ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.5

# A Survey Paper on Influence Maximization in Online Social Network

Praneetha .G. N[1], Bhavatarini .N[2], Chaithra .B .M[3], Tejaswini .B .S[4]
Assistant Professor[1, 2, 3, 4]
Department of IS&E
Sapthagiri College of Engineering, Bengaluru, India

**Abstract:**
People are becoming more interested in online social network and they depend on the social network for many purposes such as finding opinions of other people about any product, movie etc. Influence Maximization is the problem of finding a small set of influential users in the online social network so that their influence in the social network is maximized. There are many diffusion models like Linear Threshold Model and Independent Cascade Model that are used to find the maximum influential user in online social network. This paper presents a survey on these two models and extensions to these models.

**Keywords:** Influence Maximization, Multiple Online Social Networks.

## I. INTRODUCTION

The advent of Online Social Network (OSN) has been one of the most exciting events in this decade. Many popular OSN such as Facebook, Twitter, Linkedln and Flickr have become increasingly popular. These networks are extremely rich in content and linkage data which can be analyzed. The linkage data is essentially the graph structure of social network and the communication between nodes, whereas the content data contains the text, images and other multimedia data in social network. The richness of this network provides opportunities for data analysis in context of Online Social Network. There are several factors due to which the OSN has gained importance by researchers[1].

Some of the factors are availability of social data that are vast, distributed, noisy and dynamic. There are some research issues with respect to mining the social network sites using data mining techniques.

One of the issues is Influence Propagation. In many markets, customers are strongly influenced by the opinions of their friends. *Viral marketing* takes advantage of this to promote a product by marketing it primarily to those with the strongest influence in the market. Further people trust and act on recommendations from friends and their further influence their friends. This is referred to as *influence propagation*. Influence propagation has become an important mechanism for viral marketing.

This further motivates the researchers to carry out extensive studies on various aspects of the influence propagation problem. *Influence Maximization problem* is a problem of finding a small set of nodes that maximizes the spread of influence. Influence Maximization problem was first studied by Domingo's and Richardson[2] and proposed first algorithm for influence propagation.

Then, Kempe et al.[3] gave two fundamental propagation models, named Independent Cascade (IC) Model and Linear Threshold (LT) Model. Many other researchers extended this basic propagation models in terms of scalability and efficiency. But most of the works focussed on a single online social network whereas users now often are found in more than one social network. Dung T. Nguyen et. Al [10] proposed an algorithm to handle this problem.

## II. RELATED WORK

### Probabilistic Model

*Domingo's and Richardson* [2] gave the first algorithm to deal with influence propagation problem. They built probabilistic models of influence for mining the data on *knowledge-sharing websites*. Knowledge-sharing sites are the sites where customer review products and advise each other about the products. Customer's have two types of values:

*intrinsic value* and *network value*. Intrinsic values of a customer is his values as a customer based on the products he is likely to purchase and the network value of a customer is high when he is expected to have a very positive influence on other's probabilities of purchasing the product. A customer with high network values is the one who is worth of marketing.

It concluded that by building the *probabilistic models* and applying those models to the knowledge sharing websites, solved the influence propagation problem and their method is scalable to large networks. But, the method mined a network from single source and not from multiple sources.

The model was built based on Epinions data. The model was first tested with respect to Boolean Marketing. Experimental result showed that viral marketing resulted in profit increase over direct marketing and no marketing. The model which was introduced was linear model and it had tremendous speed over a non-linear model.

Then, the model was tested against Continuous Marketing, where viral marketing was advantageous over direct marketing. It was also showed that even with less network knowledge, viral marketing methods was better than direct marketing. The table shows the profit results for Boolean Marketing and Continuous Marketing scenario for various

costs of marketing based on [1]. There if lift of 3.24% profit in Continuous Viral Marketing over Boolean Viral marketing.

**Table.1. Profit result for Boolean marketing and Continuous marketing.**

| | Boolean Marketing | Continuous Marketing |
|---|---|---|
| No Marketing | 37.78 | 37.78 |
| Direct Marketing | 66.08 | 68.38 |
| Mass Marketing | 70.23 | 71.28 |

**Diffusion Models**

Then, *Kempe et al.* [3] studied influence propagation by focusing on two fundamental propagation models, named *Independent Cascade (IC) Model* and *Linear Threshold (LT) Model*. Here, a social network is modelled as a graph with nodes representing individuals and edges representing connections or relationship between two individuals. Influences are propagated in the network according to the model, such as the independent cascade (IC) model. Kempe et al. proved that the optimization problem is NP-hard, and present a greedy approximation algorithm guaranteeing that the influence spread is within the optimal influence spread.

