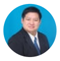
A Comparative Study of HITS vs PageRank Algorithms for Twitter Users Analysis

Conference Paper · August 2014

DOI: 10.1109/ICCST.2014.7045007

Poo Kuan Hoong
Multimedia University



Chiung Ching Ho
Multimedia University

A Comparative Study of HITS vs PageRank Algorithms for Twitter Users Analysis

Ong Kok Chien, Poo Kuan Hoong and Chiung Ching Ho
Faculty of Computing & Informatics
Multimedia University
Cyberjaya, 63100, Selangor, Malaysia
Email: ong.kok.chien14@student.mmu.edu.my, khpoo@mmu.edu.my, ccho@mmu.edu.my

***Abstract*—Social Networks such as Facebook, Twitter, Google+ and LinkedIn have millions of users. These networks are constantly evolving and it is a good source of information, both explicitly and implicitly. The analysis of Social Network mainly focuses on the aspect of social networking with an emphasis on mapping relationships, patterns of interaction between user and content information. One of the common research topics focuses on the centrality measures where useful information of the connected people in the social network is represented in a graph. In this paper, we employed two link-based ranking algorithms to analyze the ranking of the users: HITS (Hyperlink-Induced Topic Search) and PageRank. We constructed Twitter user retweet-relationship graph using 21 days worth of data. Lastly, we compared the ranking sequence of the users in addition to their followers count against the average and also whether they are verified Twitter accounts. From the results obtained, both HITS and PageRank showed a similar trend, and more importantly highlighted the importance of the direction of the edges in this work.**

***Index Terms*—Twitter, Social Network, HITS, PageRank, Big Data, Graph**

I. Introduction

Twitter is one of the largest social networking platforms that provides free microblogging (i.e. tweets) services. A recent IPO filing by Twitter has reported that it has 218 million active users comprising 163.5 million monthly mobile users and 100 million daily users. 500 million tweets per day are generated by these users [1]. Twitter has been used extensively for customer relationship management, marketing and branding, propagation of news, as well as election campaigns. Twitter users are not limited to individual social media users but also include businesses, news agencies, civil authorities and even political entities. Twitter's ability to instantaneously broadcast and receive information is the main reason it is broadly used by users from multiple domains.

As Twitter is widely used by different types of user, the interaction between user is not restricted to friends and families. One of the unique features of Twitter is the ability to broadcast messages not only to mass audiences, but as well, to selected groups of people. In order to simplify the topics categorization, a Hashtag is introduced by adding # symbol before a relevant keyword or a CamelCase phrase without spaces (i.e.: "#NowWatching" instead of "#nowwatching") and it can be placed in any position within tweets [2]. However, an irrelevant hashtags may degrade the quality of the topic clusters.

Tweets generated by Twitter users can undergo data-mining to generate new information and insights. These information and insights may be considered as a form of refined Twitter resources. The use of Twitter resources is demonstrated by Twitter through their *Discover* functionalities such as *Trend Recognition* and *Personalized Tweets/Users recommendation*. By following the @*MagicRecs* Twitter account [3][4], Twitter users will receive Personalised Tweets/Users recommendation. The economic worth of Twitter resources can be indirectly proved by the existence of Certified Data Reseller (CDR) like Gnip, DataSift and NTT DATA. These agents re-sell reliable Twitter resources to their clients, via their access to Twitter's Firehose. The Firehose provides 100% access to public profile tweets without any loss, whereas Twitter Streaming API captures 43.5% of tweets available in Firehose [5].

With the vast amount of tweets published by users, it is crucial to rank users in order to differentiate relevant important information providers from those provided by spam accounts. In this paper, we employed two Link-Based Ranking (LBR) algorithms namely HITS and PageRank that are mainly used for web page ranking to rank Twitter users. We focus on the analysis of Retweet (RT) relationship constructed at the Twitter users level with the overall quality of the tweets posted by the users. Every Twitter user is represented as the node/vertex and the RT action is represented as the directed-link/directed-edge.

This paper is organized as follows: Section II explains the basic working mechanisms of the HITS algorithm and PageRank algorithm, and is followed by examples of applications utilizing the HITS and PageRank algorithms. We subsequently highlight and discuss other approaches for Twitter users ranking as described in published literature. Section III describes the details of the authors' proposed implementation of the HITS and PageRank algorithms for Twitter users rank analysis. Section IV provides the comparison results of the top 20 ranked Twitter users using both the HITS and PageRank algorithms. Lastly, section V concludes the research findings and provides future work to be conducted.

