Identification Of Seed Users Via Short Messages Based On Hadoop

Gan Shicheng
School of Software Engineering
South China University of Technology
Guangzhou, China
ganecheng@126.com

Zhang Pingjian
School of Software Engineering
South China University of Technology
Guangzhou, China
pjzhang@126.com
* Corresponding author

Abstract—Short messages are an important part of social media network. Much work has been devoted to mining useful information and knowledge from short messages via classification and clustering approaches. However, in many situations, it is vital to identify some influential users whose messages will result in a large amount of forwarding, these users are called seed users. In this paper, directed graph has been setup to modelling short message forwarding, then, a "Reverse PageRank" scoring policy is presented to evaluate users. Finally, the Hadoop platform is introduced to facilitate processing of seed user identification. Experimental results show that the model and the scoring policy are effective in finding out seed users.

Keywords-Seed user; short message forwarding; scoring policy; Hadoop; PageRank

I. INTRODUCTION

With the popularity of mobile phones, more and more persons use their mobile phones in daily life. One of the main functionalities of mobile phones is to send and receive messages, which produces a very large volume of text messages every day, with topics covering private communications, information announcement from governmental department, education, entertainment and even business. Now, telecom operators, scientists and entrepreneurs as well have come to realize the huge potential value hidden in the big data of short messages. There is already much work on mining the information and knowledge from short messages and other social media network such as Twitter, Facebook in the literature. Clustering approaches, for example, have been utilized to extract the hierarchy that exists in groups [1],[2] and to determine the relationship among users [3],[4]. Short text classification has been tackled by various authors using SVM [5] and coherence constraints [6] to resolve the issue. In [7], a probabilistic knowledge base is employed to conceptualize short text with applications to clustering Twitter messages. Much attention has also been paid to the mining of opinions [8],[9] and sentiments from short messages [10],[11],[12].

Another interesting topic is to identify influential users. Among the short messages, some are original and some are just forwarded. Especially, some particular users' short messages could result in a large amount of forwarding by the receivers. This is similar to the hub pages in WWW where some pages are referred to by many other pages. Such influential users are called seed users, whose messages that result in a large amount of forwarding are called seed messages. Identifying seed users will create great value in business, culture and social science since these users own powerful propagation ability. Nevertheless, there is surprisingly little research work on this direction, comparing to the extensive studies on identifying hub pages. One possible reason might be that, due to privacy policy, the short messages records accessed by researchers usually do not contain the message body text, which is a big obstacle to deep analysis. Among the few related work, a tree-like model is set forth to find seed users [13]. However, their approach seems problematic with regard to the evaluation function. According to equation (1) and equation (2) in [13], the scoring function is monotonic along propagation paths, i.e., if a user is a seed user, then, all his ancestors are seed users also. This might not be true, since a non-seed user could just happen to send a seed message to a seed user.

In this paper, the directed graph is used to model the propagation of short messages, with nodes being users that send/forward seed messages. Nodes that belong to the same user are merged to represent a candidate seed user. To evaluate those candidates, a "Reverse PageRank" scoring function is established. Finally, these approaches are implemented in a Hadoop platform and applied to a real short message record dataset from some telecom operator to verify the effectiveness of the directed graph model and the "Reverse PageRank" scoring function.

II. SETUP OF THE FORWARDING MODEL

A. Terms

For the sake of convenience in the following descriptions, some definitions are introduced first.

Definition 2.1 A group message is the message that is sent by some user to many different receivers within a short period.

Definition 2.2 A seed message is a group message that involves large amount of forwarding among mobile phone users within a short period.

The term "seed message" is a little vague by now, since the precise meaning of "large amount" and "short period" is not given, this will be made clear in the model.

Definition 2.3 A candidate seed user is the one that sends or forwards some seed messages. A seed user is a candidate seed user that has powerful propagation capacity.

The candidate seed users are chosen from the ones that are involved in seed messages, thus, non-active mobile phone users are easily filtered out. This greatly simplifies the modelling and identification issues. Again, the exact meaning of "powerful propagation capacity" will be made clear in the scoring policy.

B. The forwarding model

A seed message has the feature that it often starts with a group message. Consider the situation that user A sent a short message to n different users in a short period $[s,t]$. Such a message is assumed to be a group message if the number n is greater than or equal to n_0, some prescribed positive integer, and the time lag between any two consecutive sending times is no more than t_0, some prescribed positive number. The two parameters n_0 and t_0 should be determined by the nature of dataset, as shown later in the experiment section. The group message can be further abstracted as a node $V(A,s)$, which stands that user A sends or forwards a group message at time s.

Before giving procedures to generate the forwarding model from some SMS dataset, another parameter called forwarding window needs to be introduced for the model. Since short messages have strong time effectiveness, only if a receiver sents/forwards a message within the δt period of his receiving time can this receiver be considered to be a child node and δt is defined to be the forwarding window. Otherwise, this will be treated as a different thread.

