International Journal of Networked and Distributed Computing, Vol. 1, No. 3 (August 2013), 167-173

Efficient Keyword-Related Data Collection in a Social Network with Weighted Seed Selection

Changhyun Byun*
Department of Computer and Information, Towson University, 7800 York Rd, Towson, MD 21093, U.S.A
E-mail: cbyun1@students.towson.edu
www.towson.edu

Hyeoncheol Lee
Department of Computer and Information, Towson University, 7800 York Rd, Towson, MD 21093, U.S.A
E-mail: hlee23@students.towson.edu
www.towson.edu

Jongsung You
Department of Computer and Information, Towson University, 7800 York Rd, Towson, MD 21093, U.S.A
E-mail: jyou1@students.towson.edu
www.towson.edu

Yanggon Kim
Department of Computer and Information, Towson University, 7800 York Rd, Towson, MD 21093, U.S.A
E-mail: ykim@towson.edu
www.towson.edu

Received 18 April 2013

Accepted 26 June 2013

Data mining in a social can yield interesting perspectives to understanding human behavior or detecting topics or communities. However, it is difficult to gather the data related to a specific topic due to the main characteristics of social media data: large, noisy, and dynamic. To collect the data related to a specific topic efficiently, we propose a new algorithm that selects better seeds with limited resources. Furthermore, we compare two data sets collected by the algorithm and existing approaches.

Keywords: social networks; twitter; seed analysis; initial nodes; crawling; presidential election 2012;

* 7800 York Rd, Towson, MD 21093, U.S.A

Copyright: the authors

1. Introduction

In recent years, online social network sites, such as Facebook, Twitter, Blogger, LinkedIn, and MySpace, have changed the way people communicate each other. People share information, report news, express opinions and update their real-time status on the online social network sites. With the increasing popularity of the online social network sites, a huge amount of data is being generated from them in real time. Analyzing the data in social media can yield interesting perspectives to understanding individual and human behavior, detecting hot topics, and identifying influential people, or discovering a group or community.[1,2]

Twitter is an online social network site based on text message of up to 140 characters, which generates 340 million tweets and handles 1.6 billion search queries per day as of 2012. It also provides Application Programming Interface (API) to allow researchers and data analyzers to access a variety of data in Twitter. Numerous researchers have paid attention on gathering and analyzing the data to detect issues, such as detecting earthquakes[3] and influenza using Twitter or recommending tags to users.[4]

However, it is impossible to gather plenty of data without automated data processing. For that reason, many researchers have developed their own data collecting tools from diverse of social media sites.[5,6,7] We also proposed a data collecting tool which enables data seekers to gather data from Twitter for a specific topic.[8] Nevertheless, it has been an issue to gather the data related to a specific topic that data seekers are interested in due to the main characteristics of social media data sets: data is large, noisy, and dynamic.

Among the many considerations to gather the data related to a specific topic, there is no doubt that selecting seed nodes used for staring point of data gathering process is the most important step to gather more relative data for a specific topic. In the previous research, we used manual seed selection process by experts,[9] which has potential for selecting seed nodes that have relatively low influence on a topic in a social network. Thus, we propose an algorithm to find suitable seed nodes, which can maximize the efficiency of data gathering process to collect more topic-related data from Twitter. The algorithm considers user influences and activities to find the best initial seed nodes dynamically with limited resources and time.

The remainder of this paper is constructed as follows: In Section 2, the related works that have done so far are summarized. Section 3 introduces the design specifications and explains details of the algorithm. Section 4 presents the results of data gathering and compares two data sets collected by the algorithm and an existing method respectively. The last part, Section 5 concludes the work by summarizing this paper and the future research direction.

2. Related Research

2.1. *Automated Twitter Data Collecting Tool*

In the previous research, we presented a java-based data gathering tool, which is characterized by following features (see Ref. 9). Firstly, it continuously and automatically collects data from Twitter. Secondly, it allows us to start the data collecting process from multiple seed nodes. Thirdly, it handles a multitude of authorized keys to increase the total number of Twitter API calls. Fourthly, it stores collected data into database for analysis. Finally, it supports intuitive user interface to interact with users. Although it is able to gather a huge amount of data from Twitter for a specific topic, initial seed-nodes are selected by users manually. If initial seed nodes that are not related to the topic are selected, lots of noisy data will be gathered from the social network, which makes data-analysis difficult. For that reason, we need initial seed node selection algorithm based on user's influence to gather data related to a specific topic.