II. Literature Review

Twitter is a social media platform that is differentiated through intentional concise messages with embedded links, and unidirectional follower-followee relationship (i.e. subscription *without* prior approval [6]. The open and assymetrical structure of Twitter's network allows every Twitter user to interact with another Twitter user, unlike the 'friending' requirement of Facebook for instance. Twitter enables user to share condensed information consisting of a maximum of 140 characters per tweet [7]. The brevity of a typical Tweet and the extensive usage of colloquialism and abbreviations present a challenge for the data-mining of tweets. Users of Twitter have the ability to *Retweet* (RT) or *Favourite* (FAV) other peoples' tweet that they find value in reading and sharing.

Retweet was started by Twitter's third party application to introduce re-sharing of others' tweet by appending "RT" in front of the original author screen name followed by the original tweet [8] [9]. However, due to the additional metadata and keywords with the restriction of 140 characters per tweet, users tend to modify the original tweet with shortforms or abbreviations. Twitter supports the retweet action natively with the intention to reduce redundancy of tweets shown on user's timeline, and at the same time promote worthy sharing of tweets and interesting users. Each retweet action contributes to the attribute of the original tweet which helps Twitter to remove redundant streams of tweets by different users who retweet the same content [10]. It also reserves the original content without any additional modification. When a retweet (noun) is retweeted (verb), the retweet_count in Twitter Retweet API contribute back to the original tweet [11]. Assuming there is a chain of retweet, Tweet A (the original tweet) is retweeted by Tweet B and Tweet C retweeted Tweet B. The flows of information are as follows: $[A] \rightarrow [B] \rightarrow [C]$. In Twitter Retweet API, only Tweet A entitled the increase in retweet_count, i.e. the flows of information are as shown: $[A] \rightarrow [B]$, $[A] \rightarrow [C]$. The flows of information mentioned above consider a origin-to-references view of the data. In this paper, we observed the data from the references-to-origin which is how the backlink of web pages work in the Internet. Thus, a connected graph can be constructed which is represented as follows: $[B] \rightarrow [A]$, $[C] \rightarrow [A]$

A. HITS Algorithms

Hyperlink-Induced Topic Search (HITS) algorithms was introduced by Jon Kleinberg in 1998 to solve the problem of ranking web pages [12]. HITS computes hub and authority score for each of the node. From a high level view, the higher the number of outgoing (incoming) edges, the higher the Hub (Authority) score of that particular node. A node's Hub (Authority) score is computed by summing up all the Authority (Hub) score of the outgoing (incoming) edges' node.

HITS is employed in varies domain, including search engines, journal citation analysis, and social networks. For example, "CLEVER" was built on top of HITS algorithm at IBMs Almaden Research Lab in San Jose, CA [12] to improves the performance of Internet search engines. Besides that, HITS is also used to analyse co-citation and co-reference in the field of citation analysis and bibliometrics [13]. In recent years, researchers have been using the HITS algorithm to solve problems such as spam detection [14].

Algorithm 1 HITS

```
G := set of users
procedure INITIALIZATION($G$)
    for each user $U$ in $G$ do
        $U_a \leftarrow 1$
        $U_h \leftarrow 1$
    end for
end procedure
for step from $1$ to k do
    procedure AUTHORITYUPDATE($G$)
        $norm \leftarrow 0$
        for each user $U$ in $G$ do
            $U_a \leftarrow 0$
            for each user $V$ in $U.RTbyUsers$ do
                $U_a$ += $V_h$
            end for
            $norm$ += $(U_a)^2$
        end for
        $norm \leftarrow \sqrt{norm}$
        procedure NORMALIZATION($G, norm$)
            for each user $U$ in $G$ do
                $U_a \leftarrow U_a/norm$
            end for
        end procedure
    end procedure
    procedure HUBUPDATE($G$)
        $norm \leftarrow 0$
        for each user $U$ in $G$ do
            $U_h \leftarrow 0$
            for each user $W$ in $U.RTingUsers$ do
                $U_h$ += $W_a$
            end for
            $norm$ += $(U_h)^2$
        end for
        $norm \leftarrow \sqrt{norm}$
        procedure NORMALIZATION($G, norm$)
            for each user $U$ in $G$ do
                $U_h \leftarrow U_h/norm$
            end for
        end procedure
    end procedure
end for
```

Note:
$U.RTbyUsers$ = List of users that Retweeted User U's tweet.
$U.RTingUsers$ = List of users that have their tweet Retweeted by User U.