Algorithm 2.1 The procedure to find out all nodes.

Input: D, an SMS dataset; n_0, minimum number of receivers; t_0, maximum time lag allowed among consecutive sending time.

Output: V, the set of all notes.

Procedure: the classical density-based clustering algorithm OPTICS.

After finding all nodes, put them in a queue Q in ascending order of start time.

Algorithm 2.2 The procedure to generate a component graph with root node N.

Input: Q, the queue that contains all nodes in ascending order of start time; N, the node to start to generate a component graph; δt, the forwarding window.

Output: G, the generated component graph.

Procedure:

1. for each child C of N
2. if child C is a node and its start time is within the forwarding window δt, then
3. add C to the node set of G and add a edge from N to C.
4. recursive on node C.
5. remove node C from Q.
6. end if
7. end for

Algorithm 2.3 The procedure to generate the forwarding model.

Input: Q, the queue that contains all nodes in ascending order of start time; δt, the forwarding window.

Output: M, the forwarding model.

Procedure:

1. while Q is not empty
2. fetch the head node of Q, apply Algorithm 2 on it and add the generated component graph to M.
3. end while

The scoring model

Let $C(u)$ be the set of children of user u, then, $S(u)$, the score of user u is calculated by

$$S(u) = \sum_{v \in C(u)} \alpha \frac{S(v)}{n_v} \qquad (1)$$

where n_v is the number of parents of v, α is some factor, representing the decaying contribution of a child node to its parents.

Equation systems (1) can be put into a more compact form. Let A be the weighted transition matrix of a component graph with node set $V = \{V_i, i = 1,2,...,n\}$ in the model. If V_i is one of the n_j parent nodes of V_j, then, $A(i,j) = 1/N_j$. Denote $S = (S(V_1),...,S(V_n))'$, then,

$$S = \alpha A S \qquad (2)$$

with α being the weight. Although Eq. (2) is quite similar to the PageRank formula in appearance, they are rather different in meaning. The PageRank formula distribute the score of parent page to children pages, while in Eq. (2), parent node absorbs scores of children nodes in the reverse direction, hence termed as "Reverse PageRank" policy. Again, to avoid the "Rank Sink", the so-called "teleporting" technique will also be adopted, and some reserving score is kept for each node. This introduces an additional term in the right hand side of Eq. (2), which then becomes

$$S = \alpha AS + \beta e \tag{3}$$

where e is the vector whose elements are all 1 and β is some small number.

III. EXPERIMENTS USING HADOOP

In this section, a real example of short message records from some tele-communication operator is studied. Before processing, the data should first be cleaned, then, some explorative data analysis is carried out to gain some understandings of the dataset. This helps greatly to determine parameters of the model. Finally, algorithms and scoring policy developed in previous sections are applied to the experimental dataset to find out seed users.

A. Preprocessing the dataset using Hadoop

The short message records comes from China Mobile Communication Corporation, the dataset contains the whole records of one month in a city, with total number of records being more than 20 millions.

To facilitate the subsequent processions, the following manipulations are performed. (i) The "Date" and the "Time" are combined to the "Sending time" so that time lags are computed more easily. (ii) Both "Sender's operator code" and "Receiver's operator code" are removed from the records because these information is redundant for the purpose of this study. (iii) All records are sorted first by Sender's number and then by Sending time in ascending order. This is accomplished using Hadoop, a Java implementation of the map-reduce mechanism.

Algorithm 4.1 The secondary sort procedure via map-reduce.

Input: D, an SMS dataset.

Output: D', the dataset sorted by "Sender's number" and then by "Sending time".

Procedure:

1. During the Mapper stage:
2. Records are split into small pieces and mapped to the desired format by the Mapper job.
3. The outputs of map() are partitioned, each partition will be handled by a Reducer job.
4. Records within a partition is sorted by Sender's number.
5. Records are grouped by Sender's number after sorting.
6. In the Reducer stage, each group will be handled by a reduce() to finish the secondary sort.

Data preprocessing will greatly improve the efficiency of Algorithm 2.1 and Algorithm 2.2.

Statistics of refined dataset

The first interesting question is how many nodes are there in the dataset? By setting $n_0 = 5, t_0 = 60$, the total number of nodes is 79929 with a total number of 1342007 receivers. Thus, the average number of receivers per node is about 16.8. To allow the nodes with less receivers, n_0 could be modified to be 10.

The average time lag for the whole month is 6.99s. Hence, it is reasonable to set $t_0 = 10$. Next is to estimate the size of forwarding window. It can be observed that most nodes begin forwarding messages within a window of 30s. These accounts for 97.4% of all nodes. Thus, the size of forwarding window is taken to be 36.

