2014 Master Thesis

Topics and Influential User Identification in Twitter using Twitter Lists

Date: July 21, 2014
Adviser: Professor HAYATO YAMANA

Department of Computer Science and Engineering
Graduate School of Fundamental Science and Engineering
Waseda University
Student ID: 5112BG05-1

GUANYING ZHOU
Abstract

Twitter, as one of the most popular social network services, draws the attention of more and more researchers worldwide. With a large amount of information tweeted every day, it turns essential to identify the influential users we are interested in. In the previous research, researchers mainly identify topics from tweets and rank users by utilizing the follow relationship; however, the following relationship is strongly related to their reputation in real world and cannot describe their influence and activity level in Twitter exactly. Instead, in this paper, to identify topics and influential users, we use “Twitter List,” whose name represents the topic of listed members. By analyzing Twitter List, we are able to detect topics and identify influential users in the corresponding topic more efficiently. Based on our experimental evaluation using the selected five topics, the influential users identified by our proposed method have the average influence score related to the topic made by interviewees of 3.20 outweigh the methods of ranking by follower numbers with the average score of 3.03 and ranking by the number of published tweet with the average score of 2.55.
Contents

1. Introduction .................................................................................................................. 1
2. Twitter .......................................................................................................................... 3
   2.1 Social Network Service .............................................................................................. 3
       2.1.1 Sina Weibo ........................................................................................................ 4
       2.1.2 Facebook ........................................................................................................... 5
       2.1.3 LinkedIn ........................................................................................................... 5
       2.1.4 Google+ ............................................................................................................ 6
       2.1.5 WeChat ............................................................................................................. 6
       2.1.6 Summary ........................................................................................................... 6
   2.2 Terminology Related To Twitter .............................................................................. 7
       2.2.1 Tweet ................................................................................................................ 7
       2.2.2 Retweet .............................................................................................................. 8
       2.2.3 Reply ............................................................................................................... 8
       2.2.4 Mention ........................................................................................................... 9
       2.2.5 List ................................................................................................................... 9
       2.2.6 Hashflag ........................................................................................................ 10
   2.3 Twitter API ............................................................................................................. 10
       2.3.1 Search API ...................................................................................................... 10
       2.3.2 Streaming API ................................................................................................. 11
       2.3.3 REST API ........................................................................................................ 11
3. Related Research ....................................................................................................... 13
   3.1 Research Related to Topic Extraction ................................................................... 13
   3.2 Research Related to Influential User Identification ............................................ 13
   3.3 Research Related to Twitter List .......................................................................... 15
   3.4 Summary ............................................................................................................... 16
4. Methodology ............................................................................................................. 17
   4.1 Twitter List ............................................................................................................ 17
   4.2 Dataset .................................................................................................................. 18
   4.3 Methodologies ....................................................................................................... 20
       4.3.1 Obtaining Twitter List Resource ................................................................. 21
       4.3.2 Tag Extraction ............................................................................................... 21
       4.3.3 Applying LDA model ..................................................................................... 21
       4.3.4 Topic Influence Calculation ........................................................................... 21
1. Introduction

Twitter is one of the most popular social network services with many registered users from all over the world. In Twitter, registered users can post less than 140-character messages, follow other users, retweet tweets published by other users, reply to other users, comment tweets and enjoy many other social activities in the platform of Twitter.

The platform of Twitter was created in the year of 2006, and gained popularity all over the world rapidly. According some technical reports\(^1\), the platform of Twitter have 500 million registered users in 2012, 340 million tweets posted per day. The service also handled 1.6 billion search queries per day\(^2\).

In the year of 2009, Twitter initialized the new function of “Twitter List”, which allows users to access the tweets of a group of users only by following the list. A Twitter List is a curated group of Twitter users. In order to create a Twitter List, first of all, the creator should name the List, then include other users into the List. The name of the Twitter List describes the common attribute of the users in the List. From the timeline of the List, users can view the tweets published by all of the members in the List just from the timeline of the List. Users can also subscribe Twitter List created by other users.

Users utilize the function of Twitter List for the following reasons\(^3\): 1) Monitor without following. Users can monitor a group of individuals in a list without following any of the folks on the list. 2) Easy user management. Users can find a list and follow the entire listed users with one button, which makes it easy to manage. 3) Promote lists. It is a great way for a group of users to promote each other in Twitter by creating one list. 4) Build a bigger following circle. Twitter users can increase their followers and raise their reputation by creating some useful lists and promote the lists in public to attract more followers.

According to the survey of the research firm Pear Analytics\(^4\), only 9% of the tweets in Twitter have values that should be paid attention to others are useless information. So it is essential to obtain information with value from Twitter in an efficient way. One of the methodologies to solve the problem is to follow the users who are authorities and often tweet valuable information in the area. The motivations of our research are to extract topics and identify influential users in Twitter.

---

In the previous research, researchers mainly extract the interest topics from the tweets published by users and rank users utilizing the follower/friends graph or retweet graph. In Twitter, some of the users rarely publish messages with value but got many followers because of their influence in real life. So the follower/friend graph cannot represent their real influence in Twitter. In our research, we utilize the Twitter List, which can better represent their influence in Twitter, to extract topics and rank users.

The rest of the paper is organized as follows: Chapter 2 introduces some knowledge related to social network services. Chapter 3 introduces related research about topics and influential user identification in social network services. Chapter 4 introduces the methodologies of our research. Chapter 5 introduces the results of our experiment. Chapter 6 introduces some evaluation methods of our experiment and the last chapter introduces the conclusions of our experiment.
2. Twitter

2.1 Social Network Service

A social network service is a web based service providing a platform to build social networks or social relations. In social network services, users can register for account with public profile, create a connection with other users to interact with other users to share common interest. Social network services also provide some additional services such as photo/video sharing, sending instant messages or emails. Users with common interests or backgrounds in social network services can also gather around in a group to communicate with each other. With the rapid development of social network services, users can communicate with others regardless of the borders of countries at any time.

Recently, social network services developed with a rapid speed all around the world. Difference kinds of social network services have different market share in different countries. For example, users in United States mainly use the social network services of Facebook, Google+, LinkedIn and Twitter. The most popular social network service in China is Sina Weibo. In Japan, the social network service of mixi also has got many users.

According to the technical report, the numbers of unique visitors of major social network services are as follows by the end of November 2011.

Table 1.1 Number of Unique Visitors of Popular Social Network Services

<table>
<thead>
<tr>
<th>Worldwide</th>
<th>Unique Visitors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>792,999,000</td>
</tr>
<tr>
<td>Twitter</td>
<td>167,903,000</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>94,823,000</td>
</tr>
<tr>
<td>Google+</td>
<td>66,756,000</td>
</tr>
<tr>
<td>MySpace</td>
<td>61,037,000</td>
</tr>
</tbody>
</table>

6 "ComScore: Google+ Grows Worldwide Users From 65 Million In October To 67 Million In November". December 22, 2011.
In the beginning stage, social network services are used by users to communicate with others. The registered users of social network services grow in a rapid speed and many commercial companies use social network services to promote their product or services as a kind of advertisement. Also, many researchers utilize the large amount of data created in the platform of social network services for research.

In the following part, we will introduce some popular social network services all over the world.

2.1.1 Sina Weibo

Sina Weibo is a Chinese micro-blogging website. Sina Weibo was created by Sina Corporation in 14 August 2009 and it is a social network services similar to Twitter. Users can access to the service to upload messages or photos by web, smart phone applications, SMS messages. Similar to the promotion method of “Sina Blog”, they ask stars and celebrities to join Sina Weibo and verify them with their real ID authentication. To differ them from normal users, the accounts of the verified celebrities show with a character of “V”. By December 2012, Sina Weibo has 503 million registered users.