As for Twitter analysis, Yang et al. [15] demonstrated the application of HITS algorithms in the Twitter domain by constructing two graphs: User Graph and Tweet Graph with RT and Follower-followee relationship. They proposed a modified version of the HITS algorithm with two additional parameters: (i) α - bonus for RT chain and (ii) β - bonus for RT from non-followers. However, in this research we believe that the direction of the edge for the graph should be directed from user that Retweeted (Retweeted tweets) towards the original users (tweets) as shown in Twitter User Rank (TURank) [16].

In this paper, each Twitter user U contains a set of *Hub* score U_h and *Authority* score U_a. Given that, V are the set of users that Retweeted user U's tweet; while W are the set of users that have their tweet Retweeted by user U. The authority score U_a is equal to the total of the Hub score V_h; where the hub score U_h reflects the total of the Authority score W_a. In a graph context, the authority score of Vertex U will impact the hub score of Vertex V if there is an Edge (V,U) which exists in Graph G. In this respect, HITS provides two insights for each user:

- How well this user acts as an information hub (by looking at its hub score)
- How informative is the content of this user (by looking at its authority score)

These two values are mutually reinforcive because a good hub is a page that points to good authorities, while a good authority is a page that pointed by good hubs [13]. Processes involved in HITS are shown in Algorithm 1 which includes: Initialization, Authority Update and Normalization.

B. PageRank Algorithm

PageRank (PR) algorithms was published in 1998 by Larry Page and Sergey Brin to solve web page ranking problems [17] which is mainly used by Google, one of the pioneers of Internet search engines. PR simulates the behaviour of a "random surfer" by performing random walk on the graph. PR is distributed from a node to their outlinks, therefore there is an issue when the "surfer" reached the dangling nodes in a graph. Damping factor (d), a constant probability value was introduced to provide the probability of "jump" to another nodes that is outside from its outlinks list.

PageRank can be summarized as a single score per user which is known as *PageRank Score*. Assuming a list of user V that Retweeted user U's tweet, the PageRank Score of each user in V is equally distributed to user U. For example, user V_i has a PageRank Score of *k*, and it has Retweeted a total number of four(4) users' tweet including U. Each of the user that Retweeted by user V_i will be receiving PageRank Score of $\frac{k}{4}$ from V_i. The detailed of this algorithm is shown in Algorithm 2.

With regard to Twitter analysis, Weng et al. [18] proposed an extension of the PageRank algorithm, named (TwitterRank), to measure the influence of users in Twitter by considering the following three attributes: Topic Distillation, Follower-followee relationship and Number of Tweets published. In addition to that, they also defined the influence of Twitter User A based on each of his/her follower as the relative amount of content the follower received from Twitter User A. They assigned different transition probability based on the similarity of topic and the probability of Twitter User B will see the tweet posted by Twitter User A. This is in contrast to TunkRank [19] which only consider the latter attribute.

Algorithm 2 PageRank

```
G := set of users
$d \leftarrow 0.85$                                  ▷ Damping Factor
$N \leftarrow TotalNumbersOfUsers(G)$
procedure INITIALIZATION($G$)
    for each user $U$ in $G$ do
        $U_{pr} \leftarrow \frac{1}{N}$
    end for
end procedure
procedure PAGERANK($G$)
    for step from $1$ to k do
        $dU \leftarrow 0$
        for all user $U$ that does not RT do
            $dU \leftarrow dU + d * \frac{1-d}{N}$
        end for
        for each user $U$ in $G$ do
            $U_{pr} \leftarrow dU + \frac{1-d}{N}$
            for all user $V$ in $U.RTbyUsers$ do
                $U_{pr} \leftarrow U_{pr} + \frac{d*V_{pr}}{NumberOf(V.RTingUsers)}$
            end for
        end for
    end for
end procedure
```

Note:
$U.RTbyUsers$ = List of users that Retweeted User U's tweet.
$V.RTingUsers$ = List of users that have their tweet Retweeted by User V.