B. Determining the seed users

So far, by some simple statistical analysi of the dataset, parameters of the forwarding model are determined as follows: $n_0 = 10, t_0 = 10$, and $\delta t = 36$. Using this set of parameters, Algorithm 2.1 - 2.3 generates 34011 nodes, which is much less than the original number. Furthermore, the average number of receivers per node is 23, and the average time lag of consecutive forwarding is 3.6s. Now, applying the scoring equation systems (3) with $\alpha = 0.5$ and $\beta = 0.2$, scores for all users are computed. The users of top 30 highest score are listed in the following table.

TABLE I. THE USERS OF TOP 30 HIGHEST SCORE

User #	Score	User #	Score	User #	Score
1101886	1.7961	298505	1.3265	291415	1.2533
737312	1.7466	1190355	1.3265	73849	1.2491
811831	1.4072	37306	1.3242	926369	1.2462
788796	1.3941	950493	1.3166	425760	1.2456
1236163	1.3727	1320437	1.314	911475	1.2455
1218885	1.343	869773	1.2905	416634	1.2387
811525	1.3371	775285	1.2727	693080	1.2168
254471	1.334	847913	1.2632	1310074	1.215
923128	1.3332	194637	1.26	1122233	1.215
1342520	1.3331	399972	1.256	1167561	1.2146

The scores of users have only relative meanings. The exact determination of seed users depends on circumstances. For example, seed user can be defined to be the users with score above 1.0, or with score in top 1% highest of all scores, or with score in top 1000 highest of all users, etc. Note that, whatever criteria, there are some "super" seed users, with scores much higher than that of other seed users. Further investigations show that these "super" seed users actively involve in many threads of group messages, and deserved great attention in real applications.

IV. CONCLUSION

This paper considers how to find out seed users from SMS records. A directed graph based model is presented for the SMS forwarding behavior, which can handle more complicated situations than tree based model. Moreover, an iterative scoring method which resemble the "Page Rank" algorithm is given. Experiments demonstrate that the model and scoring method can achieve the goal. An interesting problem that is not touched in the current work is, how can one mine the community or the clique from the SMS records? Such sort of issued will be discussed in subsequent work.

ACKNOWLEDGMENT

This work is supported by National Supercomputer Center in Guangzhou (No. 2013Y2-00036).

REFERENCES

[1] [1] Yusuf L., Ramachandran U., Community membership management for transient social networks, st International Conference on Computer Communications and Networks, Munich, pp. 1–7, 2012.

[2] [2] Baratchi M., Meratnia N., Havinga P. J. M., On the use of mobility data for discovery and description of social ties, Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Niagara Falls, pp. 1229–1236, 2013.

[3] [3] Moon Yang-Sae, Choi Hun-Young, Kim Jinho, Choi Mi-Jung, A data mining approach to analyzing student-peer relationships from communication history records, International Journal of Innovative Computing, Information and Control. :3497–3513, 2013.

[4] [4] Kamath K. Y., Caverlee J., Transient crowd discovery on the real-time social web, Proceedings of the 4th ACM International Conference on Web Search and Data Mining, Hong Kong, pp. 585–594, 2011.

[5] [5] Dilrukshi I., De Zoysa K., Caldera A., Twitter news classification using SVM, Proceedings of the 8th International Conference on Computer Science and Education, Colombo, pp. 287–291, 2013.

[6] [6] Dinu A., Short text categorization via coherence constraints, Proceedings of 13th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, Timisoara, pp. 247–250, 2012.

[7] [7] Song Yangqiu, Wang Haixun, Wang Zhongyuan, Li Hongsong, Chen Weizhu, Short text conceptualization using a probabilistic knowledgebase, International Joint Conference on Artificial Intelligence, Barcelona, pp. 2330–2336, 2011.

[8] [8] Kaschesky M., Sobkowicz P., Bouchard G., The application and research of classification of POCP SMS based on text mining, International Journal of Digital Content Technology and its Applications. :59–65, 2011.

[9] [9] Hangya V., Farkas R., Target-oriented opinion mining from tweets, Proceedings of the 4th IEEE International Conference on Cognitive Infocommunications, Budapest, pp. 251–254, 2013.

[10] [10] Fink C. R., Chou D. S., Kopecky J. J., Llorens A. J., Coarse- and fine-grained sentiment analysis of social media text, Applied Physics Laboratory., :22–30, 2011.

[11] [11] Leong Chee Kian, Lee Yew Haur, Mak Wai Keong, Mining sentiments in SMS texts for teaching evaluation, Expert Systems with Applications., :2584–2589, 2012.

[12] [12] Sykora M. D., Jackson T. W., O'Brien A, Elayan S., Emotive ontology: Extracting fine-grained emotions from terse, informal messages, International Journal on Computer Science and Information Systems., :106–118, 2013.

[13] [13] Li Yongli, Wu Chong, Wang Xudong, Wu Shitang, A tree-network model for mining short message services seed users and its empirical analysis, Knowledge-Based Systems. :50–57, 2013.