>13962</td>
</tr>
<tr>
<td>6</td>
<td>parliament</td>
<td>12422</td>
</tr>
<tr>
<td>7</td>
<td>actor</td>
<td>12139</td>
</tr>
<tr>
<td>8</td>
<td>investigation</td>
<td>11919</td>
</tr>
<tr>
<td>9</td>
<td>police</td>
<td>11869</td>
</tr>
<tr>
<td>10</td>
<td>sale</td>
<td>11616</td>
</tr>
</tbody>
</table>

It is easily understood that Twitter users like the sharing and discussion of the news about computer/IT, mobile phone and Twitter services. After removing the words about World Cup, sports oriented words drop out of Table VI, which also proves that the categories of often shared news vary visibly and scandal news are more accessed.

8http://ultimania.org/sen/
2) **Dwell Time and Categorization:** Topic and word frequency are often used in statistics like a social media trend survey. However, it is difficult to clearly view the overall trends of news browsing behaviors and objectively give a quantititative analysis by abovementioned metrics for following reasons.

- Topics of news articles are newly generated and frequently updated, which cannot well reflect classification of news and its variety trends. Word frequency also has similar problem.
- Each user’s concern for a piece of news is not equal. “in-depth reading” and “reading at a glance” represent different concerns and attitudes of users.
- Each news has different emphasis and focus on multi-topics. For example, both “Eight Badminton Players Disqualified From Olympics” and “Badminton Players Get The Most Out Of Olympic Medals” are categorized as sports news. The latter news talks not only the sports but also economic effect brought by Olympics. So, we can assume that some users access latter news for more concerns of economy though it is a sports news mainly and basically. In our analysis, the former news would be considered to contain relatively more essence of sports than the later one.

Here, we assume that the more time (dwell time) a user spends viewing a piece of news, the more concern this user has on this news. The more essence of sports a piece of news contains, the more concern of sports the user viewing this news has. We classify the news into different categories and precision of categorization is used to calculate essence as follows:

1) 24,241 pieces of news articles are randomly collected from Yahoo! News as base clusters and averagely contain various topics. The full content of each news article is divided into a list of words by method described in Section IV-B1 and each list represents a news document. Repeated bisection clustering is used to partition all lists into clusters, and each list is assigned to only one cluster. The lists in each cluster are arranged in descending order of the similarity values between the vectors of lists and the vector of the cluster centroid. Cosine similarity is used to measure similarity between two vectors by measuring the cosine of the angle between them, and term frequency-inverse document frequency model (TF-IDF) is used to calculate the weight (how important a word is to a document) of word as follows:

\[
sim(A, B) = \frac{A \cdot B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n} W_{i,A} \times W_{i,B}}{\sqrt{\sum_{i=1}^{n} W_{i,A}^2 \times \sum_{i=1}^{n} W_{i,B}^2}}
\]

\[
W_{i,j} = tf_{i,j} \cdot idf_i
\]

\[
tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}
\]

\[
idf_i = \log \frac{|D|}{|\{ j : t_i \in d_j \}|}
\]

where \( \sim(A, B) \) is the similarity between document \( A \) and \( B \). \( tf_{i,j} \) is frequency of word \( i \) in document \( j \), \( n_{i,j} \) is the number of word \( i \) in document \( j \), and \( \sum_k n_{k,j} \) is sum of words in document \( j \). \( idf_i \) is inverse document frequency. \(|D|\) is the total number of documents, and \(|\{ j : t_i \in d_j \}|\) is the number of documents containing the word \( t_i \).

2) Another 1500 pieces of news are selected as learning examples and manually classified into following 10 categories exclusively.

- Society (e.g. crime and incident)
- Culture/Arts (e.g. history and world heritage)
- Politics (e.g. cabinet meeting)
- Education/School (e.g. university tuition fees)
- Science/Technology (e.g. hayabusa spacecraft)
- Entertainment (e.g. cinema and singer)
- Economics/Business (e.g. company and industry)
- Sports (e.g. World Cup and baseball)
- Health/Live (e.g. influenza and weather)
- Computer/IT (e.g. PC virus and Internet service)