We introduce some basic functions of Sina Weibo, which are quite similar to Twitter. 1) Tweet. Users can publish a tweet less than 140 characters and in the tweet users can use “@UserName” to reply or mention other users. Also it is possible to publish a tweet with “#HashTag” to highlight some words. 2) Retweet. Sina Weibo users can re-post with “//@UserName” to retweet tweets published by other users and then the tweet retweeted can be seen from the timeline of the user. 3) Follow. In the platform, users can construct the follow relationship to other users to get the most recent tweets published by the user without the verification of the user followed. In Sina Weibo, the follow relationship is a one-way relationship similar to Twitter.

A feature that differs Sina Weibo from the platform of Twitter is micro-topics. In Sina Weibo, users can create a community to publish their opinions about a specific topic. Any user can read the messages published in the community and take part in the discussion.

<table>
<thead>
<tr>
<th>Others</th>
<th>255,539,000</th>
</tr>
</thead>
</table>

2.1.2 Facebook

Facebook is an online social network service created by Mark Elliot Zuckerberg and his friends from Harvard University in February 2004. In the beginning stage, the services of Facebook were only provided for students from Harvard University and now it is public for any users whose age is above 13. In the platform of Facebook, users can send pictures, videos, sound files and other files. Only registered users can use the service of Facebook and registered users can send a request to be their friends, join a group with users who have similar interests. The service of Facebook has one billion active users as of September 2012.

In Facebook, users can publish statuses in their own timeline and all of the friends of the user will see the update. Other users can reply to the user and click the button of “Like” to interact with the user. Users can also send messages to others which are only public between the users similar to e-mail. In Facebook, users can also create some events and send others about the details about the event to help connect user offline. The users of Facebook can make voice calls by using Facebook Chat and users can chat with each other from all over the world from April 2011. The service increases the interaction level of users because users can chat in real life.

2.1.3 LinkedIn

LinkedIn is a business-oriented social network service, which is different from Twitter and Facebook. LinkedIn was founded in December 2002 and launched on May 2003. In the platform of LinkedIn, users can create their own profiles, including their education experience and work experience similar to resume. Then users can connect with their friends, find jobs and business opportunities, follow the companies they are interested in to get the recent job opportunities of the company. Employers can list jobs and search for potential candidates. The recommendation system of LinkedIn also recommends users with possible job opportunities.

The site of LinkedIn is available in 20 languages. As of 2 July 2013, LinkedIn has 33.9 million unique U.S. visitors and 178.4 million globally.

---

9 Carlson, Nicholas (March 5, 2010). "At Last - The Full Story Of How Facebook Was Founded". Business Insider.
10 "Information For Parents and Educators". Facebook. Retrieved November 22, 2011.
2.1.4 Google+

Google+ is a social network service created by the company of Google. It is an authorship tool that associates web-content directly with its owner. The service of Google+ was started in November 2011\(^\text{15}\). The services of Google+ can integrate with other services such as YouTube, Google Maps, Gmail, Google Maps, Android, Google Play, Google Music, Google Voice, and others by providing background information for the users. The service of Google+ has the feature of Circles. It allows users to organize people into groups or lists for sharing across various Google products and services. In October 2013, Google counted 540 million active users who used at least one Google+ service\(^\text{16}\).

2.1.5 WeChat

WeChat is an instant message application of smart phone based in China provided by the company of Tencent Holdings in January 21, 2011\(^\text{17}\). WeChat is one of the most popular smart phone applications in China. Users can chat with their friends by voice or instant message via the application. Users can also share blogs, photos, video with their friends by publishing them in the “Circle of Friend”. The application is available on Android, iPhone, BlackBerry, Windows Phone and Symbian phones\(^\text{18}\). WeChat also provides services of contact and interact with users from other services such as Tencent QQ.

According the survey\(^\text{19}\), WeChat is the fifth most used smart phone application worldwide and in August 2013, it is reported that WeChat had had 100 million registered international users and 300 million registered Chinese users.

2.1.6 Summary

Here in Table2 we make a comparison of the above five popular social network services by listing the number of users, generally used area, initialized year and unique feature of each service. From Table2, we know the services of Facebook have the most number of registered users worldwide and the services of LinkedIn started first in the year of 2002. Sina Weibo and Weixi are Chinese-based services and mainly used in the Chinese circle. Other services are mainly used in the English circle and have users worldwide. Different services have different kinds of features: Sina

\(^{15}\) "Facebook’s Newest Challenger: Google Plus". Retrieved June 29, 2011.
\(^{18}\) Mittal Mandalia (2014-02-28). "WeChat announces native Mac client; Windows version may follow soon". technews.co.uk. Retrieved 2014-03-06.
\(^{19}\) "More 100 million WeChat users outside China". August 16, 2013.
Weibo encourages users to take part in the discussion of the hot topics. Facebook has the unique feature of creating real life event linking users in their real life. LinkedIn is mainly business-based and used for job hunting. Google+ can integrate users with other services and WeChat is a smart phone application and mainly used for sending instant messages similar to Line.

<table>
<thead>
<tr>
<th>SNS</th>
<th>Number of Users</th>
<th>Area</th>
<th>Initialized Year</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sina Weibo</td>
<td>503,000,000</td>
<td>China</td>
<td>2009</td>
<td>Topic Discussion</td>
</tr>
<tr>
<td>Facebook</td>
<td>1,000,000,000</td>
<td>United States and others</td>
<td>2004</td>
<td>Creating Real Life Event</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>212,300,000</td>
<td>United States and others</td>
<td>2002</td>
<td>Job Hunting</td>
</tr>
<tr>
<td>Google+</td>
<td>540,000,000</td>
<td>United States and Others</td>
<td>2011</td>
<td>Integration with Other Services</td>
</tr>
<tr>
<td>WeChat</td>
<td>400,000,000</td>
<td>China</td>
<td>2011</td>
<td>Mobile Application</td>
</tr>
</tbody>
</table>

### 2.2 Terminology Related To Twitter

Here we introduce some terminologies related to the social network service of Twitter.

#### 2.2.1 Tweet

A tweet is a message published by Twitter users within 140 characters. The tweets published by a user can be seen from the timeline of the user. In default situations, tweets are public and can be seen by any other users. However, the tweet publisher can adjust the settings so the tweets only can only be seen by the followers of the tweet publisher. Twitter users can publish tweets from the website of Twitter or some third-party applications.
Figure 1 shows an example of a tweet published by the user of The New York Times.

### 2. 2. 2 Retweet

A retweet is a re-posting of someone else’s Tweet. Twitter’s retweet feature can help users quickly share the tweet with all of the followers of the user. In Twitter, we have got the official function of retweet. However, Twitter users can publish tweets started with the character of “RT” to show that they are retweeting some tweets published by other users. Even though it is different from the official function of retweet in style, it is also a kind of retweet.

Figure 2 shows an example of official retweet of New York Times.

### 2. 2. 3 Reply

In Twitter, a reply is a public tweet started with “@UserName”. Reply is used to answer a specific tweet or talk to a specific person. Reply appears in the timeline of the users who are the followers of both the tweet creator and replied user. A reply is mainly a part of a conversation and usually the creator expects an answer for the reply.
Figure 3. Example of Reply

Figure 3 shows an example of reply of user New York Times.

2.2.4 Mention

A mention is a tweet containing “@UserName” starts with anything other than “@”, which is different from Reply. The tweet appears in the timeline of all of the followers of the tweet creator. Mention provides links of the profile of the user mentioned and creator do not expect an answer for the mention. The Mention is also used to evaluate the influence of the user mentioned.

