GA with *k*-Medoid Approach for Optimal Seed Selection to Maximize Social Influence

Sakshi Agarwal and Shikha Mehta

Abstract In this rapidly rising field of Web, volume of online social networks has increased exponentially. This inspires the researchers to work in the area of information diffusion, i.e., spread of information through "word of mouth" effect. Information maximization is an important research problem of information diffusion, i.e., selection of *k* most influential nodes in the network such that they can maximize the information spread. In this paper, we proposed an influence maximization model that identifies optimal seeds to maximize the influence spread in the network. Our proposed algorithm is a hybrid approach, i.e., GA with *k*-medoid approach using dynamic edge strength. To analyze the efficiency of the proposed algorithm, experiments are performed on two large-scale datasets using fitness score measure. Experimental outcome illustrated 8–16% increment in influence propagation by proposed algorithm as compared to existing seed selection methods, i.e., general greedy, random, discounted degree, and high degree.

Keywords *k*-Medoid · Genetic algorithm · Social influence · Seed selection · Topical affinity propagation

1 Introduction

Web market share of online social networks is increasing exponentially. Social network is becoming tremendously substantial for various applications, i.e., educational, matrimony Web sites, government division, job search portals, recommendation system [1], health care, viral marketing [2], and numerous other businesses. In all these applications, Social network is commonly considered as a platform for information propagation in network, i.e., social influence. In online social networks, activities of

S. Agarwal (✉) · S. Mehta
Computer Science & Information Technology, Jaypee Institute of Information Technogy, Noida, India
e-mail: sakshi.officialid@gmail.com

S. Mehta
e-mail: mehtshikha@gmail.com

A. Khanna et al. (eds.), *International Conference on Innovative Computing and Communications*, Advances in Intelligent Systems and Computing 1087,
https://doi.org/10.1007/978-981-15-1286-5_9

a person can lead to change in another person's behavior, i.e., social influence. This change in user's behavior depends on the other user's influence strength and information spreads from one user to another user. Spread of information depends on the user's position in the network. Selection of appropriate nodes becomes a challenge in order to gain maximum influence spread in the system, i.e., influence maximization. Influence maximization problem was proved to be NP-hard for numerous propagation models [3]. Therefore, this paper targets to maximize influence extent by determining k optimal nodes in the system using dynamic edge strength. Therefore, information propagation through these nodes will maximize the effect of influence in the network.

In the past studies, various models and algorithms have been introduced with respect to social influence [4]. Aslay et al. [5] presented the current state of the art of the influence maximization (IM) in the field of social network analysis (SNA), i.e., existing algorithms and theoretical developments in the field of IM. Anagnostopoulos et al. [6] described social influence identification problem methodically. They explained various social correlation models and proposed two methods that can identify influence in network using time-dependent user action information. Chen et al. [7] extended discounted degree approach to improve influence propagation in the system. Mittal et al. [8] identified that the centralities are the major elements to discover the important authors in collaboration networks. Chen et al. [9] proposed an algorithm which is scalable with respect to size of networks, i.e., social networks. In their algorithm, they used one tunable parameter which provides balance between the running time and influence spread in the network. Similarly, Khomami et al. [10] proposed learning automaton-based solution to identify minimum positive influence dominating set (MPIDS) to maximize influence propagation. Goyal et al. [11] solved influence maximization problem by their proposed model, i.e., credit distribution. Credit distribution method uses the propagation traces of each action in the time interval Δt and estimates the expected influence flow in the network. Chen et al. [12] introduced a directed acyclic graph-based scalable influence maximization approach that is modified with respect to linear threshold model. So, influence maximization is a rising field for which various theories and models have been introduced in the recent past years. Therefore, to optimize scope of influence in the network, we have presented a hybrid approach by selecting optimal seeds from the network. Kumar et al. [13] applied the concept of influence propagation to detect the rumor on social media. The remaining part of this paper is arranged as follows: In Sect. 1, we explained the importance of influence in social network. We have also described the role of good seeds in information propagation, and the related work has been done in this field. In Sect. 2, we explain the proposed algorithm, i.e., GA with k-medoid approach for optimal seed selection to maximize social influence. After this, in Sect. 3, we illustrate performance analysis of proposed methodology with respect to other methods on two datasets. Lastly, we enclose conclusion and future work.

