2011 Second International Conference on Networking and Distributed Computing

Intelligent Viral Marketing algorithm over online social network

Yin Gui-sheng Wei Ji-jie Dong Hong-bin Li Jia
Harbin Engineering University
Harbin, China
E-mail:weijijie@hrbeu.edu.cn

Abstract—**As the online social network become increasingly popular nowadays, performing viral marketing over it has become the focus of many marketing management. How to select a fixed number of initial users from the total population with the purpose of maximizing the profits has long been open as a typical discrete approximation problem. However most of the existing solutions under the setting of online social network tried to traverse every node using network properties which is time-consuming and ineffective. This paper attacks the problem successfully by implementing intelligent algorithms such as GA, DE, PSO. Considering of the huge search space, we sharply decrease the scalability of the network through analyzing the datasets and sampling the data according to a power law property. Experiment results showed that the model we designed for solving viral marketing problem outperform other current search methods.**

Keywords-Social Network; Viral Marketing; Intelligent Algorithm

I. INTRODUCTION

The viral marketing problem was originally studied in the area of economics. Marketing mangers tried to promote the product by offering free samples to a marginal size of users, (Presume that the product is in good quality and useful) the product will soon be accepted by consumers by virtue of word-of-mouth effect. In recent years, online social networks websites are popular among young people, such as Facebook, Myspace, and Twitter which enable us to connect with friends and colleagues, share photos, stories and new events. While users received from their friends invitations or information about using some interesting application, they probably will try to use it too. This phenomenon is what we called influence spreading over social network. Online social network is an ideal destination for implementing viral marketing in that it can reach a large number of users through its friends-like links and the influence spreading is quite efficient through online network. Thus, spread of influence through online social networks can be overwhelmingly more efficient than that of interpersonal processes.

However, how to choose these initial users as seeds to maximize the profit is challenging, because the large scale of the network prohibit us from implementing search algorithms freely. To address this problem, this paper took an in-depth analysis of two popular social networks, classified these networks according to the connectivity distribution. power law shows that the scalability of these networks can be reduced sharply. Then we introduced the intelligent algorithms to search optimal seeds which achieved optimistic results in comparison to current methods.

A. Paper organization

Section 2 presents an overview of relevant backgrounds and related work. We propose our model in section 3 where the diffusion model and the analysis of two social networks are included as well as the search algorithms used. Section 4 shows our experimental results and the last section concludes the paper.

II. RELATED WORK

The problem of viral marketing was originally introduced to the field of computer science by Domingos and Richardson [1] and formalized by Kempe, Kleinberg, Tardos [3] who not only proved that the optimization problem is #N-P hard but also provided a greedy approximation algorithm with a provable approximation guarantee (1-1/e-ε of the optimal solution) based on submodular property. In considering of the large-scale social network with tens of thousands of users, Leskovec et al [9] proposed a "lazy-forward" optimization in selecting seeds, which significantly speed up the process. Similarly out of this consideration, Wei Chen et al [7] proved that the problem of computing influence spread given a seed set is #P-Hard and presented a new heuristic scheme using a local arborescence structure which showed to be the most efficient and scalable algorithm in their experiment. The huge number of users among the network prevent optimizing this problem effectively through using intelligent heuristic algorithms, thus, all the above work dedicated to solve the problem by using the characteristics of physical structure of social network and the result is not inspiring.

Forrest Stonedahl et al [8] are among the first to use intelligence algorithm to solve viral marketing optimization. However, GA was used to search weight of combination strategies (such as degree, clustering coefficient etc) instead of the seed itself since search a certain number of initial users from an extremely large population is unrealistic. Forrest Stonedahl et al [8] defined a problem named local viral marketing problem (LVMP) which is opposed to global viral marketing problem (GVMP). Unlike LVMP, GVMP provide us with a directed graph G, set of vertices and edges,