Then, after words division and similarity calculation as mentioned above, each news is assigned into a cluster. In each cluster, there are news of different categories. Here, we designate the category containing most news in a cluster as the main category of this cluster. So, the news articles whose categories are same as main category of its assigned cluster are judged as correctly categorized ones while the rests are wrongly categorized ones. There are two conventional methods of calculating the performance of a categorization system based on precision. Micro-averaged values are calculated by constructing a global contingency table and then calculating precision using these sums. In contrast macro-averaged scores are calculated for each category and then taking the average of these. The notable difference between these two calculations is that micro-averaging gives equal weight to every document (document-pivoted measure) while macro-averaging gives equal weight to every category (category-pivoted measure) \([5]\). We change the sizes of base clusters and centroids to get a proper precision \( P_{macro} \). In Figure 1, different sizes of centroids do not bring much fluctuation and \( ClusterSize = 50 \) is an elbow point of precision \( P_{macro} = 75.78\% \). We select 50 as final size of clusters and calculate each \( P_{micro} \) (from \( ClusterSize = 60 \), the \( P_{macro} \) does not become visibly higher but cost of calculation is much higher). For a panel user \( M \) and a news \( N \), Table VII gives \( DwellTime_{MN} \) and \( DwellTime_N \) by similar method described in Section IV-A1 and sum of time is calculated if a user views one piece of news twice or more.

4) Each accessed and shared news is assigned into one of 50 base clusters based on similarity calculation. Because of the different number of news article in different category, we normalize the category-based concerns of access \( ConAC \) and sharing \( ConSC \) as follows:
TABLE VII
STATISTICAL RESULT OF DWELL TIME

<table>
<thead>
<tr>
<th>Panel User ID</th>
<th>article1</th>
<th>article2</th>
<th>......</th>
<th>article1671</th>
</tr>
</thead>
<tbody>
<tr>
<td>1410***</td>
<td>56</td>
<td>0</td>
<td>......</td>
<td>0</td>
</tr>
<tr>
<td>5118***</td>
<td>0</td>
<td>71</td>
<td>......</td>
<td>45</td>
</tr>
<tr>
<td>1793***</td>
<td>57</td>
<td>0</td>
<td>......</td>
<td>0</td>
</tr>
<tr>
<td>2599***</td>
<td>0</td>
<td>104</td>
<td>......</td>
<td>54</td>
</tr>
<tr>
<td>DwellTime</td>
<td>......</td>
<td>......</td>
<td>......</td>
<td>......</td>
</tr>
</tbody>
</table>

Con\(_{AC}\) = \(\frac{\sum_{\text{CL}_i \in C} \sum_{\text{N} \in \text{CL}_i} \text{DwellTime}_N \times P_{\text{micro}_i}}{\sum_{\text{CL}_i \in C} |\text{CL}_i|}\)  
(10)

\(\text{Con}_{SC} = \frac{\sum_{\text{CL}_i \in C} \sum_{\text{N} \in \text{CL}_i} \text{NewsShareCount}_N \times P_{\text{micro}_i}}{\sum_{\text{CL}_i \in C} |\text{CL}_i|}\)  
(11)

where \(\text{CL}_i\) is cluster \(i\), \(P_{\text{micro}_i}\) is micro-averaging precision of documents in \(\text{CL}_i\), and \(|\text{CL}_i|\) is the number of news categorized in \(\text{CL}_i\).

The comparison of ranking is given in Table VIII. Our category-based analysis can give an clearer overall view of news access and sharing compared to topic-based rankings or word frequency rankings objectively. These results show the entertainment news are most accessed, and the computer/news are most shared on Twitter. We use the same algorithms to calculate the category-based news access/sharing for the rest three weeks respectively. Figure 2 and Figure 3 show the fluctuation of rankings: the entertainment news keep the most accessed and the computer/IT news keep the most shared during the four continuous weeks. Besides, compared to the news access, the fluctuation of news sharing is more visible and reflects that users do not always share news fully according to their concerns or interests, which keeps relatively stable and unchangeable.

The collected tweets are just a small part of overall tweets because Twitter opens the partial tweets (about 1%-2%) to each API-based collection job and we cannot collect all the tweets. We calculate the 95% confidence interval as shown in Table IX. The value of \textit{Percentage}_2 of computer/IT news is larger than the corresponding maximum value of the interval. It also proves that the computer/IT news are highly shared on Twitter than accessed by panel users.

