User Selection Mechanism for Precision Marketing based on Social Networks

Longhua Guo, Bachelor[1,2], Jie Wu, Ph.D.[2], Wei Chang, Ph.D.[3], Jun Wu, Ph.D.[1], and Jianhua Li, Ph.D.[1]

[1] Shanghai Jiao Tong University, China, staring@sjtu.edu.cn, junwuhn@sjtu.edu.cn, lijh888@sjtu.edu.cn

[2] Temple University, USA, jiewu@temple.edu

[3] Saint Joseph's University, USA, wchang@sju.edu

***Abstract*—Businesses purse improved profits with various marketing events. It's very important to utilize appropriate precision marketing methods. In many cases, a company has several products with various user groups. In a market event, it attach great significance to select right users as potential customers. Besides, right product should be selected when recommended to an individual user using corresponding marketing strategy. In this paper, we systematically models the user selection problem to address the above issues. An interest inference approach in social networks is proposed to build an economic model. According the calculated probability of each customer's interest, two mapping relationship models are built between customers, interests, and products. The potential customer-selection problem is eventually reduced to a transportation problem . Finally, experiments on real-world data show the effectiveness and accuracy of the proposed method.**

***Index Terms*—User selection, interest inference, social network, transportation problem.**

I. INTRODUCTION

Marketing events has become increasedly popular for businesses which aim to attract potential customers to participate in the consumption activities. Advertisement, offline free try and other form activities are utilized for attracting potential customers [1]. It's significant to utilize cost-effective precision marketing method for improved profits for business owners. In many cases, a company has several products towards different customer groups. As shown in the example in Fig. 1, the business has three products. There are seven users with different interests. Based on the interest, they may choose corresponding products. Obviously, the users who don't have the interest won't be chosen as potential customers in marketing. For example, the interest of user 2 is l_1 and he may buy products including b_1 and b_2. From the perspective of the company, several essential problems arise when selecting potential customers in a marketing event. Who are chosen to be the potential customers? Which product is recommended to each potential customer? In a market event, it attach a great significance to select right users as potential customers.

Social networks have been growing which digitizing and modeling real-world connection. In the case of Facebook, there are over 1.79 billion monthly active worldwide Facebook users (Facebook MAUs) which is a 16 percent increase year over year [2]. Users spend increasing time using social networks and establishing connections with friends, and these online relationships reflect and leverage the relationships in the real world [3]. As a result, businesses are increasing their attention to social networks. Utilizing the information mined and acquired from social networks, the links and attributes of the users are analyzed to effectively achieve the marketing objectives of businesses [4]. For example, location-based social networks are regarded as a valuable tool in promoting products and services with the increasing popularity of smartphones and social network services. Restaurants, hotels, and other location-based businesses utilize the location attribute of the users to maximize the benefit and profits of marketing events.

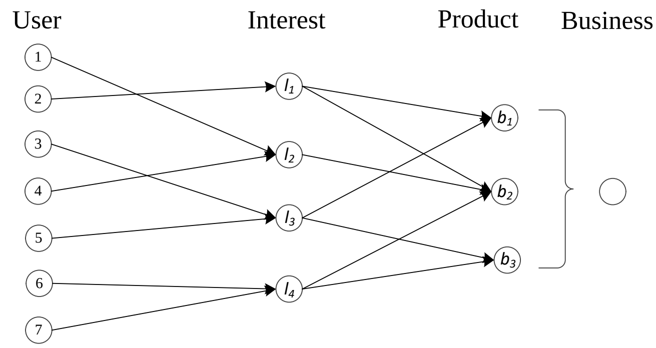

Fig. 1: Example of user selection problem.

In this paper, we proposed our method to the user selection problem utilizing the inference approach in social network. Clearly, the inference has become one of the most important parts for social network based marketing. Unfortunately, previous works have several limitations. Most works concentrate on marketing methods for companies with a single product. In social-network-based marketing, how to maximize the profits of a company with *several* products by targeting different customer groups has barely been discussed. To address the user selection problem, the interest information of each customer is inferred through social networks. According to the inferred interest, the targeted marketing strategy is utilized with marketing cost and consumer gain. An economic optimized model is built to solve the above problem.

This paper systematically models the user selection problem utilizing an interest inference approach. Our main contributions are summarized as follows.

- Two bipartite graphs are built between the customer, the interest, and the product, which model the user selection problem. Thus, the problem can eventually be reduced to transportation problem [5].