III. Experiment

In this paper, we used the Twitter Streaming Application Programming Interface (API) to crawl tweets based on specified keywords on publicly available content which is accessed via the API. The Twitter Streaming API contains an attribute called "retweeted_status" that wraps the information of the original tweets which provides us a certain level of confidence that the RT is genuine. All native RT action (including protected account) contributes to the retweet_count entity in the meta-data of the original tweets; while non-native RT will not. It's noted that the amount of information available from this meta-data is very limited. As a result, only 1282 tweets with this meta-data were captured throughout the period of 21 days of crawling. In general, there are a variety of methods to identify a RT, however, in this paper, we focus on the tweets that have the character "RT" in front.

A. Environment Setup

The following infrastructure to analyse the collected data.

- Hadoop [20]: Four Native Hadoop Clusters that consists of one Master (QUAD-CORE Processor, 2GB RAM) ma-

TABLE I
KEYWORDS USED FOR TWITTER CRAWLING

	Keyword
1	HyppTV
2	Streamyx
3	UMobile
4	Digi
5	Maxis
6	Yes4G
7	Celcom
8	xpaxsays
9	TMCorp
10	TMConnects
11	TeamMsia
12	every1connects
13	tmrewards
14	yellowpages_my
15	tmsmebiz
16	MaxisComms
17	MaxisListens
18	DiGi_Telco
19	DiGi_Youths
20	helloUMobile

TABLE II
TOP 20 RANKED TWITTER USERS FOR HITS AND PAGERANK

PR	**Rank**	**HITS**
TeamMsia	1	**TeamMsia**
ManOlimpik	2	**Khairykj**
Khairykj	3	**ManOlimpik**
OKS_HARIMAUMUDA	4	**OKS_HARIMAUMUDA**
BrooksBeau	5	**FIH_Hockey**
TMCorp	6	BB_Johor
LawakLegend	7	AtletMalaysia
WTFSG	8	**TMCorp**
JanganPanas	9	Faif_D
FIH_Hockey	10	**BBST15**
James_Yammouni	11	**JanganPanas**
MuizzKarting	12	**asianadotmy**
Konami	13	**LawakLegend**
YanaSamsudin	14	**fizoomar**
asianadotmy	15	**YanaSamsudin**
3RMalaysia	16	sportsmalaysia
fizoomar	17	MaxisComms
Stranahan	18	**BajetJer**
BajetJer	19	DiGi_Telco
BBST15	20	**3RMalaysia**

chine and three Slave machines (DUO-CORE Processor, 1 GB RAM)

- Flume [21]: Apache v1.4.0 , using Twitter Streaming API.
- Hive [22]: External Table in Hive and export data processing results to a local machine.

B. Dataset

There is a total of 20 keywords related to telecommunication providers were selected to be configured in our Apache Flume module for Twitter crawling as shown in Table I. These keywords consist of the subset of products and official customer support & communication account used by major telecommunication companies in Malaysia. For the purpose of this paper, we focused on a subset of the collected data that consists of RTs. There are many variations of the RT pattern, for instance:

- *RT “@<user_name>: <original message>”*
- *RT via @<user_name>: <original message>*
- *<original message> via @<user_name>*

It’s noted that there are some users that append additional content into their RTs which make the filtration process even more challenging, e.g.: “*RT <comments> @<user_name>: <original message>*” Hence, it is extremely difficult to capture every single pattern since there is no proper control over how the users compose a tweet. In this paper, the main pattern for recognizing a tweet as a RT is defined as “*RT @<user_name>: <original message>*”.

The dataset that was collected over the period of 21 days of crawling has the following statistics:

- # Total Tweets: 230,116
- # Total Unique Users: 121,461 (screen_name)
- # Total Verified Users: 113 (verified)
- # Average Followers Count : 983 (followers_count)

After the tweets were extracted, they are categorized as RT, 56,727 number of tweets were obtained. Among these tweets, it is identified 50,636 unique users (nodes) and constructed a graph with 50,794 RT (edges). An edge from user A to user B is constructed if user A RT user B’s tweet.

IV. RESULTS AND DISCUSSION

Table II shows the list of users screen name that are ranked in the TOP 20 solely based on the their RT relationships for both HITS and PR algorithms. As shown in Table II, it is observed that both the HITS and PR algorithms show a similar subset of users in their respective TOP 20 range with the percentage of 70%. Although only 10% (2 out of 20) of exact matches ranking for both algorithms in their respective TOP 20, out of 37,730 users exact matches from the overall 50,636 unique users (approximately 74.5%).

