Bringing Order into the Samples: A Novel Scalable Method to Influence Maximization.

[1] Nikita M. Sable, [2] Jagdish Pimple
[1] Student, Department Of Computer Science and Engineering, Nagpur Institute of Technology
[2] Professor, Department Of Computer Science and Engineering, Nagpur Institute of Technology
[1] niksabby4003@gmail.com, [2] pimplejagdish@gmail.com

Abstract: As a key problem in the social network, Influence Maximization (IM) has received extensive study. Since it is a well-known NP-complete problem, it is a great challenge to determine the initial diffusion seed nodes especially when the size of social network increases. In viral marketing, influence maximization has been extensively studied in the literature. Given a positive integer k, a social network G and a certain propagation model, it aims to find a set of k nodes that have the largest influence spread. The state-of-the-art method IMM is based on the reverse influence sampling (RIS) framework. By using the martingale technique, it greatly outperforms the previous methods in efficiency. However, IMM still has limitations in scalability due to the high overhead of deciding a tight sample size. In this paper, instead of spending the effort on deciding a tight sample size, we present a novel bottom-k sketch based RIS framework, namely BKRIS, which brings the order of samples into the RIS framework. By applying the sketch technique, we can derive early termination conditions to significantly accelerate the seed set selection procedure. Moreover, we provide a cost-effective method to find a proper sample size to bound the quality of returned result. In addition, we provide several optimization techniques to reduce the cost of generating samples order and efficiently deal with the worst-case scenario. We demonstrate the efficiency and effectiveness of the proposed method over real world datasets. Compared with the IMM approach, BKRIS can achieve up to two orders of magnitude speedup with almost the same influence spread. In the largest dataset with 1.8 billion edges, BKRIS can return 50 seeds in 1.3 seconds and return 5,000 seeds in 36.6 seconds. It takes IMM 55.32 second and 3,664.97 seconds, respectively.

Keywords: Reverse influence sampling (RIS), Influence maximization(IM) bottom-k reverse influence sampling (BKRIS),Threshold difference greedy algorithm(TDG).

1. INTRODUCTION.

Collective behavior refers to the behavior that is diffused or dispersed over large distances. The web and the social media have allowed for such rapid distribution of information around the world. Research shows that people trust information obtained from their close social circle far more than information obtained from general advertisement channels. Thus a minor piece of information can pass from ear to ear in a network and become a viral phenomenon. This type of information dispersal has led to an increase in interest among researchers in modeling the spread of such diffusion among a given population. A social network can be modeled as a graph with nodes representing individuals and edges representing connections, or relationships, between individuals. Information, behavior (e.g., joining a protest, adopting a fad), and other entities that can be propagated through a social network are referred to as contagions. There are two models of diffusion that are widely studied in the area of social networks - simple and complex. In a simple contagion model, each individual can contract a contagion if only one of its neighbours possesses it.

An example of a simple contagion would be the spread of a contagious disease like the flu in a community wherein each individual has a chance of getting infected if he/she comes in contact with another infected person. In a complex contagion model, multiple sources of exposure are needed before an individual adopts the change of behavior. Online social networking websites like Twitter and Facebook have provided an effective medium for diffusing ideas and spreading influence. These social networking websites provide a platform for marketing products and businesses online. Due to budgetary constraints in marketing, the ideal strategy is to influence a set of users who will start using the product and who will in turn influence their friends to use the product and so on. Informally, the problem of influence maximization as defined is the problem of finding a small set of seed nodes (that initially possess a contagion) in a social network that maximizes the spread of influence. Kempe et al. proved that this optimization problem is NP-hard, and presented a greedy approximation algorithm guaranteeing that influence spread is (1-1/e-€) of optimal influence spread. Just as important, their work has motivated the development of a huge body of literature on the topic of influence maximization. Influence maximization has applications in viral marketing, feed ranking, recommendations and several other areas. A complex contagion requires multiple contacts for an individual to change his/her state and exhibits a different diffusion pattern compared to a simple contagion. Let us take a real-world example that exhibits complex contagion: a tense atmosphere prevailing in a city which can potentially lead to a public protest. It has been argued and demonstrated empirically that the possibility of an individual participating in a protest depends on the number of that person's neighbors who are already part of the protest. In particular, for this and other scenarios, we focus on the linear threshold model.