V. NEWS RELATED QUERY ANALYSIS

Aimless browsing or accidental clicking is an unavoidable problem in analysis of webpage access and it is also not seldom in news access. Most news sites recommend some highlighting news at their top pages and many users could click links to these news pages easily and aimlessly as a behavior like “try viewing”. Some news would attract users’ concerns and the others would be ignored after a glance.
Obviously, these accessed but ignored news cannot represent users’ concerns.

Here, news related queries are employed to verify whether access reflects interests of users since aimless query is a rare behavior. If users access a news and search for related information about this news after access, we can make certain users have deep concerns/interests in this news. We find the valid related queries as follows:

1) For each news access $N_{\text{news access}}$, query parameters are extracted from URLs if user $M$ uses search engines provided by Google $^{10}$ or Yahoo! $^{11}$ or Infoseek $^{12}$ or Amazon $^{13}$ or Wikipedia $^{14}$ after accessing news $N$.  
2) Each query parameter $N_{Q}$ is divided into a list of word $W_{NQ}$ by Sen. If a word $W_{NQ}$ occur within title or any paragraph of news $N$, $N_{Q}$ is considered as a possible “related query”.  
3) Whether query $N_{Q}$ is really related to news $N$ or not is checked manually by using data of the first week. If $N_{Q}$ is related to news $N$, its index value $I_{NQ}$ given as 1 (related query), otherwise 0 (unrelated query).  
4) Actual valid query period $P$ is calculated to satisfy break-even point (precision = recall) between related queries and unrelated queries as follows. As a result, $P$ is calculated as 829 seconds and precision = recall = 78.99%.  
$$\sum_{AT \in P} I_{NQ} / |NQ| = \sum_{AT \in P} I_{NQ} / \sum I_{NQ} \Rightarrow |NQ| = \sum I_{NQ}$$  
(12)
where $AT$ is the time of query $N_{Q}$ sent to search engines (time of accessing search result page) and $|NQ|$ is the sum of possible related queries during period $P$.  
5) Possible related queries of all four weeks submitted during $P$ are considered as related queries. Category-based average query count $AC_{Q}$ is calculated as follows and ranking is given in Table X.

$$AC_{Q} = \frac{\sum_{N \in C} I_{NQ}}{|C|}$$  
(13)
where $|C|$ is the sum of news in category $C$.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Category</th>
<th>Average Query Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Entertainment</td>
<td>2.34</td>
</tr>
<tr>
<td>2</td>
<td>Education/School</td>
<td>0.98</td>
</tr>
<tr>
<td>3</td>
<td>Sports</td>
<td>0.87</td>
</tr>
<tr>
<td>4</td>
<td>Computer/IT</td>
<td>0.82</td>
</tr>
<tr>
<td>5</td>
<td>Culture/Arts</td>
<td>0.65</td>
</tr>
<tr>
<td>6</td>
<td>Economics/Business</td>
<td>0.58</td>
</tr>
<tr>
<td>7</td>
<td>Society</td>
<td>0.45</td>
</tr>
<tr>
<td>8</td>
<td>Politics</td>
<td>0.39</td>
</tr>
<tr>
<td>9</td>
<td>Science/Technology</td>
<td>0.33</td>
</tr>
<tr>
<td>10</td>
<td>Health/Live</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Comparing rankings of news access, news sharing and news related query, Spearman’s rank correlation coefficient $^{[14]}$ is used to measure similarity value $\rho$ as follows:

$$\rho = 1 - \frac{6 \sum d_{i}^{2}}{n(n^{2} - 1)}$$  
(14)
where $d_{i}$ represents differences between the ranks of each observation on the two rankings, and value of $n$ is 10. $\rho$ of news access and news related query $\rho_{AQ} = 0.84$, and $\rho$ of news sharing and news related query $\rho_{SQ} = 0.63$. They show news access ranking reflects users deep concerns to a great extent or degree, and news sharing ranking is much different to them.