Figure 4. Example of Mention

Figure 4 shows an example of user Durham Cathedral published a tweet mentioning the user of The New York Times.

2.2.5 List

A list is a curated group of Twitter users. Twitter users can create their own lists or subscribe to lists created by others. Viewing a list timeline will show a stream of Tweets from only the users on the list.
Figure 5. Example of List

Figure 5 shows the example of which lists the user of New York Times is belong to.

2.2.6 Hashflag

In order to increase the amount of users during the football gram of World Cup, the platform of Twitter started the new function of “Hashflag” in June 2014. Hashflag is a popular feature from the 2010 tournament that allows users to embed flag emoticons into their tweets\(^{20}\). Figure 6 shows an example of the new feature of “Hashflag”.

Figure 6. Example of Hashflag

2.3 Twitter API

Twitter is a public platform with more than 200 million tweets created a day\(^{22}\), which are of research and commercial values. Twitter developers can access to the data of Twitter via Twitter API. There are mainly three kinds of API in Twitter: Search API, Streaming API and REST API. Twitter developers are able to build and extend their applications integrating these Twitter APIs. Another import point about Twitter API is that the APIs are evolving by time.

2.3.1 Search API

The Twitter Search API allows Twitter developers to access to the data with specific keywords, tweets referencing a specific user or tweets from a specific user. However, the Search API is not exactly the same as the Search function in Twitter web services. It is important to notice that the Search API focus on relevance and some data may be missing. And the Streaming API can provide a dataset with relatively higher level of completeness.

\(^{20}\) https://twitter.com/nytimes/memberships.
\(^{22}\) “Getting Started [Twitter Developers]”, https://dev.twitter.com/start.
2. 3. 2 Streaming API

Twitter Streaming API provides real-time data with large amount for Twitter developers, which is different from REST API. In order to get access to Streaming API, Twitter developers need to build a long-lived HTTP connection and maintain it. In the process of utilizing Streaming API, developers can also set up some criteria to collect data related to some keywords or geotagged tweets from a specific region. The major disadvantage of Streaming API is that is only provides a sample of the real-time data.

2. 3. 3 REST API

The REST API provides developers with data related to a certain user, such as user timelines, status updates and other information. The REST API provides the history data of a certain user, different from the real-time data of Streaming API. Besides, developers can also interact with the platform by publishing tweets, following other users or updating the profiles of the user via REST API. In the utilization of Twitter REST API, very strict limit for developers and the rules of the REST API is evolving.

In our research, we mainly use the LIST API of Twitter. So we will try to introduce LIST API in detail.

Table 3. The Function of LIST API

<table>
<thead>
<tr>
<th>API Name</th>
<th>Function of the API</th>
</tr>
</thead>
<tbody>
<tr>
<td>GET lists/list</td>
<td>Returns all lists the authenticating or specified user subscribes to, including their own.</td>
</tr>
<tr>
<td>GET lists/statuses</td>
<td>Returns a timeline of tweets authored by members of the specified list.</td>
</tr>
<tr>
<td>GET lists/memberships</td>
<td>Returns the lists the specified user has been added to.</td>
</tr>
<tr>
<td>GET lists/members</td>
<td>Returns the members of the specified list.</td>
</tr>
<tr>
<td>GET lists/ownerships</td>
<td>Returns the lists owned by the specified Twitter user.</td>
</tr>
</tbody>
</table>
3. Related Research

In the previous research, most of the research about topics extraction in Twitter is based on the tweets published by users. The research related to influential user identification is based on friend/follower graph or retweet graph.

3.1 Research Related to Topic Extraction

In the previous research, most of the research of topic extraction is based on the tweets published by Twitter users. Our research solved the problem of topic extraction from a new perspective, using Twitter Lists.

Weng et al. [1] extracted user interest topic by collecting the tweets published by the users and applying the LDA model to calculate the probability of user interest. In the research, firstly they collected a set of top-1000 Singapore-based Twitter users from twitterholic.com, then they crawled all the followers and friends of each individual user, at last they crawled all the tweets user had published so far. In order to identify the topics that twitter users are interested in automatically according to the tweets users have published, they utilize the LDA (Latent Dirichlet Allocation) model, which is a popular topic model in text mining to extract topics from the tweets published by the users. In the process of applying LDA model, they regard all of the tweets published by one user as a document then in the results they obtain the topic interest of each user represented by extracted topics and the corresponding probability.

3.2 Research Related to Influential User Identification

In the previous research, most of the research ranked Twitter users based on friend/follower graph or retweet graph. Our research try to rank users and identify influential users based on the feature of Twitter List.

Weng et al. [1] proposed an algorithm called TwitterRank to measure the influence of users in Twitter. Considering both the factors of topical similarity between users and the link structure, the proposed topic-specific TwitterRank calculate the user influence score. It constructed the graph based on follower or friend relationship. Figure7 shows the general framework of TwitterRank. In topic-sensitive user influence ranking, they ranked Twitter users utilizing an algorithm similar to
However, it has been proved that in-degree, i.e. the number of followers, represents a user’s popularity, but is not related to other important notions of influence such as engaging audience, i.e., retweets and mentions \[2\]. In the research, they collect a large amount of data from Twitter and they measure the user influence from three aspect of indegree, retweets, and mentions. Based on these measures, they investigate the dynamics of user influence across topics and time. At last, they make the observation that most influential users can hold significant influence over a variety of topics.

Yamaguchi et al. \[3\] showed the concepts that to identify authoritative users, it is important to consider actual information flow in Twitter and existing approaches only deal with follower or friend relationship among users. They also proposed TURank (Twitter User Rank), in which an algorithm for evaluating users’ authority scores in Twitter based on link analysis based on graph called user-tweet graph which models information flow is introduced and ObjectRank in applied to evaluate users’ authority scores.

Kwak et al. \[4\] ranked Twitter users by the total number of followers, PageRank in the following/follower network and number of retweets in the diffusion network respectively. The results indicate that there exists a gap between the number of followers and the popularity of one’s tweets. Ranking Twitter users by the retweets of the original tweets they published show a new perspective of measure user influence.

Nakajima et al. \[5\] proposed methods to identify influential bloggers based on the knowledge level of the blogger. It calculates the scores of each blogger in each topic based on the domain-specific words. Blog is one kind of user-generated content. In the research, they measure the influence of blog authors based on the bloggers’ knowledge level considering the credibility of blog. The method calculates knowledge level. Bloggers’ knowledge level is evaluated based on their usage of domain-specific words in their past blog entries. In their method, blog entries written by knowledgeable bloggers have higher rankings than those written by common bloggers.

Wu et al. \[6\] ranked Twitter users by the frequency of being listed in each topic category. As exploiting the function of Twitter List is a method of classifying users \[6\], the name of the list can be
considered a meaningful label for the listed users. In their research, they define and identify certain predetermined topics including celebrity, media, organization, blog, ordinary, rather than extracting topics from the data automatically. However, defaults exist in their method: by defining the predetermined five categories of topics, they can only identify users related to the predetermined topics and it is less possible to discover categories other than the predetermined topics.

### 3.3 Research Related to Twitter List

Twitter List is an official function of Twitter initialized on November 2, 2009. The function of Twitter List is to help group users with a common feature and thereby realizing easy user management. In order to create a Twitter List, the creator needs to name the List and make a simple description about the List. Moreover, the creator can decide whether the List is public or private. After creating the List, the creator can add/edit/delete members of the List. Users other than the creator can subscribe to the List to receive the timeline of tweets published by the members of List.