978-0-7695-4427-4/11 $26.00 © 2011 IEEE
DOI 10.1109/ICNDC.2011.69

where each vertex represents a consumer in the network and each edge labeled a number represents the connection between two consumers through which certain amount of influence can be exerted. It is conventionally believed that LVMP is more relevant to the real-world than GVMP for the reasons of privacy protection and computational complexity. Nevertheless, Amit Goyal, Francesco Bonchi, Laks V.S.Lakshmanan [4] addressed the issue discussed above, which makes possible the access of GVMP from a social graph with a log of actions. They not only proved the phenomenon of influence propagation among the real social network but also get the probabilities of influence between two users. Thus this paper attacked the problem of GVMP and it would give us a better view of how a user is influenced by others.

In contrast to previous work, this paper solved the problem via directly using the popular intelligence algorithms, such as GA, DE, PSO to search seeds, to accomplish this task and overcome the problem of scalability, we made an in-depth analysis of two real-life social networks through which the search space reduced exponentially.

III. MODEL

A. Diffusion Model and Adoption Function

There are two fundamental diffusion models generalized by Kempe et al.[3], namely, Independent Cascade Model and Linear Threshold Model, which are proved to be equivalent after generalizing to the general threshold model and general cascade model. Given a social network in form of a directed graph *G(V, E)*, with its nodes in set *V* representing users and edges in set E representing social relations through which the effect of word-of-mouth happens. At first, all users are presumed to be inactive except for the users selected as seeds, as time unfolds, more and more inactive users become active (or adopt the product) influenced by the initial seeds, which will in turn trigger further adoptions.

A. Independent Cascade Model (IC)

Given a seed set *S0*∈ *V*, the IC model works as follows. At each time stamp, every node St ∋ u (St denotes the set of users activated at time *t*) will exert an independent probability of *p(u, v)* to influence their currently inactive neighbors. It is worth noting that whether or not user *u* can successfully activate user v to accept the product at time t, it cannot make any further attempt to activate *v*. This means the influence exerted by user u cannot last long or be accumulated. This process proceeds until *St'=Φ* at some time t'.

B. Linear Threshold Model (LT)

In the LT model, each user's tendency to become active increases monotonically as more of its neighbors become active. At each time stamp, each node *v* check its out neighbors, if formula (1) is satisfied, then user *v* become active, aset represents the set of active users among v's neighborhood, *f(u,v)* is the influence that user *u* exert over user *v*, θv is the threshold that would make user v to accept the product, normally, this value is set arbitrary (say 1/2 or over) in the system for simple of computation. Similarly, this process succeeds until no more users become activate. In considering of the characteristics of our datasets, we choose LT model as our diffusion model, what's more, it is a relatively more immediate translation of word-of-mouth effect and similar to human thinking.

$$\sum_{u \in aset} f(u,v) \geq \theta_v \quad (1)$$

C. Adoption Function

Let Act(*S0*) denote the number of users be activated given a seed set of *S0*, the adoption function can be defined as follows:

$$Act(S_0) = \sum |S_t| \text{, where } |S\ t+1|=0.$$

D. Datasets Analysis

In examining the social networks that we used, we found that at least over half of the total users were alone, that means such users cannot affected by others nor can they exert influence to their friends, so time spend in searching these users is a large waste. However, how large a proportion of users is useless or can be deemed as nearly useless is connected to the topology of the social network. Motivated by this, we made an in-depth analyze of the real-life social networks.

Two datasets are used in this research, which is extract from real-world online social networks based on a daily snapshot. They are robots_net and squeakfoundation,.The first one is a Web Community site related to robotics, a daily snapshot of it has been collected since May 10,2008. In this dataset, edges between users are specified in three levels: Apprentice, Journeyer and Master. We map it to the value of influence probability in the following way: Apprentice=0.6, Journeyer=0.8, Master=1.0. The second dataset is similar to the robot.net.