2.2. *Measuring user influence in Twitter*

Cha et al. measured the influence of users in Twitter using three interpersonal activities: in-degree, retweets, and mentions.[10] In-degree influence is the number of followers of a user, which indicates the size of audience for that user. Retweet influence is the number of retweets contacting one's name, which indicates the ability of that user to generate content with pass-along value. Mention influence is the number of mentions containing one's name, which indicates the ability of that user engage others in a conversation. While they showed that user influence can be measured by three perspectives, they analyzed the influence based on the

Copyright: the authors

data that has been gathered, which needs a lot of pre-processing time. Also, they did not consider other factors that are related to user's influence in Twitter.

2.3. *Influence maximization in a social network*

Influence maximization is to find individuals who have the most influence in a social network.[11] Researches have paid a lot of attention to this problem. Domingos and Richardson firstly studied influence maximization problem and proposed a probabilistic solution.[12,13] Kempe et al. presented diffusion models to solve the problem and proved the influence maximization problem to be NP-hard.[14] They also proposed a greedy approximation algorithm that guarantees performance of the optimal influence, which outperforms the existing node-selection approaches. Chen et al. proposed a new greedy algorithm and heuristics to maximize influence in a social network.[15,16] However, existing influence maximization approaches need to be improved to reflect the characteristics of a social network and find initial seed nodes for a specific topic with limited time and resources.

3. Dynamic Seed Analysis in a Social Network

In this section, we present the main idea of this paper to maximize data gathering results using seed analysis. This section consists of two subsections. Section 3.1 shows architecture of the Twitter Data gathering tool with the Seed Analysis module, and Section 3.2. describes how the algorithm quantifies activities of a node.

3.1. *Architecture of Twitter data gathering tool with seed analysis module*

In this paper, we extended the existing architecture of the Twitter data collecting tool. The given data collecting tool consists of the Account Handler, the Data Gathering Controller, the Database Handler, Data Gatherer Thread Pool, as well as the Data Filtering Handler. Fig. 1 shows you the constituent modules of the Twitter data collecting tool including the Seed Hander.

The novel module, the Seed Hander, is in charge of selecting an initial node and calculating each node's activities based on frequency of use of a certain keyword in its most recent postings. The Seed Handler module is running on two main algorithms that are the Initial Node Selecting Algorithm, and the Node's Activity Calculating Algorithm to satisfy its purpose of existence.

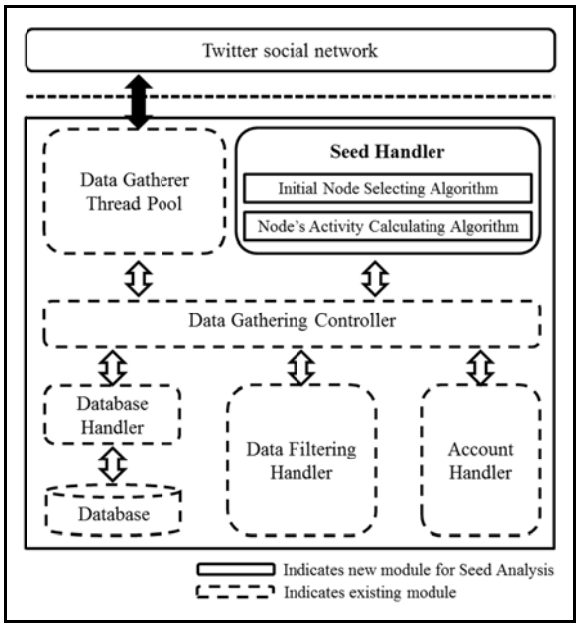

Fig. 1. Architecture of the Twitter data gathering tool with seed analysis module.

3.2. *Algorithm for selecting influential nodes*

For the most part, recent critical debates about social analysis have tended to center around the question of finding the most influential node in social network, also known as finding seed nodes. Social influence analysis aims either to verify the existence of social influence or to quantify the strength of the influence.[17] Dynamic Social Network Analysis is about to model how friendships drift over time using a dynamic model or to investigate how different pre-processing decisions and different network forces such as selection and influence affect the modeling of dynamic network.

These two related research areas in social network encouraged us to build an algorithm finding seed nodes by calculating node's activity in dynamic social network. Our approach of gathering keyword-related data more efficiently is collecting data from only qualified nodes. This goal can be achieved by giving activity weight to each node and checking if the node has enough activity weight before collecting tweets from the node.