Yamaguchi et al. [7] proposed a method of tagging Twitter users using Twitter list. Firstly, they extract tags from the name of Twitter List, which is the common feature of members of the List. Then they analyze the correlations among Twitter Lists, tags and users and proposed a method of tagging users and calculating their relevance level with the tag. However, there are many synonymous tags and name tags. Unfortunately, the proposed method ranks these many tags in higher ranks.

Ghosh et al. [8] proposed the method of expert sampling: If and only is a user has been listed at least 10 times for a certain topic, the user will be regarded as a ‘topical expert’. The conclusions show that the experts’ tweets are significantly richer in information content, cover more diverse topics, and more popular content. Experts’ tweets are also more trustworthy and they often capture breaking news stories earlier than random sampling. From the conclusions above, we know the method of expert sampling is an efficient method for generating content for content-centric applications, such as topical search, breaking news detection, trustworthy content recommendation and other applications.

Rakesh [9] proposed a method for recommending Twitter List to users considering the personal interests of users. They developed a ListRec model and the network-based LIST-PAGERANK model to recommend Lists to users who are interested in. Our research focuses on ranking Twitter users and recommending them to utilize the Twitter List.
3.4 Summary

The research above shows that the name of list is closely related to the interest topic of user and the number of times user has been listed can show the influence of the user. The previous research mainly focus on tweets published by users, friend/follower graph and retweet graph to identify topics and influential users in Twitter. In our research, we utilize the official function of Twitter List, to extract topics and rank Twitter users from a new perspective.

Our research is useful in the following aspects. 1) Personal user recommendation. Everyday a large amount of data is published in Twitter, and we need to find the information that is valuable for us. One of the methodologies of solving the problem is to follow the influential users in the area we are interested in. 2) Commercial Utilization. Commercial companies always use come celebrities to help promote their products. By utilizing the results of our research, companies can use the influential users in the area of the product to help promote the products effectively.
4. Methodology

In our research, we propose a new methodology to extract topics argued in Twitter by analyzing the name of Twitter List, then rank Twitter users based on their influence in each topic. By using Twitter List, we can easily extract influential users in each topic, because the name of Twitter List usually describes the attributes of its members, besides that the subscriber count of Twitter List reflects the influence of the members.

In the previous research, it has been proven that Twitter List is a function that can help identifying influential users in Twitter [7] [8]. In our research, we proposed a new way of extracting topics and identifying influential users by using Twitter List. We will introduce our proposed methodologies in details in the following part.

4.1 Twitter List

Twitter list, a new function adopted officially by Twitter can help user to manage their account. For example, Twitter user can create a list and include other users in the list then they can read the tweets published by the members included in the list directly. Users can also follow lists created by other users to read the tweets.

Figure 8. The Profile of New York Times

For example, in Figure 8, from the profile page of Twitter user New York Times, we can see many users include the user of New York Times as the list of news.

23 https://twitter.com/nytimes/memberships.
In the utilization of Twitter List, user can subscribe lists created by other users to read the whole time line tweets of users included in the list. The subscriber count of a List can reflect the influence of the list.

4.2 Dataset

Here we introduce a simple example to explain the words of user, list and tag in our research.

Figure 9. Example of Relationship between Twitter user, list and tag

Figure 9 explains the relationship between Twitter user, list and tag. For example, the Twitter user of “New York Times”, has been listed for 5 times. The name of Lists are “news”, “business”, “media”, “world news”, “US news” so we have the tags of “news” for three times, “business” for once, “media” for once, “world” for once and US for once.

In our research, we only focus on the users whose language is English, and the name of the list is English. In our dataset, we collected users whose language is English randomly by using Twitter API. Then, we collected all of the Twitter List in which the users had been included into. The Summary of collected data is shown below:

- 8,351 users
- 194,369 lists
- 231,376 tags

In the dataset, we only collected users who have been listed at least once. In one list, the name of the list contains several words and we treat one word as one tag. Moreover, in the process of extracting tags from the name of the list, we exclude characters other than English. We have 231,376 tags.
tags in total and they contain 37,125 unique English words.

Table 4. Top Tags in the Dataset Listed by Frequency

<table>
<thead>
<tr>
<th>Tag</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>my</td>
<td>8,053</td>
</tr>
<tr>
<td>news</td>
<td>5,960</td>
</tr>
<tr>
<td>music</td>
<td>4,515</td>
</tr>
<tr>
<td>you</td>
<td>2,664</td>
</tr>
<tr>
<td>gaming</td>
<td>2,242</td>
</tr>
<tr>
<td>i</td>
<td>2,044</td>
</tr>
<tr>
<td>the</td>
<td>2,024</td>
</tr>
</tbody>
</table>

In Table 4, we list the top tags extracted from the name of all of the lists in our dataset. From the table, we can understand the keyword “news”, “music”, and “gaming” stand for popular topics often appear in Twitter. The keywords “my”, “you”, “i”, and “the” are meaningless stop words, which need to be removed before applying the LDA model.
Figure 10 shows the distribution of Twitter List counts of the users in our experiment dataset. From Figure 10, we can understand that most of Twitter users have been listed below 500 times and only a few users have been listed for over 2,000 times.

4.3 Methodologies

Our research relies on the official function of Twitter List to identify topic and influential user in Twitter, which has not been done in the previous research. In our previous research [10], we tried to extract possible interest topics from the tweets published by users and rank users based on retweet relationship graph using LDA and PageRank algorithms.

Figure 11 shows the general flow of our methodologies. Firstly, we obtain the Twitter List resources from Twitter API. Secondly, we extract English tags from the name of the List resources. Next, we apply of LDA model to extract topics from the name of the lists. In the next step of influential user identification, we calculate the topic influence which is related to the subscriber count of Twitter List, and then calculate the user influence. We will introduce each step in details in the following parts.

![General Flow of Methodologies](image)
4. 3. 1 Obtaining Twitter List Resource

We obtain the Twitter List Resource by utilizing the Twitter API\(^2^4\). In our research, we mainly use the List API of REST API. We utilize the “GET list/memberships” API to obtain the lists the specified user has been added to. In the process, the id or screen name of the specified user is essential. In the List Resource, the name, slug, member count, subscriber count, information about the creator of the list and other information are provided. In the acquisition of List Resource, we have the important problems of the limitation of Twitter API. In the utilization of “GET list/memberships” API, we can only access it 15 times in every 15 minutes.

4. 3. 2 Tag Extraction

The creator of the Twitter List names the list and adds other users into the list. The name of the List can reflect the attributes of the members included in the List from some perspectives. We treat every word included in the name of the List as one tag.

4. 3. 1 Applying LDA model

Latent Dirichlet Allocation (LDA) model is a topic model proposed by David Blei, Andrew Ng, and Michael Jordan in the year of 2003. In the process of applying LDA model, all we need is the set of documents and the number of topics training with. In the results of LDA model, each document is viewed in the form of probability distributions and each topic is assumed to be characterized by the keywords and corresponding probability.

LDA model is widely used for topic extraction from texts. Recently, it is popularly used in the research of Twitter and other social network services for topic extraction.

In our research, we only focus on English word so before applying the LDA model, we use the regular expression matching to exclude the words other than English. We regard each tag extracted from the name of List as one word. If all of the characters composing the word are English characters or Arabic numerals, we regard it as an English word. Moreover, in order to improve the performance of topics extraction of LDA model, we also exclude the stop words defined in WordNet\(^2^5\), which is a lexical database for the language of English because the stop words are meaningless in the process of topic extraction.