In [1] Weiwei Yuan et al have proved that the datasets presented above were all small world networks. Specifically speaking, its large clustering coefficient and relative short diameter quantified by average path length is consistent with the characteristic of small world network. In [10] L. A. N. Ameral classified the small world network into three categories, (a) scale-free network characterized by a vertex connectivity distribution that decays as a power law; (b) broad-scale network characterized by a connectivity distribution that has a power law regime followed by a sharp cutoff; and (c) single-scale network, characterized by a connectivity distribution with a fast decaying tail. In order to further recognize the specific type of the small world network, the structure of the datasets has been probed through distribution of connectivity.

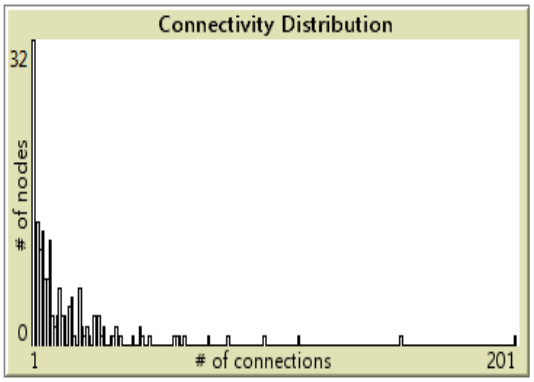

Figure 1. connectivity distribution of squeakfoundation

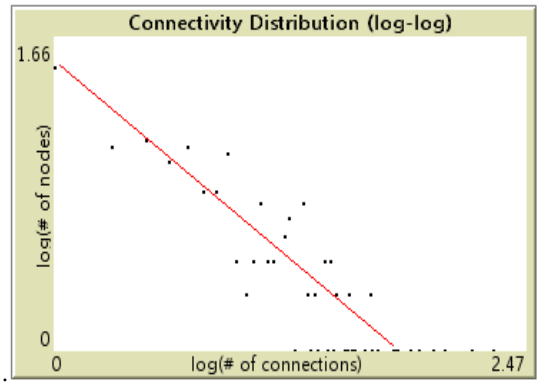

Figure 2. connectivity (in log-log) distribution of squeakfoundation

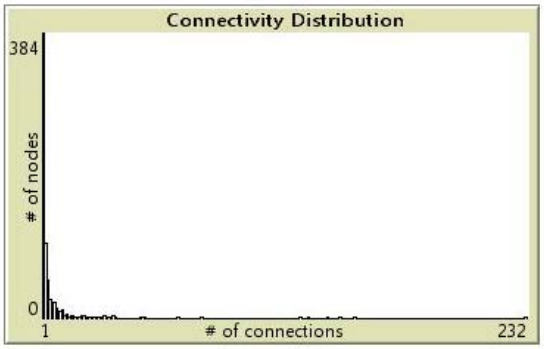

Figure 3. Connectivity distribution of robots_net

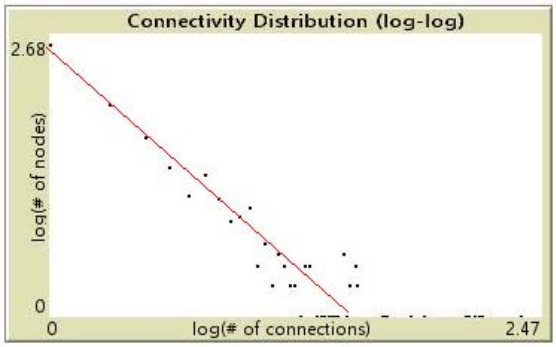

Figure 4. Connectivity (log-log) distribution of robots_net

Seen from Figure 1, Figure 2, Figure 3, Figure 4, it is obvious that the connectivity distribution is extremely disproportionate. That is, most of the nodes in the network only connect to small amount of other nodes while a few huge nodes connect to large number of other nodes. Therefore, however random the algorithms is, there is an extremely high possibility to select those alone nodes. That's why the direct use of intelligent algorithms which exert upon the total population in the network would be ineffective. Figure 2, Figure 4 showed that there is power law dominates among the two datasets which indicate that the small world presented above were all scale-free networks. It is especially apparent in robots.net while not so obvious in squeakfoundation, which may stem from the population size which is of great importance (squeakfoundation is smaller in size) since it is a statistical approximation. This information is particularly useful in that if there are 10 users with 1000 connections in the scale-free network, then there are 1000 users with only one connection. Theoretically speaking, when we try to choose initial users as seeds, we can set aside these 1000 users since they can at most influence one person to accept the product. The table below give us a more clear view of the connectivity distribution of the two networks.