In LT model, a node v is influenced by each neighbour w according to a weight $b_{v,w}$ such that

$$\sum_{w \text{ active neighbor of } v} b_{v,w} \leq 1 \quad \ldots\ldots \text{Eq.1}$$

Step 1: Each node chooses a threshold from the interval [0,1].
Step 2: Initially A0 is a set of nodes that are active.
Step 3: All nodes that were active in step 2 remains active, and these active nodes tries to activate its neighbour nodes. This is according to the equation give below

$$\sum_{w \text{ active neighbor of } v} b_{v,w} \geq \theta_v \quad \ldots\ldots\ldots \text{Eq.2}$$

In IC Model, we again start with an initial set of active nodes A0, and the process unfolds in discrete steps according to the following randomized rule. When node v first becomes active in step t, it is given a single chance to activate each currently inactive neighbour w. If v succeeds, then w will become active in step t+1; but whether or not v succeeds, it cannot make any further attempts to activate w in subsequent rounds. Again, the process runs until no more activation is possible. The main disadvantage of the algorithm is that it was less efficient. This model was tested using a collaboration graph which was obtained from co-authorships in physics publications. It was considered that co-authorship network captured many of the important features of social network. This resulted in a graph of 10748 nodes and edges between 53000 pairs of nodes. LT Model and IC Model was compared against degree and centrality-based algorithms. The algorithm for LT Model outperforms high-degree node heuristic by about 18% and central node heuristic by 40%. This showed that better marketing results can be obtained by considering the dynamics of information in network than just concentrating on structural properties of graph. The IC model was tested with probabiltity 1% and 10%. It activated a large fraction of network, which was almost 25% better than other algorithms.

**Heuristic Algorithm**

*Chen et al*. proposed a new propagation model similar to the greedy algorithm[4,5] but with a better efficient result. Scalability problem of IC Model is addressed by proposing a new heuristic algorithm. The main idea of heuristic scheme is to use *local arborescence structures* of each node to approximate the influence propagation. *Maximum influence paths* (MIP) are computed between every pair of nodes in the network via a *Dijkstra shortest-path algorithm*, and ignore MIPs with probability smaller than an influence threshold. The union of the MIPs starting or ending at each node into the arborescence structures, which represent the local influence regions of each node, is constructed. Influence propagated through this local arborescence is only considered, and this model is referred as the *maximum influence arborescence (MIA) model*. *Chen et al.* proposed the first scalable heuristic algorithm for influence maximization in the LT model which was referred as LDAG algorithm [6].

They constructed a local DAG for every node in the network and restricted the influence for the node to be within the local DAG structure. To construct local DAG, they proposed a fast greedy algorithm where the nodes were added one by one to local DAG such that the influence of these individual nodes is greater than θ. They conducted experiments on four real-world networks and synthetic data sets. They conducted experiments to illustrate the performance of their algorithm with respect to 3 aspects: Scalability, Influence Spread and tuning of control parameter θ. Experimental result showed that Greedy and SP1M algorithm are not scalable and PMIA algorithm can scale up quite well. For testing influence spread, k values were taken as 50. PMIA algorithm was 3.9% and 11.4 % better than Greedy and SP1M algorithm. They investigated the effect of parameter θ on the running time and influence spread of PMIA algorithm. Running time increases when θ value decrease. Experimental result showed that as running time increases, the influence spread also increases. *Chen et al*. proposed the first scalable heuristic algorithm for influence maximization in the LT model which was referred as LDAG algorithm [6]. They constructed a local DAG for every node in the network and restricted the influence for the node to be within the local DAG structure.

To construct local DAG, they proposed a fast greedy algorithm where the nodes were added one by one to local DAG such that the influence of these individual nodes is greater than θ. The experiments were conducted on same four datasets as in [5]. Experimental results show that the LDAG algorithm is more than three orders of magnitude faster than the greedy algorithm and other methods have poor scalability. LDAG algorithm performs consistently among best algorithms in all the tests that were being performed. They also verified LDAG construction algorithm and it was efficient than RandDAG algorithm which randomly generated LDAGs.

**General Threshold Model**

*Goyal et al.* [7] also had made a study of the problem of learning influence probabilities using an instance of the General Threshold Model. The idea is that when a user sees their social contacts performing an action such as joining an online community, that user may decide to perform the same. In truth, when a user performs an action, there may be many reasons: heard it outside of OSN, action is very popular or genuinely influenced. All previous papers assumed that they are given as input a social graph with edges labelled by the probability with which a user's action will be influenced by her neighbour's action. How or from where probabilities are known is an issue. Hence Goyal et al. gave *action log* as an input along with the social graph. An action log is a table that contains any actions performed by every user. Algorithms are optimized to minimize the scans over the action log, a key input to the problem of inferring probabilities of influence. Three types of models were proposed to capture the probability with which one user influences its neighbour.