2. LITERATURE REVIEW.

A. Bring Order into the Samples: A Novel Scalable Method for Influence Maximization
Xiaoyang Wang, Ying Zhang, Wenjie Zhang, Xuemin Lin, Fellow, IEEE 2016 , Chen Chen.
In this paper, we investigate the influence maximization problem. Based on our analysis, the state-of-the art method IMM still has limitations in scalability in term of k and graph size. In this paper, we provide a more efficient solution, BKRIS, which integrates the bottom-*k* sketch with the RIS framework. Particularly,we bring the order of the samples into the RIS framework, which enables us to achieve possible early termination before materializing all the samples. To bound the quality of returned seeds, we propose an efficient heuristic approach to obtain a lower bound of OPT. We also propose several optimizations to deal with sample order generation and the worst case processing. We conduct extensive experiments on 10 real social network datasets. Compared with IMM, we can achieve up to 2 orders of magnitude speedup.

B. Bring Order into the Samples: A Novel Scalable Method for Influence Maximization (Extended abstract) 2017 IEEE 33rd International Conference on Data Engineering.
In this paper we have evaluated that the RIS framework, the computation cost is proportional to the sample size to obtain a result with theoretical guarantee, the sample size should be at least λ□/OPT, where OPT is the influence spread of the optimal seed set, and λ□ is related to the input graph, k and the error parameters. However, in IMM, the cost of deriving a tight lower bound of OPT becomes the bottleneck of this algorithm when k is large, and limits its scalability. To reduce the cost, in this paper, the developed method is based on a two step framework. In the first step, a reasonable large sample size is computed to bound the quality of the returned result. In the second step, we materialize the samples in the order and terminate when the stop condition is satisfied. In this case, even though the sample size may be large, we can accelerate the search by terminating in advance.

C. Data Mining-Based Decomposition for Solving the MAXSAT Problem: Toward a New Approach. 2017 IEEE INTELLIGENT SYSTEMS Published by the IEEE Computer Society.

This paper indicates that the Apriori DPLL outperforms other algorithms in terms of success rate and has a very competitive runtime. For future work, we plan to investigate other heuristics to deal with the conflict problem caused by the separator variables. We are also planning to apply the two suggested approaches to other optimization problems such as weighted MAXSAT, the coloring problem, and the CSP problem. Finally, proposing a parallel version that explores high-performance computing to solve very large MAXSAT instances is also in our agenda.

D. An Algorithm for Influence Maximization and Target Set Selection for the Deterministic Linear Threshold Model Influence maximization, complex contagion, linear threshold Copyright 2014, Anand Swaminathan.

The research in the field of influence maximization is growing rapidly. We have looked into a broad range of algorithms from the literature, and have provided a brief summary for each of the algorithms examined. In this thesis, I have presented the threshold difference greedy (TDG) algorithm for the deterministic linear threshold model which addresses both the influence maximization problem as well as the target selection problem. With extensive experiments on 14 real-world networks of varying size and density, I have shown that the novel approach using vertex thresholds is better than the seven other algorithms taken from the literature. Since the execution time is a crucial factor for any influence maximization heuristic, the tuneable parameters are essential in controlling the execution times. The performance improvement attained through threading was important in reducing the execution times for several graphs. Through this thesis, I have shown that the threshold of a node is an important input for a graph and data mining techniques to compute threshold values for real world networks can provide more accurate results for the influence maximization problem.

3. PROBLEM STATEMENT.

Given a social network G, a constant integer k and a probabilistic diffusion model M, the problem of influence maximization is to find a set S of k nodes in G which has the largest influence spread.