- The interests' probability distribution for each customer, which achieves an improved accuracy.
- Extensive real-data-based experiments show the effectiveness of our method.

The remainder of the paper is organized as follows. Section II describes the related work. Section III presents the user selection model and formulates the problem. In Section IV, the details of the proposed method are given including the calculation of interests' probability distribution and the reduced transportation problem. Experiments on real-world data are conducted to evaluate the proposed method in Section V. Finally, we draw our conclusion in Section VI.

II. Related Work

Social networks are utilized for businesses to achieve their marketing goals in the previous works [6]. Using viral marketing to maximize influence is widely utilized in social network marketing. [7] exploits viral marketing to maximize influence and promotion benefits. [8] proposes an efficient algorithm utilizing mathematical programming to find optimal seedings for medium-sized networks. What's more, [1] proposes a participant selection method that leverages the location-based social network in offline event marketing. However, all these previous works concentrate on the marketing method for only one company. How we can maximize the marketing benefit for one company with several products is barely studied.

Inferring a customer's hidden interests is also very important which has been widely studied in the previous works. From the perspective of the information types, the methods can be categorized as content-based and network-based methods. Content-based methods try to infer hidden interests from hints in published content based on attributes. For example, Mahmud *et al.* [9] proposed a probabilistic model to infer a Twitter user's location attribute based on his tweets. The inference result shows having 64 percent accuracy in the city-level location attribute. Network-based methods try to estimate a user's attribute value by his neighbors. Tang *et al.* [10] focus on estimating birth years. But the conjectures are limited by user location and are not suitable for other attributes and other social networks.

III. Model and Problem Formulation

Previous works concentrate on the marketing method for one company with only one product. We build our model to maximize the profits of one company with several products towards different customer groups. We first present a needed social network data model. We formulate the interest inference method and group utility maximization-based selection.

A. Social Network Data Model

In social network data models, structural data and users profiles are contained in a social network dataset. In this paper, social networks is modeled using an undirected and unweighted graph $G=\langle V, E \rangle$, in which vertex set V represents the users. The total number of users is N_v. We refer to an undirected edge for the relationships of the users. An edge $e_{uv} \in E$ represents a link corresponding to the relation between nodes u and v, where $e_{uv}=1$ means the link exists. Node u is called a one-hop neighbor of v, and v is also a one-hop neighbor of u. Attribute information of the users, such as gender or location, is also an important part of a social network. A denotes an attribute set where N_a is the quantity of attribute categories. a_v^i represents the node v's value of a_i.

We assume that the social network information, including structural information and attribute information, is known from the background knowledge.

B. User Selection Model

To model and formulate the marketing goals for one company with several products, we make following definitions:

- **Products.** The company has N_b products which are denoted as $b_i, i \in [1, N_b]$. Each product has different customer groups. At the marketing event, the goal of all products is to maximize the overall benefit of the company.
- **Interest.** The customer group is only related to interest set L which has N_l values, denoted as l_j. Pr_v^j represents the probability of customer v with an interest value l_j. According to the background knowledge, nodes whose $Pr_v^j = 1$ are called "seed nodes" which means their interests are previously known.
- **Marketing Strategy.** The marketing strategy m_v is decided based the interests of user v.
- **Cost.** The marketing cost depends on the products and the marketing strategy, which are previously known. d_j^i represents the cost of b_i for each customer whose interest is l_j.
- **Gain.** The gain is related to the interests of the customer. For the company, the gain from each customer with l_j is denoted as g_j. We assume that each customer will consume one product at most at a marketing event.
- **Quantity.** The total quantity of the users for marketing is constant and is denoted as Q. Considering the capability and balance of different products, the maximum quantity for product b_i is q_i. Only one product can be marketed to a single customer.

From the view of the business, the utility u_v of customer v is the maximal benefit in a marketing event.

$$u_v = \max(g_j * Pr_v^j - d_j^i), \forall i \in [1, N_b], j \in [1, N_l] \quad (1)$$

X_v^i represents that the product b_i is marketed to customer v. In the final decision, if v is selected as a user and recommended with b_i, $X_v^i = 1$. Otherwise, $X_v^i = 0$.

The objective is to maximize the total utility U_{all} with constraints of the selected customer quantity for each product.