Fig. 2 shows the data gathering process from selecting an initial node through storing tweets into database. Once the tool is initiated, a list of candidates for an initial node is organized based on their most

Copyright: the authors

recent tweet, which contains a certain keyword at least once. When the list of candidates is built, an algorithm calculates activity weight of each candidate, and the list is organized by the number of followers of each candidate node. The first node in the list, which has highest the number of followers and is a qualified node, will be the initial node. If the first node in the list is not the qualified node, then check the next available node to see whether it is qualified. This process of finding a qualified node is iterated until an initial node is selected. Rebuilding a new list of candidates is required in case of the absence of a qualified node in the list. Once an initial node is selected, the tool generate a list of the initial node's followers information, such as each follower's unique id, language, number of followers, number of friends, etc. Then, each follower's activity weight is calculated, and only tweets from qualified users are collected, until there is no more follower information in the list.

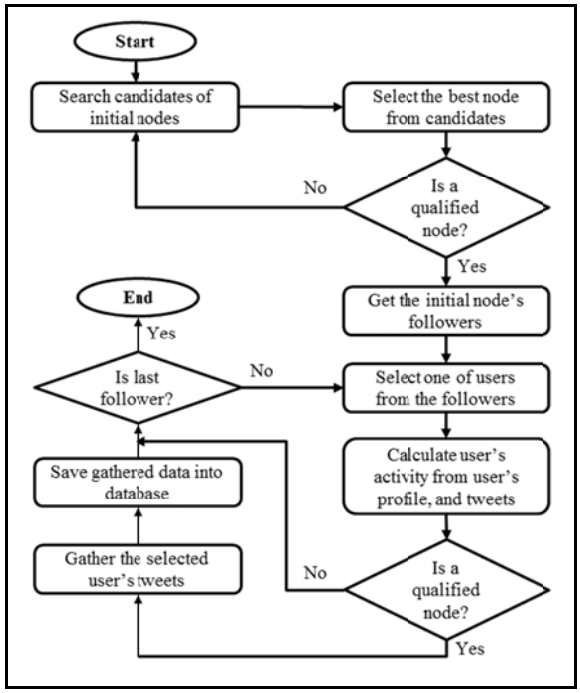

Fig. 2. Flow chart of data gathering process by qualified nodes.

Fig. 3 shows the simple algorithm to find a node, which has more followers than others. If the selected node is not a qualified node, the tool removes the node from the list and runs the algorithm to find a suitable node again.

Notation: *A* is a set of nodes such that recently posted their tweets about a certain keyword. *B* is a pointer indicating the best initial nodes in the set *A*.

0: Set *B*=*A*[0].
1: **Loop** *I* from 1 to *A*.size – 1.
If *B*.no_of_followers < *A*[*I*].no_of_followers, **then**
3: Set *B*= *A*[*I*].
4: **End If.**
5: **End Loop.**
6: **End of Algorithm.**

Fig. 3. Algorithm of selecting initial node from a list of candidates.

In a user's profile, there are properties to be considered as factors of the user's activities in Twitter, such as number of followers, number of friends, number of keyword-related tweets, date tweeted, and favorite count. Among the user's properties, we use the number of followers, the number of keyword-related tweets, and date tweeted as main factors for calculating user's activity. Fig. 4 shows the algorithm of calculating user's activity based on existence of a target topic in its tweet texts within a 30 day time period.

Notation: *T* is a set of tweets such that recently posted by a user within 30 days from search date and time. *K* is a string variable containing a keyword. *W* is a float variable containing user's activity value calculated by this algorithm. *M* indicates the number of tweets and *N* implies the number of tweets containing the keyword *W*.

0: Set *M*=*T*.size.
1: Set *N*=0.
2: **Loop** *I* from 0 to *M*
3: **If** the tweet ,T[*I*], contains the keyword K, **then**
4: Set *N*=*N*+1.
5: **End If.**
6: **End Loop.**
7: **If** *M* is not 0, **then.**
8: *Set W=N/M.*
9: **End If.**
10: **End of Algorithm.**

Fig. 4. Algorithm of calculating user's activity weight.

4. Experiments

In this section, we present an evaluation of the performance of building the Twitter dataset with the Twitter data gathering tool with our influential node selecting algorithm on the real Twitter network. The following sub sections show an experimental test bed and the result of test.

4.1. *An experimental test bed*

- **Keywords:** The presidential election of the United States of America was held at the end of 2012. There is no doubt that the presidential election is the most popular event in the U.S.A. The name of one of the candidates, Obama, has been chosen as a keyword to build a new dataset for a popular event to analyze user's behavior in Twitter about elections in the future.
- **Dataset:** Dataset will be built from real Twitter network in real-time. Two different types of data gathering approaches will be used. One approach is using the seed analysis algorithm we proposed in this paper, and another method is to start gathering process from an initial node manually selected by a data analysis specialist. Particularly, the specialist chooses possible seed candidates from Twitter accounts, such as BarackObama, MittRomney, VP, TheDemocrats, and etc. From the candidate list, one of Twitter account is selected for each attempt. For instance, Obama's Twitter account is picked as an initial node for first attempt because Twitter users following Obama tend to post tweets about Obama more than otherwise. Then, Target users are arbitrarily selected from all of Obama followers for each attempt to generalize result of data collection.