\(^2^4\) Twitter Developers. https://dev.twitter.com/
\(^2^5\) http://wordnet.princeton.edu/
In applying the LDA model, the tags extracted from the name of the lists including the same user is regarded as one document and train the LDA model with 50 topics. As the results of the LDA model, we obtain two probability distribution metrics: $UserTopicProbability_{ij}$ and $KeywordTopicProbability_{kj}$. $UserTopicProbability_{ij}$ represents the probability $User_i$ has interests in $Topic_j$. In the results of the LDA model, the topics extracted from the tags of the list are represented by several keywords and corresponding probability. $KeywordTopicProbability_{kj}$ stands for the probability $Keyword_k$ occupies in $Topic_j$. In the following step, we will utilize these probability metrics to rank Twitter users.

4.3.2 Topic Influence Calculation

In order to evaluate the influence of each topic extracted from the LDA model, we utilize $KeywordTopicProbability_{kj}$ to calculate the influence of each topic.

![Figure 12. Topic Influence Calculation](image)

Figure 12 shows the way how we calculate the influence score of each topic. In the influence calculation of $Topic_j$, $Topic_j$ is consist of $k$ keywords from $Keyword_1$ to $Keyword_k$ and their corresponding probability $KeywordTopicProbability_{kj}$. The name of $List_l$ consists of some keywords from $Keyword_1$ to $keyword_k$. We select the maximum value of $KeywordTopicProbability_{kj}$, where $keyword_k$ is included in the name of $List_l$, then multiply by the subscriber count of $List_l$ to calculate the topic score of $Topic_j$. 
\[
\text{Influence}(\text{Topic}_j) = \sum_{l=1}^{L} (\text{subscriber\_count}(\text{List}_l) \times \max_{k \mid \text{the name of List}_l \text{ consists of keyword}_k} (\text{KeywordTopicProbability}_{kj}))
\]

(Equation 1)

Equation 1 shows the calculation of topic influence. \(\text{subscriber\_count}(\text{List}_l)\) represents the number of users who subscribe the list which shows the importance of the list.

### 4.3.3 User Influence Calculation

In order to calculate the influence score of each user and identify the influential users, we utilize \(\text{UserTopicProbability}_{ij}\) and influence score of \(\text{Topic}_j\) to calculate the influence score of each user.

Figure 13 shows the methodologies of how to calculate the influence score of each user. In the results of the LDA model, \(\text{UserTopicProbability}_{ij}\) represents the probability \(\text{User}_i\) has in \(\text{Topic}_j\).
Influence(User) = \sum_{j=1}^{l} UserTopicProbability_{ij} \times Influence(Topic_j)

(Equation 2)

Equation 2 shows how to calculate the influence score of each user. We calculate the user general score according to how much probability the user has in each topic and the influence score of each topic. In this way, we can rank Twitter users and identify influential users.

We proposed a new methodology to identify topics and influential users in Twitter using Twitter List, which is a relatively new function of Twitter. In the previous research, it has been proven that Twitter List is helpful in expert identification, but only little research has been done to identifying topics and influential users in Twitter using Twitter Lists. That is where our research outweighs the previous research.
5. Results

In this part, we will show the results of LDA model. In 5.1, we list some of the topics and the corresponding keywords in the results of the LDA model. In 5.2, we list top 5 Twitter users ranked by our proposed methodologies. In 5.3, we list the top 5 users in ten of the selected topics.

5.1 Extracted Topics

Here, we list some of the results of the LDA model. In the results of the LDA model, we train the whole texts formed by the tags of Twitter List with 50 topics. In the results of the LDA model, we can obtain the keywords of each topic and the probability it occupies. Here we list 10 of the 50 topics and the corresponding keywords. We try to analyze the results of the LDA model from the keywords of each topic. Table1 shows the results of the LDA model.

Table 5. Topics in the LDA model (Sample)

<table>
<thead>
<tr>
<th>Topic ID</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>writers, gaming, authors, bloggers, asmsg, writing, political, startup, libertarian, fabulous</td>
</tr>
<tr>
<td>27</td>
<td>news, media, general, kesehatan, detik, arab, portal, sehat, detikhealth, magazines</td>
</tr>
<tr>
<td>5</td>
<td>health, big, digital, medical, woah, belovedcome, kesehatan, funnys, fitness, godspeedgreatcommission24</td>
</tr>
<tr>
<td>21</td>
<td>xbox, battlefield, apps, basketball, garena, notable, georgia, randomness, visual, shopping</td>
</tr>
<tr>
<td>29</td>
<td>info, favorite, entertainment, life, public, somerset, local, kesehatan, ios, expert</td>
</tr>
</tbody>
</table>
In Table 5, we can understand the topics from the related keywords. In Topic 37, the keywords “writers”, “authors”, “bloggers”, and “writing” mean the topic stands for “writing”. In Topic 27, the keywords “news”, “media”, and “magazines” mean the topic stands for “media”. In topic 5, the keywords “health”, “medical”, and “fitness” mean the topic stands for “health care”. In Topic 21, the keywords “xbox”, “apps”, “basketball”, and “visual” mean the topic stands for “video games”. In topic 33, the keywords “music” and “dj” mean the topic stands for “music”.

However, in some of the topics, we cannot understand the topics by analyzing the keywords of the topic, such as Topic 29, 40, 2, 8, 14 listed in Table 5. Even though we utilize the name of the Twitter List, which better describes the attributes of users, not all of the topics are understandable and how to extract more understandable topics remains for future work.

5.2 Top Users ranked by our methodologies

We ranked Twitter users of our dataset by our proposed methodologies described in Section 4.

Table 6. Top 10 Users Ranked by Proposed Methodology

<table>
<thead>
<tr>
<th>Rank</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>music, dj, berita, funny, art, follow, frasil, ceo, jokes, online</td>
</tr>
<tr>
<td>40</td>
<td>twitter, pretty, teamfollowback, fresh, twitterblades, arizona, promising, girl, sufc</td>
</tr>
<tr>
<td>2</td>
<td>top, cool, artist, tweets, sexy, federal, promoters, promo, engineers, money</td>
</tr>
<tr>
<td>8</td>
<td>edm, politics, bristol, boobs, official, law, aggies, legal, positive, australian</td>
</tr>
<tr>
<td>14</td>
<td>win, industry, support, quotes, happy, test, blogs, resources, ecommerce, logistics</td>
</tr>
</tbody>
</table>
All of the Twitter users listed above are normal Twitter users, not celebrities or news media. However, they have got a large number of friends and publish many tweets so they take active part in the platform of Twitter. By following these top users who got a higher score over all of the topics, we are more likely to start a conversation and make friends with them.