TABLE Ⅰ. STRUCTURE OF TWO DATASETS

	Total number of nodes	Alone nodes	Average connection	Nodes above average connection
robots_net	15966	15268	0.223	698
squeakfoundation	769	592	3.507	120

It can be seen from the table that over half of the nodes were alone which can be definitely excluded from the population. The average connection of dataset robots_net's was sparse while the other was relatively denser signifying a better connectivity between nodes. Based on the above analyze, we can sharply reduced the scale of the social network, next, by sampling the nodes according to the connectivity distribution which can help to keep the diversity of the solutions, it is possible to search the seeds using intelligent algorithms.

E. Search algorithms

Fitness is evaluated by the total number of ultimate active users, so our goal is to maximize it through its genetic search mechanism. The search process is started by randomly selecting 100 solutions from search space which is optimized and sampled according to connectivity distribution. Each solution consists of a certain number of users which is defined as seed sizes and each user included in the solution is analogue of a gene, through simulation of crossover and mutation in GA, best solution can be selected. And DE is different from GA primary in crossover process. PSO simulate the action of birds, the group of users selected as seeds are likened to the position of the bird, as the algorithm unfolds, the best position is recorded through adjusting of both speed and position of each individual in the population.

The essence lies behind those intelligent algorithms is that this powerful random search algorithms enable us to concentrate on the solution itself other than the topology of the network which is complex and uncertain.

Ⅳ. EXPERIMENT

The experiment was carried out by NetLogo, we first scan the datasets and simulate the social network by creating agents corresponding to users in the network. Then model the LT diffusion process ($\theta v = 0.8$) and embed Java code to implement the intelligent algorithms. We also perform the CP (connectivity preferential) algorithm and random algorithm as comparison.

In GA search, we set crossover probability as 0.85, mutation probability 0.08, this process is repeated for 200 generations in both datasets. In DE search, the parameters setting is similar to that of the GA search and in PSO search, the inertia weight ω is set to be 0.9, cognitive parameter c1=c2=1. All the above running result in a total of 2 datasets × 3 algorithms × 100 individuals ×200 generations× 50 searches = 6 million fitness evaluations. The table below shows the final results of those algorithms.

TABLE Ⅱ. RESULT OF DIFFERENT ALGORITHMS

	Squeak-foundation 20	Squeak-foundaton 50	robots_net 20	robots_net 50
Random	168	195	65	278
CP	166	176	393	408
GA	178	230	425	480
DE	170	177	410	425
PSO	174	205	417	452

For each algorithm we took an average of 50 runs, the result in Table 2 shows that the algorithm random performs better in dataset squeakfoundation than in dataset robots_net even take an average of 50 searches. The reason might lies in the two datasets' average connection. The denser the dataset is (squeakfoundation), the larger the final active users will get through random search. For the dataset robots_net random search is more inclined to get those alone notes which reduce its number of adoption. Moreover, it can be see clearly that each of the three intelligent algorithms outperform the former traditional methods. And among those three algorithms, GA get the best result with PSO followed and DE is not that good in performance, especially when the seed sizes increased the enhance of adoption is marginal. It is worth noting that for connectivity preferential (CP) algorithm the adoption increase is marginal as the seed size grows. This may attribute to the reason that huge nodes (users with big connectivities) often tend to share same friends. So choosing those popular users will not leads to wider acceptance but is a kind of waste since one of them is enough to activate the same set of users.