2 Proposed Methodology

In this paper, we propose an influence maximization technique by discovering appropriate set of seeds that maximize the influence spread in the network. Our proposed algorithm is a hybrid approach and divided into three steps as shown in Fig. 1. In the first step, edge strength score is updated by topical affinity propagation (TAP) technique. In the second step, initial population set is generated using *k*-medoid algorithm. This initial population is given as an input to genetic algorithm (GA) to trace the optimal set of seeds that can increase the impact of influence up to maximize extent.

Given a directed network $G = (E; V; S, k)$, where E denotes the set of links/edges (u, v), i.e., u, v ЄG, set of nodes is symbolized by V, and k is a variable that denotes the number of seeds. S is the set of edge strength score w.r.t edge set E, where edge strength score s_{uv} of an edge (u, v) is

$$s_{uv} = \frac{1}{\text{out\_degree}(u)}$$

Objective of our algorithm is to find k optimal nodes as seed set V_1, where $V_1 \subseteq V$ and influence spread is maximized in network G through set V_1.

2.1 Dynamic Edge Strength Score

For a given network $G(V, E, S)$, dynamic strength score of every edge S_d is calculated using topical affinity propagation algorithm. TAP is an algorithm which computes the topic-wise influence likelihood of each node with respect to different node attributes as defined in algorithm 1. In this paper, we considered one topic per node. Therefore, no node attribute required to input in TAP. Table 1 defines the variables involved in the estimation of dynamic influence likelihood.

Node_score function is the base component of the TAP algorithm as defined by Eq. 1.

$$\text{node}(v_i, r_i) = \begin{cases} \frac{s_i r_i}{\sum_{j \in N(i)} (s_{ij} + s_{ji})} & r_i = \text{i} \\ \frac{\sum_{j \in N(i)} (s_{ji})}{\sum_{j \in N(i)} (s_{ij} + s_{ji})} & r_i \neq \text{i} \end{cases} \tag{1}$$

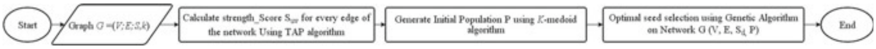
Start
Graph G =(V;E;S,k)
Calculate strength_Score S_{uv} for every edge of the network Using TAP algorithm
Generate Initial Population P using K-medoid algorithm
Optimal seed selection using Genetic Algorithm on Network G (V, E, S_d, P)
End

Fig. 1 Flowchart of the proposed algorithm

Table 1 List of symbols

v_i	A specific node in G
r_i	Representative node of node v_i with highest edge_sum for node set $\{N(v_i) \cup v_i\}$
e_{ij}	A link between nodes v_i and v_j
s_{ij}	Strength score of an edge e_{ij}
I_{ij}	Influence score of node v_i on node v_j
T_{ij}	Influence likelihood assumed by node v_j to node v_i, initial value = 0
A_{ij}	Influence likelihood node v_j approves to take on self from node v_i
$N(i)$	Set of all nodes having incoming edge from node v_i
Ln_{ij}	Logarithm of the edge_score of edge (i, j)

where r_i is the node with highest edge_sum value in the set $\{N(i) \cup i\}$ for node i identified using Eq. 2.

$$\text{Edge\_sum}(v_i) = \sum\nolimits_{k \in N(i)} s_{ik} \tag{2}$$

Similarly, logarithm of normalized edge_score Ln_i for node i, A_{ij}, T_{jj}, T_{ij} and I_{ij} calculated using Eqs. 3, 4, 5, 6, and 7, respectively.

$$\mathrm{Ln}_{ij} = \log \frac{\mathrm{node}^{(v_i, r_i)|_r}{}_{i} = j}{\sum_{k \in N(i) \cup \{i\}} \mathrm{node}^{(v_i, r_i)\,|_r}{}_{i} = k} \tag{3}$$

$$A_{ij} = \mathrm{Ln}_{ij} - \max\nolimits_{k \in N(j)} \{\mathrm{Ln}_{ik} + T_{ik}\} \tag{4}$$