S = argmax

S□V {σ(S) | |S| = k}

In this paper, we study the case when M is set as IC model or LT model.Problem Hardness: The influence maximization problem is NP-Hard in both IC model and LT model. Chen et al.have proved it is P-Hard to calculate the influence spread of a seed set in both models. Fortunately, the influence spread function σ(S) satisfies the following two properties.

• Monotonic property. For S, T ⊆ V and S ⊆ T, we have σ (T) ≥ σ(S).

• Submodular property. For S, T ⊆ V, S ⊆ T and u ∈ V ∧ u /∈ S ∪ T , we have σ(S ∪ {u}) − σ(S) ≥ σ(T ∪ {u}) − σ(T).

Based on these two properties, we can iteratively select the node with the largest marginal influence until k nodes have been found. It can return a result with an approximation ratio of 1−1/e, if the influence spread is exactly calculated.

4. PROPOSED APPROACH.

As a key problem in viral marketing, influence maximization has found many important applications in real life. Kempe et al. first formalize the influence maximization problem.

In the seminal paper, it defines two models, the independent cascade (IC) model and the linear threshold (LT) model, to simulate the influence spread. Also, a greedy algorithm is proposed in to return a result with 1-1/e-€ approximation ratio. However, the naive greedy method is known to be inefficient in practice. Due to the importance of the problem, it motivates a lot of follow-up works to improve the performance under both models. Recently, Borgs et al. develop an elegant framework, called reverse influence sampling (RIS), to solve the influence maximization problem. IMM further reduces the sample size in RIS by using novel techniques. However, IMM still has limitations in scalability. In this paper, we will propose the BKRIS framework, which accelerates the RIS framework by introducing the order of samples based on the bottom-k sketch. In addition, an efficient method is developed to derive a sufficient and reasonable large sample size. Also, several optimization techniques are given to reduce the cost of generating and processing samples.

5. EXPECTED OUTCOME.

In this paper we will extract the live data from twitter through the tweets on the particular book we have entered and based on the tweets we will analyze the results. From the analyzed results we will come to know that how many people refer that particular book and their views for that book, what people talk about that book and how many people speak positive and how many speak negative about that book. This will help us to know or analyze that how good the book is and what it is related to so that it becomes easy for the readers to refer or read the books. It will also show the analyzed result in the form of bar graph.

REFERENCES.

[1] Bring Order into the Samples: A Novel Scalable Method for Influence Maximization. Xiaoyang Wang, Ying Zhang, Wenjie Zhang, Xuemin Lin, Fellow, IEEE 2016 , Chen Chen.

[2] Bring Order into the Samples: A Novel Scalable Method for Influence Maximization (Extended abstract) 2017 IEEE 33rd International Conference on Data Engineering.

[3] Data Mining-Based Decomposition for Solving the MAXSAT Problem: Toward a New Approach. 2017 IEEE INTELLIGENT SYSTEMS Published by the IEEE Computer Society.

[4] An Algorithm for Influence Maximization and Target Set Selection for the Deterministic Linear Threshold Model Influence maximization, complex contagion, linear threshold Copyright 2014, Anand Swaminathan.

[5] Summarizing Data using Bottom-k Sketches.Edith Cohen AT &T Labs. Research 180 Park Avenue Florham Park, NJ 07932, USA. Haim Kaplan School of Computer Science Tel Aviv University Tel Aviv, Israel August 12.15, 2007, Portland, Oregon, USA.Copyright 2007.

[6] Java, Java Development Kit (JDK) and related terms are all copyright by Oracle Corporation. http://www.oracle.com.

[7] D. Kempe, J. M. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in SIGKDD, 2003, pp. 137–146.

[8] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in KDD, 2009, pp. 199–208.

[9] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in ICDM, 2010, pp. 88–97.

[10] C. Borgs, M. Brautbar, J. T. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in SODA, 2014, pp. 946–957.

[11] E. Cohen and H. Kaplan, "Summarizing data using bottom-k sketches," in PODC, 2007.