4.2. *Experimental Results*

Data gathering results by two different types of data gathering approaches are illustrated in Table 1 and Table 2. Each data gathering approach attempts five-times to see if the algorithm performs evenly with different sets of candidates of initial node. As shown in two table records, the average portion of keyword-related tweets in the dataset built by our approach is much larger than another approach (9.88% keyword-related tweets in our approach compared to average of only 1.27% keyword-related tweets in the manual pick approach). In other words, this result means that data collection from qualified seed node and follower nodes collects more keyword-related tweets than otherwise.

Table 1. Data gathering results from seed analysis algorithm

Attempts	Total number of tweets	The number of keyword-related tweets (%)	The number of no keyword-related tweets (%)
1	14,149	1,906 (13.47%)	12,243 (86.53%)
2	12,045	758 (6.29%)	11,287 (93.71%)
3	11,615	1,109 (9.55%)	10,506 (90.45%)
4	10,779	1,152 (10.68%)	9,627 (89.32%)
5	10,115	875 (8.66%)	9,240 (81.46%)
Average	11,741	1,160 (9.88%)	10,581 (90.12%)

Table 2. Data gathering results from selecting an initial node by manual pick

Attempts	Total number of tweets	The number of keyword-related tweets (%)	The number of no keyword-related tweets (%)
1	11,847	13 (0.11%)	11,834 (99.89%)
2	11,666	17 (0.15%)	11,649 (99.85%)
3	14,082	27 (0.19%)	14,055 (99.81%)
4	13,087	409 (3.13%)	12,678 (96.87%)
5	10,549	309 (2.93%)	10,240 (97.07%)
Average	12,246	155 (1.27%)	12,091 (98.73%)

To perceive the statistical significance of differences between data gathering results from two different algorithms, we applied the Chi-square test to the data gathered by two different data gathering approaches and derived the results, as shown in the Table 3 and Fig. 5. The result of the Chi-square test shows that there is less than 0.001% chance that this deviation is due to chance alone. This implies that the algorithm we developed gathers significantly more data than gathering data from an initial seed selected by manual pick.

Table 3. Result of Chi-square test to data gathered by two different data gathering approaches

Approach	Keyword-related		Not-keyword-related	
	Freq.	*%*	*Freq.*	*%*
Seed Analysis	1,160	88.2%	10,581	46.7%
Selected by Specialist	155	11.8%	12,091	53.3%
Total	1,315	100%	22,672	100%

[a.] $\chi^2 = 858.40$; Degree of freedom = 1; Probability < 0.001

Copyright: the authors

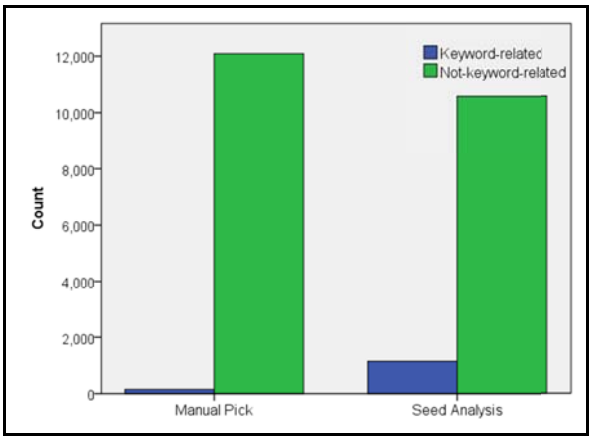

Fig. 5 Graph visualization of the result of Chi-square test

Organizing user nodes by each node's activity weight allows us to discover the number of *k* seed nodes. Table 4 illustrates top 5 seed nodes from data gathering result using the seed analysis approach. To protect user's privacy, first four characters of each user's screen name are shown. During data collecting process, the ordered list of the nodes can be dynamically changed due to the dynamic seed selecting algorithm.