VI. DISCUSSION

Based on analysis of news access, sharing and related query, we find that Web users do not always share their most interested news on Twitter. News articles about entertainment, sports and culture attract interests of users in general life, but users share news articles about politics, IT and economics/business on Twitter. Moreover, news sharing trends vary more visibly than actual interests of users. The difference of Web news oriented browsing behavior between

---

10http://www.google.co.jp
11http://www.yahoo.co.jp
12http://www.infoseek.co.jp
13http://www.amazon.co.jp
14http://www.wikipedia.org

TABLE IX
ANALYSIS OF 95% CONFIDENCE INTERVAL

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of News in Net View Data</th>
<th>Percentage1</th>
<th>Number of News on Twitter Data</th>
<th>Percentage2</th>
<th>Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Society</td>
<td>287</td>
<td>0.172</td>
<td>68</td>
<td>0.1351</td>
<td>0.1361038804, 0.207403019</td>
</tr>
<tr>
<td>Culture/Arts</td>
<td>18</td>
<td>0.011</td>
<td>5</td>
<td>0.0116</td>
<td>0.0001049535, 0.020529033</td>
</tr>
<tr>
<td>Politics</td>
<td>245</td>
<td>0.147</td>
<td>55</td>
<td>0.1279</td>
<td>0.113184802, 0.13605278</td>
</tr>
<tr>
<td>Education/School</td>
<td>41</td>
<td>0.025</td>
<td>10</td>
<td>0.0233</td>
<td>0.0099913868, 0.039159026</td>
</tr>
<tr>
<td>Science/Technology</td>
<td>53</td>
<td>0.032</td>
<td>11</td>
<td>0.0359</td>
<td>0.0151535233, 0.040281815</td>
</tr>
<tr>
<td>Entertainment</td>
<td>228</td>
<td>0.136</td>
<td>73</td>
<td>0.1698</td>
<td>0.104000377, 0.168890108</td>
</tr>
<tr>
<td>Economics/Business</td>
<td>222</td>
<td>0.133</td>
<td>60</td>
<td>0.1395</td>
<td>0.100727974, 0.164936182</td>
</tr>
<tr>
<td>Sports</td>
<td>459</td>
<td>0.275</td>
<td>109</td>
<td>0.2533</td>
<td>0.232496446, 0.316875138</td>
</tr>
<tr>
<td>Health/Live</td>
<td>88</td>
<td>0.055</td>
<td>22</td>
<td>0.0512</td>
<td>0.031551182, 0.07377497</td>
</tr>
<tr>
<td>Computer/IT</td>
<td>30</td>
<td>0.018</td>
<td>17</td>
<td>0.0395</td>
<td>0.005402842, 0.010503801</td>
</tr>
</tbody>
</table>
Many users partially hide their own concerns, hobbies and interests, and emphasize the concerns about social phenomenon, such as political viewpoints and IT knowledge. They may consider whether shared news would bring negative elements to their impression left to their followers. For example, in our analysis we find the scandals of immorality of female entertainers is most accessed but less shared because many users think that sharing of such news would make themselves look like people who spread rumors and gossip.

Retweet means sharing a message sent from followings, and tweet means (active) sharing a new message found outside of Twitter. There is a lot of content about celebrities in entertainment on Twitter, but commonly there are photos and related chatters (reply and retweet). General Twitter users prefer to repost messages (existing tweets) rather than to post an external Web news (new tweets) about entertainment.

Some users share news in an aimless way for a kind of “presenting existence of self among others” not their concerns or interests. We get some interesting examples and give a simple analysis. For example, a user who usually only posts tweets about manga (Japanese comics) continuously shared 4 pieces of sports news in less than 90 seconds. Through access log in net view data, we found this user just accessed headline pages (title and abstract) and did not access full content pages, which shows this user did not have true concern in these sports news and just wanted to present an existence of self with specific intentions like attracting more retweets/replies/followers.

We do not think this user would like to read sports news even if sports news are recommended. This also happens to sharing of general webpages. A detailed user-profile-based analysis is given in Appendix.

Some users share news (mention by @), which could attract interests of friends not themselves (recommend news to others not self).

Some news articles attract users’ concerns easily but it is not a general public discussion topic such as funeral.

More and more users prefer browsing on smart phone (mobile social), which is easily affected by small trivial matters beyond users’ strong concerns or interests.

The difference between the news access and sharing on Twitter would make the researches based on the Twitter data, such as the news discovery and recommendation, be easily limited to some special topics. The abovementioned examples also show that we cannot easily make certain that a user sharing many sports news must have interest in sports news. It is challengeable to find tweets reflecting user true interests from large quantities of tweets containing chatters and aimlessly shared news. Besides news tweets, followers/followings, user profile information and interval time of twitter posts are not negligible elements. Moreover, we find holiday and festival pattern (valentine’s day, etc) is ignored though weekend pattern is analyzed in many twitter oriented researches.