### 5.3 Influential Users of Selected Topic

In order to verify our proposed method can identify influential users in the understandable topics, we selected the top 5 users in each topics of the above table. Table 7 shows the users who have got the top 5 scores in the selected topic.
## Table 7: Influential Users in Each Topic

<table>
<thead>
<tr>
<th>User Rank</th>
<th>Topic 37</th>
<th>Topic 27</th>
<th>Topic 5</th>
<th>Topic 21</th>
<th>Topic 29</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Custard_Corn er</td>
<td>MalaysiaGaz ette</td>
<td>Nutta_Kim</td>
<td>OHAshlaay_</td>
<td>stebuzzm15</td>
</tr>
<tr>
<td>2</td>
<td>Deathstrike2014</td>
<td>maalaimalar</td>
<td>_<em>ImaBIGDeal</em></td>
<td>claudinblazin</td>
<td>Marcus_Salnic__ky</td>
</tr>
<tr>
<td>3</td>
<td>fitrian</td>
<td>abarandun</td>
<td>racingrandy80</td>
<td>ayeitsangelic_aaa</td>
<td>muralipiyyer</td>
</tr>
<tr>
<td>4</td>
<td>Braahna</td>
<td>dinamalarweb</td>
<td>Kitch2202</td>
<td>ogmob_</td>
<td>__bnger__x</td>
</tr>
<tr>
<td>5</td>
<td>GhostFaceK eylw</td>
<td>YarmoukNews_s</td>
<td>P01YN0NYM0U55</td>
<td>bugctti</td>
<td>captainrupert</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User Rank</th>
<th>Topic 33</th>
<th>Topic 40</th>
<th>Topic 2</th>
<th>Topic 8</th>
<th>Topic 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ballin_Shibby</td>
<td>ShanCabrera</td>
<td>jaanecortes</td>
<td>EsteebanVilla</td>
<td>cesar_m67</td>
</tr>
<tr>
<td>2</td>
<td>teqnop</td>
<td>kijefemuqas</td>
<td>K_Street01</td>
<td>Leslie_Beatdown</td>
<td>Prince_Taylor91</td>
</tr>
<tr>
<td>3</td>
<td>CancelDatB_tch</td>
<td>zesuqemorun</td>
<td>senshine_wind</td>
<td>leeners_</td>
<td>princessjass16</td>
</tr>
<tr>
<td>4</td>
<td>Syuhazulkifli</td>
<td>andr3wkell</td>
<td>hanna_robertss</td>
<td>OscarSteezz</td>
<td>JinXs__</td>
</tr>
<tr>
<td>5</td>
<td>RagingNeurotic</td>
<td>qucaxukyzojo</td>
<td>Kawtarares</td>
<td>Frantavator</td>
<td>Kaykillaaaaa_Ho</td>
</tr>
</tbody>
</table>
6. Evaluations

In the previous researches, they often refer influential Twitter users as the users who have many followers. However, the users with many followers are often famous celebrities or news media or organizations. Some of them are inactive Twitter users and they rarely have activities in Twitter. We evaluate our proposed methodologies from two perspectives: 1) Top users with higher scores over all of the topics are active users using the platform of Twitter actively. 2) In each of the selected topic, the top users always publish messages with value related to the topic but get relatively few followers.

In order to verify our proposed method can identify influential users which cannot be identified by ranking by the number of followers, we regard the method of ranking users by the number of followers as baseline and compare the two methods in general score.

6.1 Evaluation by Questionnaire

As tentative evaluation, we selected five topics for evaluation. In each topic, we select top 5 users in both our proposed method and baseline. In each of the selected topic, we treat both the top 5 users ranked by our method and the baseline as one group. We asked 6 students to grade each group of users making a score from 1, 2, 3, 4, to 5 according to their influence in the topic, then calculate the average score of each group of users. Here, higher score represents higher influence.

Table 8. Questionnaire Results of Each Topic

<table>
<thead>
<tr>
<th>Topic Id</th>
<th>37</th>
<th>27</th>
<th>5</th>
<th>21</th>
<th>29</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Method</td>
<td>3.70</td>
<td>3.33</td>
<td>3.33</td>
<td>3.80</td>
<td>1.83</td>
<td>3.20</td>
</tr>
<tr>
<td>Rank by Follower Number</td>
<td>3.22</td>
<td>3.27</td>
<td>2.83</td>
<td>3.63</td>
<td>2.23</td>
<td>3.03</td>
</tr>
</tbody>
</table>
In each of the selected topic, we calculate the average of each user by the total score received from the respondents divided by the number of respondents. Table 8 shows the average score of the users ranked top five by our proposed method, ranking by the number of followers and the tweet amount. The table shows our proposed method can identify influential users related to the topic with higher average score in more topics than ranking by the number of followers and the number of published tweet. Our proposed method identified the influential users who often publish related messages but have relatively less followers.

6.2 Evaluation by User Message Amount in General

Score

![Comparison of Tweet Amount](image)

In this part, we regard the method of ranking by number of followers as baseline. Figure 14 shows the comparison of the number of tweets tweeted by the top users ranked by our proposed method and
the baseline. From the figure, we can easily understand users ranked top by our proposed method outweigh the baseline from the aspect of published tweet amount. Our method can identify the influential users who take active part interaction in Twitter and publish a large amount of information.

6.3 Evaluation by Topical Difference

In order to prove the specialty of the top users ranked by our proposed method, i.e., to prove that they are expert in a few areas, not those who publish messages randomly in a wide range of areas, we use the topical difference defined in paper [1] to help us evaluate our proposed methodologies and baseline.

Firstly, we normalize the metrics of UserTopicProbability as $UserTopicProbability'$ such that $||UserTopicProbability'|| = 1$ for each row $UserTopicProbability_i'$. We defined the topical difference between two Twitter user $u_i$ and $u_j$ as Equation 3.

$$
\text{dist}(i,j) = \sqrt{2 \ast D_{JS}(i,j)}
$$

(Equation 3)

$D_{JS}(i,j)$ is the Jensen-Shannon Divergence between the two probability distributions $UserTopicProbability_{i'}$ and $UserTopicProbability_{j'}$, which is defined as:

$$
D_{JS}(i,j) = \frac{1}{2} \left( D_{KL}(UserTopicProbability_{i'} || M) + D_{KL}(UserTopicProbability_{j'} || M) \right)
$$

(Equation 4)

$M$ is the average of the two probability distributions, i.e. $M = \frac{1}{2} (UserTopicProbability_{i'} + UserTopicProbability_{j'})$. $D_{KL}$ is the Kullback-Leibler Divergence which defines the divergence from distribution Q to P as: $D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$. 

32
We use the methods described above to calculate the topical difference of every pair of users in our experiment dataset. We calculate the total topical difference of top-k users between every other user in the dataset. We select top-k users ranked by both our proposed methodology and baseline.

In Figure 16, we can understand the top users ranked our proposed methodologies have more topical difference between other users in the dataset than the top users ranked by the number of

<table>
<thead>
<tr>
<th>K</th>
<th>Proposed Method</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9476.45</td>
<td>8941.47</td>
</tr>
<tr>
<td>5</td>
<td>47382.25</td>
<td>44464.85</td>
</tr>
<tr>
<td>10</td>
<td>94764.51</td>
<td>89198.66</td>
</tr>
<tr>
<td>20</td>
<td>189404.30</td>
<td>179816.50</td>
</tr>
<tr>
<td>30</td>
<td>283977.46</td>
<td>270679.70</td>
</tr>
</tbody>
</table>
followers. It means the top users identified by our proposed methodologies have relatively limited range of topical areas than the users ranked by the number of followers.

We compare the top users with the baseline from the above three perspectives. It has been shown that the users ranked by our proposed methodologies are more active in interaction, more involved in the platform of Twitter and have more topical difference with other users in the dataset.

6.4 Summary

We evaluate our proposed methodology from the above two perspectives by three methods. From the evaluation of general score by comparing tweet amount and topical difference, we know our proposed methodologies can identify influential users who use the platform of Twitter actively and have limited topic area. From the evaluation of user influence in five of the selected area by questionnaire, we know our proposed methodologies can identify influential users related to the specific topic with higher scores compared to ranking by tweet amount and follower number. Our method can successfully extract topics and identify influential users in Twitter.
7. Conclusions

In our research, we proposed a new method of topics and influential user identification in Twitter by using Twitter Lists. Twitter List, which is relatively new function of Twitter, has been studied recently. With the help of the name of list, we can successfully identify the interest topics of Twitter users. Moreover, with the help of subscriber count of Twitter List, namely, the follower of the Twitter List, which reflects the popular level of Twitter List, we identify influential Twitter users who cannot be identified by other methodologies.