Table 4. Seed nodes discovered by our algorithm on dynamic twitter network

Number	Screen Name	Follower Count	Friend Count	Activity Weight
1	Noma*******	13604	6345	0.99
2	Kath*******	2937	2972	0.36
3	Rock*******	7638	5236	0.36
4	Want********	13475	13364	0.33
5	Boud*******	1051	1175	0.32

5. Conclusions

In this paper, we proposed a new algorithm to find the best initial seed nodes with limited time and resources to gather the data that is related to a specific topic or keyword that data seekers are interested in. The algorithm evaluates user's activities and updates the seed node list dynamically.

After the gathering process, we compared two results, one from this algorithm and one from manual pick by human. The result proved that the efficiency of the algorithm for collecting more keyword-related data is higher than the existing approach.

The algorithm presented in this paper supports only one keyword. In future work, the algorithm need to be improved to find an initial node based on multiple keywords. Also, this algorithm can be enhanced to analyze user influence in a social network that extends dynamically.

References

1. J. C. Cortizo, F. M. Carrero, J. M. Gomez, B. Monsalve, and P. Puertas, *Introduction to mining social media*, In F. M. Carrero, J. M. Gomez, B. Monsalve, P. Puertas, and J. C. a. Cortizo, editors, Proceedings of the 1st International Workshop on Mining Social Media, pages 1–3, 2009.
2. I. King, J. Li, and K. T. Chan, *A brief survey of computational approaches in social computing*, IJCNN'09: Proceedings of the 2009 international joint conference on Neural Networks, pages 2699–2706, Piscataway, NJ, USA, 2009. IEEE Press.
3. T. Sakaki, M. Okazaki, and Y. Matsuo, *Earthquake shakes Twitter users: real-time event detection by social sensors*, Proceedings of the 19th international conference on World Wide Web (WWW '10), Raleigh, North Carolina., 2010, pp.851-560.
4. E. Aramaki, S. Maskawa, and M. Morita. *Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter*, Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK.,2011, pp.1568-1576.
5. M. Bošnjak, E. Oliveira, J. Martins, E. Mendes, L. Sarmento. *TwitterEcho - A Distributed Focused Crawler to Support Open Research with Twitter Data*, WWW 2012 – MSND'12 Workshop, April 16–20, 2012, Lyon, France.
6. H. Kwak, C. Lee, H. Park, and S. Moon. *What is Twitter, A Social Network or A News Media?*, Proceedings of the 19th International Conference on World Wide Web (WWW), pages 591-600, 2010.
7. P. Noordhuis, M. Heijkoop, and A. Lazovik. *Mining Twitter in the Cloud*, IEEE 3rd International Conference on Cloud Computing, 2010.
8. C. Byun, H. Lee, and Y. Kim. *Automated Twitter Data Collecting Tool for Data Mining in Social Network*, Proceedings of the 2012 ACM Research in Applied Computation Symposium, pages 76-79. 2012.
9. C. Byun, H. Lee, Y. Kim, K. K. Kim, *Automated Twitter Data Collecting Tool and Case Study with Rule-based Analysis*, 14th International Conference on Information integration and Web-based Application & Services, Bali, Indonesia, IIWAS, 2012, pp. 196-204.
10. M. Cha, H. Haddadi, F. Benevenuto, K. P. Gummadi, *Measuring User Influence in Twitter: The Million Follower Fallacy*, 4th International AAAI Conference on Weblogs and Social, ICWSM, 2010, pp. 10-17.
11. X. Shang, X. Chen, Z.Jiang, Q.Gu, D.Chen, *Factor Analysis for Maximization Problem in Social Networks*, 13th ACIS International Conference on Software

Copyright: the authors

Engineering, Artificial Intelligence, Networking and Parallel & Distributed Computing, Nanjing, China, SNPD, 2012, pp.95-101.

12. P. Domingos and M. Richardson, *Mining the network value of customers*, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), San Francisco, CA, 2001, pp. 57–66.
13. M. Richardson and P. Domingos, *Mining knowledge-sharing sites for viral marketing*, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), Edmonton, Alberta, Canada, 2002, pp. 61–70.
14. D. Kempe, J. Kleinberg, and E. Tardos, *Maximizing thespread of influence through a social network*, Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), Washington,D.C., 2003, pp. 137–146.
15. W. Chen, Y. Wang, S. Yang, *Efficient influence maximization in social networks*, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), Paris, France. 2009, pp.199-208.
16. W. Chen, Y. Wang, S. Yang, *Scalable influence maximization for prevalent viral marketing in large-scale social networks*, Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), Washington,D.C., 2010, pp.1029-1038.
17. M. Richardson, P.Domingos, *Mining Knowledge-Sharing Sites for Viral Marketing*, Eighth International Conference on Knowledge Discovery and Data Mining 02, 2002.

Copyright: the authors