Web news reading is considered as an approach to acquire the latest information/knowledge. We make a statistic of news access range by category. For each user, we calculate the sum of accessed category and give the statistical result as shown in Figure 4. Here, 20.4% of all the users accessed news from only one category, and 4.2% of all the users accessed all the ten categories. For the analysis of possible influences to user’s information acquisition, we also collect the information of the followers of each Twitter user who shared the news, and make a statistics of possible news access of followers. If we suppose that each shared news is accessed by all the followers (or we can say each Twitter user accesses all the news shared by followings), Figure 5 gives the statistical result of possible news access. It proves that the Twitter users only could access news limited into a small range if they use Twitter as their news portals and social networks still cannot replace the important place of traditional news sites in the news-based information/knowledge delivery currently. Compared to the users who use Twitter as their main approach of news acquisition, the users who directly access the traditional Web news sites can acquire more fresh information and widen their knowledge.

![Fig. 4. Percentage of Sum of Accessed Category](image)

VII. CONCLUSION AND FUTURE WORK

In this paper, we analyzed the news access/sharing and related query behavior on Twitter and the usual Web. We found that the highly accessed news are not necessarily shared on Twitter. Twitter users tend to share the politics, IT and economics/business news and the fluctuation of category of often shared news is more visible. We also explained the reasons causing this state and discussed the possible influences brought to users’ information acquisition and news recommendation on Twitter.

As the future work, we will widen our analysis range to world news and other browsing logs such as vision movements.

1526 of “Twitter top 50 most followed Japanese” are entertainers on September 7, 2012 at http://meyou.jp/ranking/follower_allcat

16http://twitter.com/settings/profile
or cursor movements [6]. We also will apply our analysis results to evaluate the information credibility on Twitter, and construct news recommendation system, news filtering system and other news-based systems.

VIII. ACKNOWLEDGEMENT

We gratefully acknowledge the Twitter data provided by Yaman Laboratory (Department of Computer Science, Waseda University, Japan). We also acknowledge the technical support provided by Feng Xiao (Department of Computer Science, Tokyo Institute of Technology, Japan). This work was supported by a Grant-in-Aid for Scientific Research A (No.22240007) from the Japan Society for the Promotion of Science (JSPS).

REFERENCES


APPENDIX

Different users have different interests and concerns on the Web. Some users have very diverse interests while others just focus on a limited spectrum. For both of these two types of users, here is a reasonable and common assumption: if users have concerns on a topic, they would often access related information (news and non-news webpages) about this topic. Some users share webpages (title and link) on Twitter with a serious consideration. They would access related/recommended internal webpages or other external webpages for previous or further acquisition of related information, which means deciding to share a webpages after accessing some related webpages, or sharing a concerned webpages firstly and continuing accessing other related webpages. In contrast, some users prefer to casually and aimlessly share webpages without deep and further interests on related information. Usually, interests are reflected in user profiles [4], which are mined from access logs. Here, we generate a profile (4 weeks period) for each user as follows:
1) Access Log Cleaning: Redundant access such as images, sound, flash files are removed. Private access such as online banking and email are also removed.
2) Page Content Extraction: An algorithm of boilerplate detection using shallow text features [8] is used to detect and extract the main textual content of a webpage. For each user, all the webpages in cleaned access log are extraction targets.
3) Profile Vector Generation: As methods described in Section IV-B1, content extracted from webpages is divided into words. An array of word list is constructed as a user profile and each word list records a vector of an accessed webpage.

We give a context oriented analysis based on user access profile. 65 active users who posted most tweets in all panel users are selected. They keep general Twitter accounts, not official company accounts or other bot or cyborg accounts. We calculate the vector similarity value between each shared webpage and all other webpages accessed within one hour [-3600s, 3600s] near to this webpage sharing (-3600s: an hour before sharing, 3600s: an hour after sharing). In Figure 6,
different users show different browsing behaviors of webpage sharing. For example, 1-2 \textsuperscript{17}, 2-3 and 3-1 accessed much related information when they shared webpages on Twitter, while 1-1, 2-4 and 5-3 acted quite the opposite: they had no deep and continuous concerns or interests on their shared webpages. All similarity values of 65 users (141,459 valid values) are collectively plotted in Figure 7. It shows that many users accessed webpages containing related information during a period of short time, which approximates [-900s, 900s] and partially confirms the valid period of “deep concerns” got in Section V.

\textsuperscript{17}1-2: the 2nd user of the 1st line