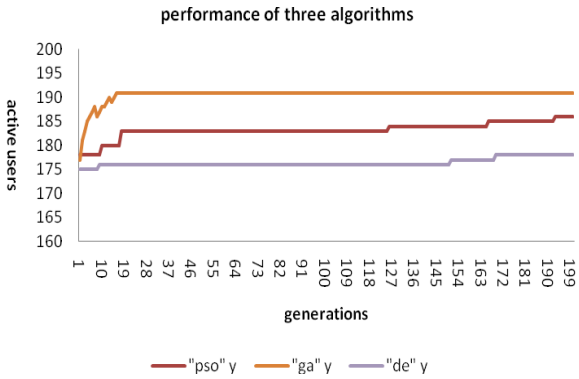

Figure 5. Performance of three algorithms over dataset squeakfoundation with seed size 20

Figure 5 shows the performance of three algorithms in dataset squeakfoundation (performance under other parameters are alike). Clearly, GA outperforms other two algorithms both in the result and generations of converge that used.

Ⅴ. CONCLUSION

To summarize, by sampling the initial individuals according to a power law code which serves to reduce the scalability of large online social network, this paper makes possible the use of several popular intelligent algorithms to solve the Viral Marketing problem. The experiment result shows that our method performs better than conventional algorithms. We believe that other properties of social network as well as improved intelligent algorithms would probably contribute to the approach of optimal results and that will be the focus of our future work.

ACKNOWLEDGEMENT

This work is sponsored by the National Natural Science Foundation of China under Grant No. 60973075, the Natural Science Foundation of Heilongjiang Province under Grant No. F200937 and the Foundation of Harbin Science and Technology Bureau under Grant No. RC2009XK010003.

REFERENCES

[1] Matthew Richardson, Pedro Domingos, Mining knowledge-sharing sites for viral marketing, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, July 23-26, 2002, Edmonton, Alberta, Canada [doi>10.1145/775047.775057] K. Elissa, "Title of paper if known," unpublished.

[2] Cong Yu, Laks Lakshmanan, Sihem Amer-Yahia, It takes variety to make a world: diversification in recommender systems, Proceedings of the 12th International Conference on Extending Database

Technology: Advances in Database Technology, March 24-26, 2009, Saint Petersburg, Russia.

[3] David Kempe, Jon Kleinberg , Éva Tardos, Maximizing the spread of influence through a social network, Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, August 24-27, 2003, Washington, D.C. [doi>10.1145/956750.956769].

[4] Amit Goyal , Francesco Bonchi , Laks V.S. Lakshmanan, Learning influence probabilities in social networks, Proceedings of the third ACM international conference on Web search and data mining, February 04-06, 2010, New York, New York, USA [doi>10.1145/1718487.1718518]

[5] Weiwei Yuan, Donghai Guana, Young-Koo Leea, Sungyoung Leea and Sung Jin Hur, Improved trust-aware recommender system using small-worldness of trust networks, Knowledge-Based Systems Volume 23, Issue 3, April 2010, Pages 232-238.

[6] Mani R. Subramani, Balaji Rajagopalan, Knowledge-sharing and influence in online social networks via viral marketing, Communications of the ACM, v.46 n.12, December 2003 [doi>10.1145/953460.953514]

[7] Wei Chen, Chi Wang, Yajun Wang, 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, July 25-28, 2010, Washington, DC, USA [doi>10.1145/1835804.1835934] .

[8] Forrest Stonedahl, William Rand, Uri Wilensky, Evolving viral marketing strategies, Proceedings of GECCO'10 Proceedings of the 12th annual conference on Genetic and evolutionary

[9] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, Natalie Glance, Cost-effective outbreak detection in networks, Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, August 12-15, 2007, San Jose, California, USA [doi>10.1145/1281192.1281239]

[10] L. A. N. Amaral, A. Scala, M. Barthélémy, H. E. Stanley, classes of small-world networks, PNAS October 10, 2000 vol. 97 no. 21 11149-11152.

[11] W ilensky, U. NetLogo. http://ccl.northwestern.edu/netlogo/, 1999.