$$T_{jj} = \max\nolimits_{k \in N(j)} \min\left\{A_{kj}, 0\right\} \tag{5}$$

$$T_{ij} = \min\left(\max\left\{A_{jj}, 0\right\} - \min\left\{A_{jj}, 0\right\} - \max\nolimits_{k \in N(j) \setminus \{\mathrm{i}\}} \min\left\{A_{kj}, 0\right\}\right), i \in N(j) \tag{6}$$

$$I_{ij} = \frac{1}{1 + e^{-\left(A_{ji} + T_{ji}\right)}} \tag{7}$$

Algorithm 1
Dynamic likelihood computation $G(E, V, S)$

```
Compute the influence value of each node $v_i$ using eq. 1
Compute $Ln_{ij}$ for each edge e(i,j) using eq. 3
For each $e_{ij,}$ initialize $T_{ij}$ = zero
Repeat till convergence
    For all $e_{ij} \in G$
        Compute $A_{ij}$ using eq. 4
    For all $v_j \in G$
        Compute $T_{jj}$ using eq. 5
    For all $e_{ij} \in G$
        Compute $T_{ij}$ using eq. 6
For all $v_i \in G$
    For every node $k \in \{i\} \cup N(i)$
        Compute $I_{ki}$ using eq. 7
Create $G^1$ (E, V, $S_d$), where $S_d$= I i.e. set of influence score of each edge.
```

Therefore, in the initial phase of our proposed algorithm, we updated the edge strength score from S to dynamic edge strength score S_d.

2.2 *Generate Initial Population Using* k-*Medoid Algorithm*

In this step, population set *P* of *k* nodes is generated such that this population easily converges in optimal time and generates optimal seeds to maximize influence propagation in the network. In the past studies, various selection approaches have been introduced such as graph-based heuristics and mathematical models with respect to various domain, i.e., shortest path optimization [14] or influence maximization [15]. In random model, selection of *k* nodes is arbitrary and does not depend on any of the property of node or edge attribute. It selects the *k* nodes in $O(1)$ time. In general greedy model, choices made on the best are at moment basis such that the objective function is optimized. Therefore, it chooses best solution at every step. Node with largest neighborhood is the strongest node of the network is the hypothesis of the next method, i.e., high degree. Therefore, this method uses the out-degree for the directed graphs and degree for the undirected graphs. Extension of high degree is known as discounted degree or single discount heuristic. This heuristic believes that neighborhood of neighboring nodes is not mutually exclusive [16]. In this paper, we applied clustering algorithm, i.e., *k*-medoid algorithm to generate *k* cluster centers as population set *P* as described in algorithm 2.

Algorithm 2
k-medoid $G(V, E, k)$

```
medoid set P = ({}$_1$, {}$_2$,........., {}$_{k-1}$,{}$_k$): k node sets
1. Randomly select k nodes as medoids from V and assign them to P such as one medoid per node set i.e. ({N}$_1$, {N$_2$},........., {N$_{k-1}$},{N$_k$}):
2. Calculate shortest distance of each non medoid node i w.r.t. each medoid N$_i$ and allocate a to closest medoid set.
3.  For all medoid{}$_i$:
4.          For all non medoid point b:
5.                  Update medoid N$_i$ with non medoid node b
6.                  Compute total distance C$_{li,v}$ i.e. shortest distance from b to all nodes of G.
7. Select the Medoid set N$_i$ with lowest cost
8. Repeat steps 2 to 7 till no change in medoid set P
```

2.3 Optimal Seed Selection Using Genetic Algorithm

Genetic algorithm is a well-known and widely accepted optimization technique. In the literature, various studies and researchers addressed and proposed solution for various optimization problems using genetic algorithm [16]. In this paper, we used genetic algorithm to solve influence maximization problem. For a given network $G(E, S_d, V, P)$, genetic algorithm starts searching for k optimal nodes which can improve the influence spread up to maximum extent using initial population set P, i.e., k nodes selected using k-medoid clustering algorithm. In this step, after each iteration we modified its population set P by replacing the least efficient candidate node i with lowest fitness score $\Psi(i)$ with node j, i.e., generated using one random mutation and 1-point crossover with same probability one [16]. Therefore, objective of genetic algorithm over dynamic edge score S_d and initial population set P is maximizing spread of influence by generating final population space P_d as set of k optimal seeds of the network.