From the analysis of the evaluation of our results in Chapter 6, our proposed methodologies can identify influential user more active in interaction, more involved and show more topical interest compared to the users ranked top by the number of followers. As a conclusion, which is also proved by other papers, there exists a gap between the users who play an important role in Twitter and those users with many followers. In Twitter, many users follow other celebrities or organizations because of their huge influence in real life, not because of their influence of information spread in Twitter.

Our research can identify the users who are really influential in Twitter, not those users with many followers. Especially, we utilize the official function of Twitter List, to help identify topics and influential users, which have not been done in the research before.
I would like to give many thanks to Professor Yamana and all of the students of Yamana Lab. They gave me a lot of advice and encouragement during my two-year master’s courses. I would like to thank them very much.
References


### Appendix1

Table 1. Topics extracted from the name of Twitter List in the results of LDA model

<table>
<thead>
<tr>
<th>Topic ID</th>
<th>Keywords and Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.066<em>stay + 0.063</em>im + 0.054<em>eminem + 0.048</em>ilysm + 0.046<em>fb + 0.044</em>filipino + 0.036<em>favesss + 0.035</em>listed + 0.031<em>thx + 0.031</em>awsome</td>
</tr>
<tr>
<td>1</td>
<td>0.205<em>ur + 0.124</em>clash + 0.073<em>important + 0.073</em>ppl + 0.041<em>vancouver + 0.032</em>lovatics + 0.025<em>chlo + 0.019</em>supernatural + 0.018<em>ultimate + 0.017</em>hometown</td>
</tr>
<tr>
<td>2</td>
<td>0.476<em>top + 0.178</em>cool + 0.065<em>artist + 0.060</em>tweets + 0.030<em>sexy + 0.023</em>federal + 0.014<em>promoters + 0.013</em>promo + 0.008<em>engineers + 0.005</em>money</td>
</tr>
<tr>
<td>3</td>
<td>0.282<em>tottenham + 0.193</em>lemme + 0.087<em>candy + 0.061</em>hip + 0.055<em>hot + 0.034</em>hiphop + 0.021<em>young + 0.016</em>favstar + 0.015<em>horny + 0.014</em>thefunny</td>
</tr>
<tr>
<td>4</td>
<td>0.050<em>jville + 0.041</em>uc + 0.038<em>khan + 0.037</em>ncat + 0.033<em>future + 0.031</em>encouraging + 0.030<em>progressives + 0.030</em>teamfollowback2 + 0.028<em>seguidors + 0.025</em>person</td>
</tr>
<tr>
<td>5</td>
<td>0.207<em>health + 0.063</em>big + 0.035<em>digital + 0.026</em>medical + 0.024<em>woah + 0.024</em>belovedcome + 0.019<em>kesehatan + 0.018</em>funnys + 0.018<em>fitness + 0.018</em>godspeedgreatcommission24</td>
</tr>
<tr>
<td>6</td>
<td>0.374<em>attractive + 0.046</em>ashton + 0.033<em>sledui + 0.032</em>seattle + 0.027<em>smile + 0.026</em>idk + 0.025<em>poltiie + 0.017</em>unt18 + 0.016<em>tl + 0.012</em>kirov</td>
</tr>
<tr>
<td>7</td>
<td>0.490<em>djs + 0.088</em>producers + 0.060<em>business + 0.027</em>watch + 0.025<em>deejays + 0.022</em>amazing + 0.021<em>rt + 0.017</em>dzj + 0.011<em>security + 0.009</em>rap</td>
</tr>
<tr>
<td>8</td>
<td>0.191<em>edm + 0.068</em>politics + 0.064<em>bristol + 0.052</em>boobs + 0.051<em>official + 0.039</em>law + 0.033<em>aggies + 0.025</em>legal + 0.025<em>positive + 0.022</em>australian</td>
</tr>
<tr>
<td>9</td>
<td>0.214<em>voices + 0.210</em>egypt + 0.050<em>ekechukwu + 0.046</em>excellent + 0.038<em>mktg + 0.033</em>games + 0.030<em>web + 0.019</em>hedge + 0.019<em>loyal + 0.019</em>community</td>
</tr>
<tr>
<td>10</td>
<td>0.253<em>list + 0.078</em>sammw14 + 0.072<em>outlast + 0.028</em>quote + 0.023<em>de + 0.022</em>alpha + 0.021<em>myfriends + 0.020</em>listei + 0.019<em>editors + 0.018</em>talk</td>
</tr>
<tr>
<td>11</td>
<td>0.134<em>tgdn + 0.131</em>iphone + 0.037<em>molon + 0.031</em>stars + 0.029<em>lifestyle + 0.028</em>yoga + 0.026<em>american + 0.025</em>kitchen + 0.025<em>alt + 0.025</em>lose</td>
</tr>
<tr>
<td>12</td>
<td>0.485<em>awesome + 0.069</em>top5000ceos + 0.069<em>chicago + 0.062</em>thisisntporn + 0.035<em>philly + 0.018</em>salute + 0.014<em>anne + 0.014</em>lds + 0.012<em>warriors + 0.008</em>oracle</td>
</tr>
<tr>
<td>13</td>
<td>0.673<em>people + 0.054</em>verified + 0.029<em>pro + 0.024</em>wrestling + 0.020<em>super + 0.013</em>engagement + 0.013<em>vol + 0.012</em>adidas + 0.011<em>wwe + 0.006</em>fblog</td>
</tr>
<tr>
<td>Page</td>
<td>Text</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>14</td>
<td>0.149<em>win + 0.043</em>industry + 0.041<em>support + 0.038</em>quotes + 0.033<em>happy + 0.027</em>test + 0.027<em>blogs + 0.025</em>resources + 0.023<em>ecommerce + 0.022</em>logistics</td>
</tr>
<tr>
<td>15</td>
<td>0.276<em>syria + 0.189</em>social + 0.083<em>automotive + 0.040</em>onedirectiongame + 0.029<em>club + 0.027</em>sos + 0.022<em>favoritos + 0.018</em>creativert + 0.018<em>car + 0.016</em>branding</td>
</tr>
<tr>
<td>16</td>
<td>0.114<em>world + 0.112</em>middleweight + 0.077<em>dope + 0.057</em>bad + 0.052<em>quality + 0.036</em>favess + 0.035<em>neo + 0.024</em>ahli + 0.023<em>periodistas + 0.022</em>hotties</td>
</tr>
<tr>
<td>17</td>
<td>0.321<em>good + 0.048</em>celeb + 0.038<em>blue + 0.034</em>patriots + 0.026<em>politic + 0.024</em>webcomics + 0.023<em>list91 + 0.023</em>knowledge + 0.018<em>ladies + 0.015</em>mostretweeted</td>
</tr>
<tr>
<td>18</td>
<td>0.467<em>gta + 0.082</em>befit4less + 0.077<em>bbm + 0.066</em>download + 0.065<em>psn + 0.057</em>pes + 0.051<em>working + 0.032</em>ps4 + 0.014<em>britney + 0.008</em>gold</td>
</tr>
<tr>
<td>19</td>
<td>0.143<em>radio + 0.102</em>austin + 0.071<em>bookworms + 0.065</em>promote + 0.054<em>tobacco + 0.042</em>mahomes + 0.036<em>south + 0.034</em>teman + 0.019<em>list44 + 0.018</em>silly</td>
</tr>
<tr>
<td>20</td>
<td>0.162<em>founders + 0.075</em>comics + 0.028<em>armenia + 0.027</em>inhai + 0.020<em>gossip + 0.019</em>listings + 0.019<em>fun2 + 0.019</em>lirocka + 0.017<em>ksa + 0.017</em>fk</td>
</tr>
<tr>
<td>21</td>
<td>0.757<em>xbox + 0.039</em>battlefield + 0.027<em>apps + 0.025</em>basketball + 0.015<em>arena + 0.011</em>notable + 0.009<em>georgia + 0.008</em>randomness + 0.007<em>visual + 0.007</em>shopping</td>
</tr>
<tr>
<td>22</td>
<td>0.389<em>giveaway + 0.167</em>influential + 0.088<em>beautiful + 0.054</em>yall + 0.043<em>workinghackccoc + 0.035</em>dms + 0.021<em>men + 0.019</em>hackclashofclans + 0.010<em>nieuws + 0.009</em>niall</td>
</tr>
<tr>
<td>23</td>
<td>0.261<em>shethatbitch + 0.124</em>creative + 0.070<em>fashion + 0.052</em>main + 0.030<em>list82 + 0.023</em>rochester + 0.022<em>gleeks + 0.019</em>boingboing + 0.018<em>teamafrica + 0.016</em>xoxo</td>
</tr>
<tr>
<td>24</td>
<td>0.858<em>free + 0.056</em>pearl + 0.010<em>football + 0.009</em>nike + 0.003<em>edc + 0.003</em>wont + 0.002<em>aion + 0.001</em>schoolmates + 0.001<em>feu + 0.001</em>horse</td>
</tr>
<tr>
<td>25</td>
<td>0.515<em>youwanttolookbetter + 0.157</em>jbsbelieve3dmovie + 0.049<em>millionaire + 0.021</em>louis + 0.020<em>upper + 0.016</em>register + 0.012<em>traffic + 0.011</em>rising + 0.011<em>diary + 0.009</em>indpolitics</td>
</tr>
<tr>
<td>26</td>
<td>0.062<em>amigos + 0.056</em>kata + 0.055<em>motivasi + 0.055</em>edc + 0.046<em>hiburan + 0.037</em>te + 0.027<em>artistas + 0.022</em>justforfun + 0.022<em>inspirator + 0.021</em>hottest</td>
</tr>
<tr>
<td>27</td>
<td>0.539<em>news + 0.120</em>media + 0.031<em>general + 0.029</em>kesehatan + 0.022<em>detik + 0.016</em>arab + 0.008<em>portal + 0.006</em>sehat + 0.006<em>detikhealth + 0.006</em>magazines</td>
</tr>
<tr>
<td>28</td>
<td>0.243<em>perfect + 0.215</em>friends + 0.103<em>fav + 0.093</em>artists + 0.081<em>favs + 0.059</em>love + 0.053<em>fave + 0.022</em>ill + 0.009<em>youre + 0.006</em>list95</td>
</tr>
<tr>
<td>29</td>
<td>0.184<em>info + 0.148</em>favorite + 0.077<em>entertainment + 0.055</em>life + 0.046<em>public + 0.025</em>somerset + 0.019<em>local + 0.019</em>kesehatan + 0.018<em>ios + 0.017</em>expert</td>
</tr>
</tbody>
</table>
0.180*favretweeters + 0.118*arika + 0.112*teamfollowback1 + 0.060*teamfollowback3 +
0.049*stanley + 0.047*con + 0.034*halers + 0.028*labournstaffing + 0.021*ukip +
0.018*freinds