There is a total of 26 unique users identified in the TOP 20 generated by both the HITS and PR algorithms. The first column of Table III represents the user (User), followed by verified accounts (V), followers count number (FC), PR ranking sequence (PR), HITS ranking sequence (HITS) and the difference of ranking sequence between both ranking algorithms (Diff) respectively. From the experiment results, 23.07% of them are verified accounts and 88.5% of them surpassed the average followers count of 983.

Figure 1 visualizes the overall RT relationship network. Based on the visualization, there are several large communities being identified and it is noted that these communities were highly clustered based on topics and locations. Figure

TABLE III
COMPARISON OF RANKING RESULT FOR HITS AND PAGERANK

User	V	FC	PR	HITS	Diff
TeamMsia	F	98469	1	1	0
ManOlimpik	F	1661	2	3	1
Khairykj	T	432259	3	2	1
OKS_HARIMAUMUDA	F	35058	4	4	0
BrooksBeau	T	1226629	5	704	699
TMCorp	F	13767	6	8	2
LawakLegend	F	49593	7	13	6
WTFSG	F	469909	8	705	697
JanganPanas	F	14642	9	11	2
FIH_Hockey	F	24628	10	5	5
James_Yammouni	T	817592	11	706	695
MuizzKarting	F	104	12	39	27
Konami	T	323840	13	707	694
YanaSamsudin	F	494408	14	15	1
asianadotmy	F	137	15	12	3
3RMalaysia	F	3198	16	20	4
fizoomar	F	492952	17	14	3
Stranahan	F	31322	18	708	690
BajetJer	F	13422	19	18	1
BBST15	F	13149	20	10	10
BB_Johor	F	26704	21	6	15
sportsmalaysia	F	4716	25	16	9
AtletMalaysia	F	4291	35	7	28
Faif_D	F	506	36	9	27
MaxisComms	T	116604	76	17	59
DiGi_Telco	T	48204	94	19	75

2 shows the most popular account in the entire network - "TeamMsia". The account accumulated up to 5,032 RTs in this period of time. One of their famous RT reads : "*Tahniah pasukan Hoki Remaja. We are all very proud of you. RT jika anda bangga dengan mereka! #TeamMsiaBoleh #HJWC pic.twitter.com/7f9t6TfF1k*" [1].

V. CONCLUSION AND FUTURE WORKS

In this paper, both Link-Based Ranking (LBR) algorithms, namely HITS and PageRank were applied to Twitter users analysis. As this is an initial research work to analyze tweets on a higher level, i.e. Twitter Users level, we found that the direction of the edges is extremely important. From experiment results obtained, we can conclude that there are differences of the ranking sequence by both algorithms which are mainly due to the consideration of both inbound and outbound links in HITS, while PR only considers the inbound link.

As for future work, we will expand the our study by including additional attributes such as Follower-Followee Relationship and Reply-Mention Relationship. Besides that, we will explore the idea of ranking the Twitter user level cascading down to tweets level and detect interesting tweets as well as spams.