They are: Static Model, Continuous Model and Dynamic Model. This algorithm takes input as both the social network and action log. They considered Flickr social network and considered joining a group as the action. They compared the different models based on ROC curves. They examined the four static models and four discrete models. Results show that Bernoulli is slightly better than Jacard model and among two Bernoulli variants, Partial Credit is better. In conclusion, discrete time model achieves same quality as continuous time model but much more efficiently.

**Negative Opinions**

All of the above works ignored an important aspect of influence propagation i.e, negative opinions may also be propagated in the network. Negative opinions are much more contagious and stronger than the positive opinions in affecting the people's decision. Wei Chen et.al [11 ] proposed a new IC model called Independent cascade model with negative opinions (IC-N) which the extension of IC model. This model is associated with a parameter called quality factor (q). Each node will have three states:

neutral, positive and negative. Quality factor indicates the probability of a particular node staying positive after it is being activated by a positive neighbour. Every node u tries to active its neighbour node v, if it gets successful, v is activated at step t+1. If v is activated by negative activated node then v becomes negative node. Otherwise, if v is activated by positively activated node then v becomes positive node with some probability q and becomes negative node with probability 1-q. The MIA algorithm was also proposed for IC-N which is more scalable than IC-N.For experimentation, they used 3 data sets: NetHEPT, WikiVote and Epinions. They conducted experiments on greedy algorithm and MIA-N. They tested the effect of quality factor on influence spread and found the when q increase, the positive influence spread also increased in superliner fashion. With respect to influence spread, MIA-N algorithm is better than greedy algorithm and it is orders of magnitude faster than the greedy algorithm.

It takes only 11 minutes to process a graph of 256k nodes and 353k edges but the greedy algorithm takes more than 2 hours to process a graph four times smaller.

**Multiple Online Social Networks**

Existing works focused only on single online social network while users nowadays are found in several OSNs. Thus, it is important to study the influence in multiple social networks. Dung T. Nguyen et.al [10] proposed a method to handle this problem. Multiple OSN network are combined into one network while preserving the influential properties of the original network. After coupling the networks, LT model was being used to find the influential users. A new metric named Influence Relay was introduced which was used to study the flow of influence between the social networks.

**Interest matching users**

Most of the related works focused only on the topology of the network but ignored the factors such as users interest in the information that is being propagated. Yilin Shen et.al [9] proposed a new method Total Seed Nodes Learning (TSNL) to capture these interest-matching users whose interest match with each other. First, a new network was constructed from k multiple networks using network coupling. Then, interest matching users were found out using semi-supervising learning and then minimum influential users were found out based on idea called Iterative Semi-Supervised Learning.

## III. CHALLENGES

Two diffusion models i.e, Linear Threshold (LT) model and Independent Cascade (IC) model is used to find the influential user in online social network. The processing time of these model increases exponentially as size of network increases. This is the major disadvantage of the model. Hence, some efficient method should be identified to decrease the processing time of the model.

## IV. REFERENCES

[1].G.Nandi, A. Das, "A survey on using data mining techniques for Online Social Network Analysis", International Journal of Computer Science, November 2013

[2].M. Richardson and P. Domingos, "Mining knowledge sharing sites for viral marketing," in Proc. of the 8th ACM SIGKDD Int. Conf on Knowledge Discovery and Data Mining (KDD'02), 2002, pp. 61-70.

[3]. Kempe, J. M. Kleinberg, and ´E. Tardos, "Maximizing the spread of influence through a social network," in Proc. Of the 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'03), 2003, pp.137-146.

[4].Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. Van Briesen, and N. S. Glance, "Cost-effective outbreak detection in networks," in Proc. of the 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'07), 2007, pp. 420–429

[5].W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks", in Proc. of the 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'09), 2009.

[6].W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks", in Proc. of the 16th ACM SIGKDD Int.Conf.

on Knowledge Discovery and Data Mining (KDD'10), 2010, pp. 1029-1038.

[7].W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model", in Proc. of the 10th IEEE Int. Conf. on Data Mining (ICDM'10), 2010.

[8]. A. Goyal, F.Bonchi, and L.V.S Lakshman, "Learning influence probabilities in social networks", in Proc. Of the 3rd ACM Int. Conf. On Web Search and Data Mining (WSDM'10), 2010.

[9].Y.Shen, T.N. Dinh, H. Zhang, and M. T. Thai. "Interest-matching information propagation in multiple online social networks in CIKM, 2012.

[10].Dung T. Nguyen, Huiyuan Zhang, Soham Das, My T. Thai, "Least cost influence in multiplex social networks: Model Representation and Analysis, 2012 IEEE 13th International Conference on Data Mining.

[11].Wei Chen, Alex Collings, Rachel Cummings, Te Ke "Influence Maximization in Social networks when negative opinions may emerge and propagate"

[12].Tereza Iofciu, Peter Fankhauser, Fabian Abel, Kerstin Bischoff "Identifying Users Across Social Tagging Systems", Proceedings of the fifth international Conference