3 Experiments and Results

We have performed detailed experiments on two datasets [17], i.e., Amazon co-purchasing network and wiki vote network. Details of these dataset are given in Table 2. We analyzed the performance of proposed approach with GA, i.e., embedded with various other seed selection methods, i.e., general greedy, random, discounted degree, and high degree.

Table 2 Datasets

Dataset	Details of dataset			
	Network definition	Node count	Edges count	Degree statistics
Wiki vote network	If person i voted for person j, an edge $i \rightarrow j$ created	7115	103,689	Lowest 0, highest 893
Amazon product co-purchased	If an item i purchased along with item j, an edge $i \rightarrow j$ created	5122	11,321	Lowest 0, highest 5

3.1 Evaluation Parameter

In this paper, we have applied fitness score, i.e., $\boldsymbol{\Psi}(\mathbf{P_d})$ as the performance parameter to compare the effectiveness of our proposed algorithm with other existing algorithms. The total number of non-influenced nodes converted from non-influenced to influenced state by node set P_d is known as fitness score of node set P_d, i.e., $\boldsymbol{\Psi}(\mathbf{P_d})$. Detailed description of cascade model to compute the fitness score is given in our previous work [15].

3.2 Analysis of Results

In this paper, we analyzed efficacy of the proposed approach with various existing seed selection methods, i.e., general greedy, random, discounted degree, and high degree methods embedded with GA using dynamic probabilities (GADP). We performed experiments on two datasets, i.e., Amazon product co-purchased dataset and wiki vote.

In our experiments, we used different seed values raging from 10 to 50. Figure 2a and b shows the experimental results for Amazon product co-purchased and wiki

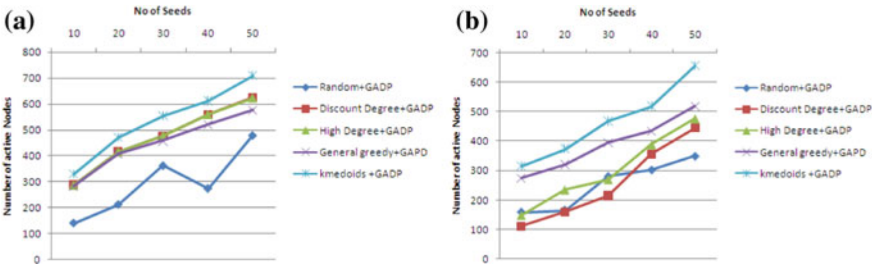

Fig. 2 Statistical results: **a** wiki vote, **b** Amazon product co-purchased

vote, respectively. It is noticed from Fig. 2a that the proposed algorithm, i.e., *k*-medoid along with GADP which is indicated by sky blue line shows the better results as compared to all other algorithms used in this paper. For wiki vote dataset, discounted degree shows the second best results because of high out-degree ratio. It is clear through the outcomes of the experiments that the proposed algorithm increased influence propagation by converting more number of nodes into influenced state from non-influenced state, i.e., up to 11%. Similar results have been observed from the Amazon product co-purchased dataset as shown in Fig. 2b. The proposed approach improved the influence propagation up to 16% for Amazon product co-purchased dataset with respect to other approaches. For this dataset, greedy approach shows the second highest fitness score because of low maximum out-degree ratio.

Overall, from the results, it can be depicted that high degree and discounted degree show the good results for the datasets of small out-degree ratio and greedy approach shows the good results for the datasets of high out-degree ratio, whereas the proposed approach presented the improved results for both types of degree ratio, i.e., low and high.

We have also performed the comparative analysis with respect to propagation value as well, i.e., fitness score of the results generated by different approaches as shown in Fig. 3. It can be easily depicted from the Fig. 3 that our proposed algorithm shows the significant improvement in fitness score as well with respect to all other approaches. Our proposed algorithm showed this improvement with respect to fitness score for both the datasets.

Overall, from experimental results we also observed an interesting behavior of our proposed algorithm w.r.t. other methods, i.e., consistency. Our proposed algorithm showed the improved results for different out-degree ratios, i.e., low and high both, whereas random, discounted degree, and high degree result are out-degree ratio dependent. Therefore, our proposed algorithm shows the improved influence propagation by 8–16% with respect to other approaches.