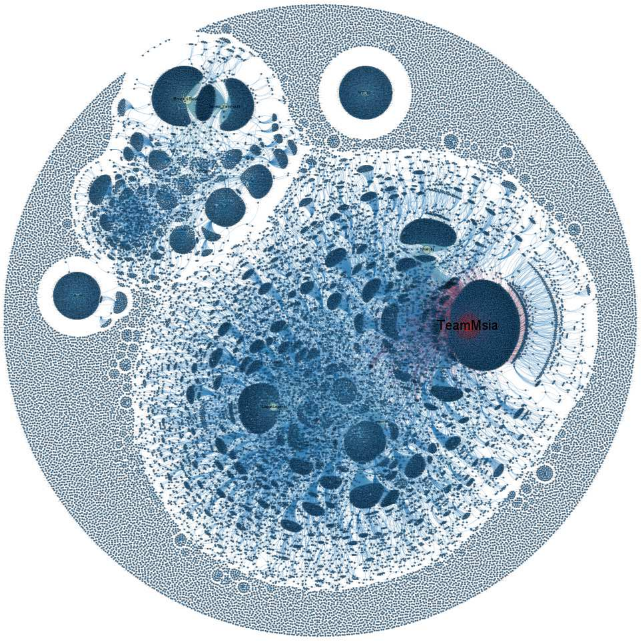


Fig. 1. *Visualization of the overall Retweet (RT) network.*

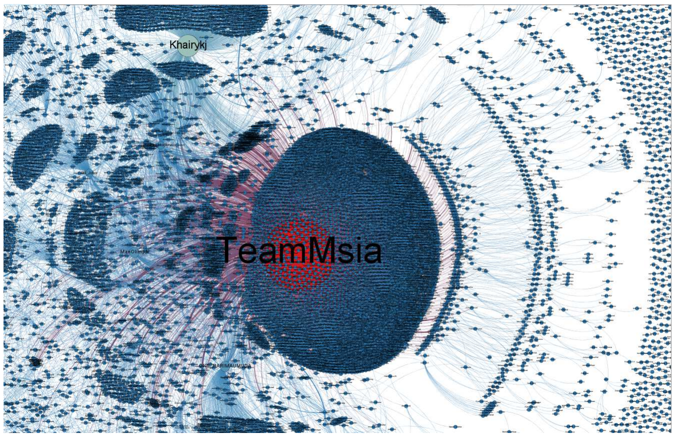


Fig. 2. *Visualization of the largest cluster in the Retweet (RT) network.*

VI. ACKNOWLEDGEMENTS

This work is partially supported by the Telekom Research & Development (TM R&D) 2014 grant.

REFERENCES

[1] Twitter IPO Filling Security Exchange Commission (SEC) http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm#toc564001_4

[2] Twitter - Using Hashtags on Twitter https://support.twitter.com/groups/50-welcome-to-twitter/topics/204-the-basics/articles/49309-using-hashtags-on-twitter

[3] Twitter - Receiving recommendations from Twitter https://support.twitter.com/groups/53-discover/topics/217-tweets/articles/20170749-receiving-recommendations-from-twitter

[4] Twitter Blog - Stay in the know https://blog.twitter.com/2013/stay-in-the-know

[1] https://twitter.com/TeamMsia/status/411470672129314816

[5] Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitters streaming API with Twitters firehose. Proceedings of ICWSM.

[6] Twitter - FAQs about following
https://support.twitter.com/groups/52-connect/topics/213-following/articles/14019-faqs-about-following

[7] Twitter - Posting a Tweet
https://support.twitter.com/groups/52-connect/topics/211-tweeting/articles/15367-posting-a-tweet

[8] Twitter - FAQs about Retweets (RT)
https://support.twitter.com/articles/77606-faqs-about-retweets-rt

[9] Twitter - Project Retweet: Phase One
https://blog.twitter.com/2009/project-retweet-phase-one

[10] Early developer preview: Retweeting API
https://groups.google.com/forum/#!topic/twitter-api-announce/HgfjMuw9RJ0

[11] Twitter - Retweet Count Incorrect for a Retweeted Tweet
https://dev.twitter.com/issues/203

[12] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46(5), 604-632.

[13] Ding, C., He, X., Husbands, P., Zha, H., & Simon, H. D. (2002, August). PageRank, HITS and a unified framework for link analysis. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 353-354). ACM.

[14] Bosma, M., Meij, E., & Weerkamp, W. (2012). A framework for unsupervised spam detection in social networking sites. In Advances in Information Retrieval (pp. 364-375). Springer Berlin Heidelberg.

[15] Yang, M. C., Lee, J. T., Lee, S. W., & Rim, H. C. (2012, August). Finding interesting posts in twitter based on retweet graph analysis. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (pp. 1073-1074). ACM.

[16] Yamaguchi, Y., Takahashi, T., Amagasa, T., & Kitagawa, H. (2010). Turank: Twitter user ranking based on user-tweet graph analysis. In Web Information Systems EngineeringWISE 2010 (pp. 240-253). Springer Berlin Heidelberg.

[17] Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer networks and ISDN systems, 30(1), 107-117.

[18] Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010, February). Twitterrank: finding topic-sensitive influential twitterers. In Proceedings of the third ACM international conference on Web search and data mining (pp. 261-270). ACM.

[19] TunkRank - A Twitter Analog to PageRank
http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank/

[20] Apache Hadoop
http://hadoop.apache.org/

[21] Apache Flume
http://flume.apache.org/

[22] Apache Hive
http://hive.apache.org/