0.209*worldshottestjeans + 0.065*harry + 0.063*filipiniggas + 0.058*wwa + 0.038*hey +
0.031*code + 0.023*mah + 0.023*ikaw + 0.022*ebay + 0.019*spotifyfreepremium

0.171*cod + 0.103*slander + 0.081*fifa + 0.073*nba + 0.070*mena + 0.036*spor +
0.033*list80 + 0.024*latest + 0.023* louisiana + 0.022*da

0.309*music + 0.297*dj + 0.076*berita + 0.022*funny + 0.017*art + 0.015*folllow +
0.013*frasil + 0.010*ceo + 0.008*jokes + 0.008*online

0.103*sailing + 0.092*bizitalk + 0.080*back + 0.049*glee + 0.045*los +
0.044*cape + 0.028*cnet + 0.028*wc + 0.026*call

0.176*fans + 0.106*auto + 0.089*rappers + 0.083*trend + 0.031*cars + 0.025*jail +
0.023*list27 + 0.020*fan + 0.019*shout + 0.017*musician

0.727*girls + 0.040*bitcoin + 0.025*major + 0.017*lost + 0.016*back + 0.013*crypto +
0.013*young + 0.007*women + 0.006*author + 0.006*read

0.152*writers + 0.090*gaming + 0.087*authors + 0.076*bloggers + 0.049*asmsg +
0.045*writing + 0.031*political + 0.030*startup + 0.029*libertarian + 0.024*fabulous

0.199*team + 0.087*musicians + 0.055*global + 0.048*politico + 0.045*filmmakers +
0.045*network + 0.038*entrepreneur + 0.034*mexico + 0.026*diacritical + 0.021*actors

0.763*follow + 0.153*baddies + 0.008*diacritical + 0.007*sweethearts + 0.004*bff +
0.004*thefreaks + 0.003*taylor + 0.002*ginger + 0.002*potatoes + 0.002*weenies

0.632*twitter + 0.169*pretty + 0.025*teamfollowback + 0.014*fresh + 0.010*2 +
0.009*twitterblades + 0.008*arizona + 0.006*promising + 0.006*girl + 0.005*sufc

0.502*real + 0.097*faves + 0.089*fab + 0.052*beliebers + 0.019*bieber + 0.015*ily +
0.015*beyonce + 0.014*true + 0.012*justin + 0.012*fuck

0.221*cute + 0.208*eat + 0.119*celebs + 0.049*ps3 + 0.049*tha + 0.039*gtav +
0.039*sports + 0.016*sas + 0.016*sp + 0.012*entertainers

0.343*myfollowers + 0.139*interior + 0.063*fulfillingtheneed + 0.061*justice +
0.044*goodspeed + 0.038*guitar + 0.035*palestine + 0.025*peeps + 0.025*marketing +
0.022*bds

0.322*imma + 0.159*active + 0.076*list1 + 0.050*alotfollowers + 0.039*tweeps +
0.032*xxx + 0.023*friend + 0.021*respect + 0.014*diacritical + 0.013*animal

0.770*check + 0.068*vip + 0.027*trill + 0.014*family + 0.012*owls + 0.012*uniteblue +
0.011*christian + 0.006*play4free + 0.005*dedicated + 0.005*yea

0.320*great + 0.099*middle + 0.063*breast + 0.062*jobs + 0.035*clashofclanshack +
| 47 | 0.027*architecture + 0.023*pelajarkhilafah + 0.022*qollector + 0.014*open + 0.014*islam |
| 48 | 0.129*chaturbate + 0.114*followfriday + 0.102*book + 0.062*nice + 0.060*retweeters + 0.047*famous + 0.044*ff + 0.039*random + 0.024*smart + 0.017*daily |
| 49 | 0.169*uk + 0.150*interesting + 0.075*tech + 0.043*tweet + 0.038*science + 0.034*zazzlers + 0.032*zazzle + 0.025*international + 0.019*education + 0.017*python |
| 49 | 0.132*list79 + 0.100*steam + 0.082*fellow + 0.066*celebrities + 0.064*leaked + 0.063*ea + 0.034*gamers + 0.024*amro + 0.021*celebz + 0.019*internet |
Appendix2 Publications