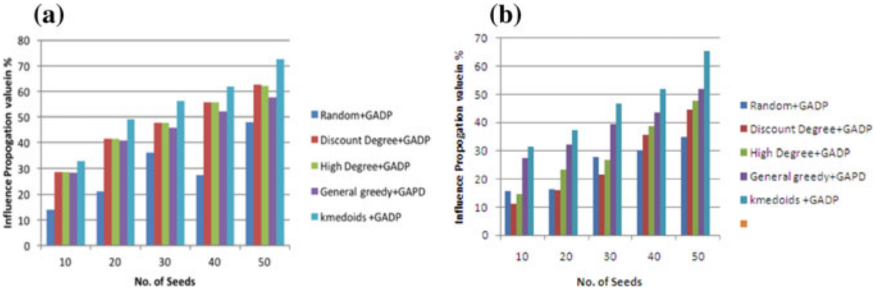

Fig. 3 Comparision of fitness score: **a** wiki vote, **b** Amazon product co-purchased

4 Conclusion and Future Work

In this paper, we proposed an influence maximization model that identifies optimal seeds that maximize the influence spread in the network. Our proposed algorithm is a hybrid approach, i.e., GA with *k*-medoid approach using dynamic edge strength. Through experiments, we analyzed efficacy of proposed approach with various existing selection methods, i.e., general greedy, random, discounted degree, and high degree methods embedded with GA and dynamic probabilities (GADP). In our experiments, we have used two types of datasets, i.e., wiki vote and Amazon product co-purchasing network with high and low out-degree ratio, respectively. Experimental results demonstrate that the proposed approach is able to achieve up to 16% improvement in influence spread with respect to other approaches. Through results, we also observed that performance of proposed algorithm is consistent, i.e., does not depend on the out-degree ratio. Therefore, our proposed algorithm maximized the influence spread by finding optimal seeds as compared to other approaches.

References

1. X. Song, Y. Chi, K. Hino, B.L. Tseng, Information flow modeling based on diffusion rate for prediction and ranking, in *WWW* (2007), pp. 191–200
2. P. Domingos, M. Richardson, Mining the network value of customers, in *KDD* (2001), pp. 57–66
3. D. Kempe, J. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network, in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM* (2003)
4. Y. Li, et al., Influence maximization on social graphs: a survey. IEEE Trans. Knowl. Data Eng. (2018)
5. C. Aslay et al., Influence maximization in online social networks, in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. ACM* (2018)
6. A. Anagnostopoulos, R. Kumar, M. Mahdian, Influence and correlation in social networks, in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM* (2008)
7. W. Chen, Y. Wang, S. Yang, Efficient influence maximization in social networks, in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM* (2009)
8. R. Mittal, M.P.S. Bhatia, Identifying prominent authors from scientific collaboration multiplex social networks, in *International Conference on Innovative Computing and Communications* (Springer, Singapore, 2019), pp. 289–296
9. W. Chen, C. Wang, Y. Wang, Scalable influence maximization for prevalent viral marketing in large-scale social networks, in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM* (2010)
10. M.M.D. Khomami et al., Minimum positive influence dominating set and its application in influence maximization: a learning automata approach. Appl. Intell. **48**(3), 570–593 (2018)
11. A. Goyal, F. Bonchi, L.V.S. Lakshmanan, A data-based approach to social influence maximization. Proc. VLDB Endowment **5**(1), 73–84 (2011)
12. W. Chen, Y. Yuan, L. Zhang, Scalable influence maximization in social networks under the linear threshold model, in *Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE*, pp. 88–97 (2010)

13. A. Kumar, S.R. Sangwan, Rumor detection using machine learning techniques on social media, in *International Conference on Innovative Computing and Communications* (Springer, Singapore, 2019), pp. 213–221
14. S. Agarwal, S. Mehta, Approximate shortest distance computing using *k*-medoids clustering. Ann Data Sci **4**(4), 547–564 (2017)
15. S. Agarwal and S. Mehta, Social influence maximization using genetic algorithm with dynamic probabilities, in *2018 Eleventh International Conference on Contemporary Computing* (*IC3*) (Noida, India, 2018), pp. 1–6
16. D. Bucur, G. Iacca, Influence maximization in social networks with genetic algorithms, in *European Conference on the Applications of Evolutionary Computation* (Springer, Cham, 2016)
17. J. Leskovec, A. Krevl, *Large Network Dataset Collection* (2015)