For all customers,

$$U_{all} = \max \sum_{v \in V} X_v^i * (g_j * Pr_v^j - d_j^i), \forall i \in [1, N_b], j \in [1, N_l]$$
$$s.t. \sum_{v \in V} X_v^i \leq q_i, \forall i \in [1, N_b]$$
$$\sum_{i=1}^{N_b} X_v^i = Q \tag{2}$$

The optimal result is a mapping relationship from *V* to *L* and then to *B*. $\{v, l_j, b_i\}$ represent that selected customer *v* will be recommended to the product b_i utilizing the interest l_j-based marketing strategy. It is essential to calculate the probability distribution Pr_v^j based on the inference of the social network. In addition, designing the optimal algorithm is another task that maximizes the total utility.

C. Interest Inference

The background knowledge, including the structural and attribute information in social networks, plays a key role in interest inference. We define the similarity score between two nodes from the perspective of the structures and attributes in the social network. Node pairs who obtain a higher similarity score have a larger probability of having the same interests. For each node, the probability of each interest is calculated in the proposed method.

In order to calculate the probability distribution Pr_v^j, a seed-based method is utilized based on similarity. As for the seed nodes, some social network users submit their personal properties about interests in the websites. What is more, the history records of businesses are also a possible method to extract the seed nodes. $S(u, v)$ represents the similarity score between two nodes u, v. The algorithm of similarity calculation is also an important part of our work.

IV. User Selection Method

A seed based algorithm is utilized for inferring the interest. The interest distribution is previously known for seed nodes according to the background knowledge. The interest probability distribution for non-seed nodes is calculated based on seed nodes' distribution. Initial interest probability distribution matrix $\boldsymbol{P}^0$ is generated in advance. According to the update sequence, the interest probability of each node is updated based on the one-hop neighbors' similarity. The result is the input of the user-selection method. This marketing problem will be reduced to a transportation problem, which provides an approach for the optimization of user selection.

A. The initial interest probability distribution matrix

A methodology is used to obtain the initial interest probability distribution matrix $\boldsymbol{P}^0$. $\boldsymbol{P}$ is an $N_v * N_l$ matrix. $\boldsymbol{P}_v$ is a $1 * N_l$ vector which represents the interest probability distribution for node v.

$$\boldsymbol{P}_v = \{Pr_v^j | j = 1, 2, ..., N_l\} \tag{3}$$

$$\sum_{j=1}^{N_l} Pr_v^j = 1 \tag{4}$$

We check the background knowledge for L to obtain seed user set V_{seed}. For seed nodes whose Pr_v^j equals 1, the *j*-th item of vector $\boldsymbol{P}_v$ equals 1 while other items equal 0. The other nodes' interest probabilities are calculated based on the seed nodes' interest distributions.

$$\boldsymbol{P}_u = \frac{\sum_{v \in V_{seed}} \boldsymbol{P}_v}{\sum_{v \in V_{seed}} 1}, \forall u \in \{V - V_{seed}\} \tag{5}$$

B. Update sequence

Based on the nontrivial number of seed users, the second step is to update the interest probability of other nodes. r_v represents v's ratio of its one-hop seed nodes in all its one-hop neighbors. The result of the interest inference will be more accurate if r_v is higher. Based on this principle, each node's proportion is calculated. The node that has the highest r_v will update in advance. If the proportion is equal for several nodes, the node that has more communities will have a higher priority. The seed nodes will not be updated. The update sequence U is generated as shown in Algorithm 1.

$$r_v = \frac{\sum_{u \in V_{seed}} e_{uv}}{\sum_{u=1}^{N_v} e_{uv}} \tag{6}$$

Algorithm 1 Update sequence generation

Input: Prior graph G, Seed user set V_{seed}
Output: Update sequence U

```
U ← ∅, V ← V − V_seed
for Each node v in V do
    Calculate the proportion r_v
    Insert v to sequence U, sorted based on r_v
end for
return U
```

C. Process of updating

Before presenting the updating algorithm, we first define the node similarity measurement. Similarity $S(u, v)$ between two user nodes u, v includes the attribute similarity and the structural similarity score.

$$S(u, v) = \alpha S^a(u, v) + (1 - \alpha) S^b(u, v) \tag{7}$$

1) Structural similarity: The structural similarity mainly considers the connection to the relationship in the social network graph. The Jaccard Index is utilized to represent the similarities between two nodes.

$$S^b(u, v) = \frac{\sum_{i=1}^{N_v} e_{vi} * e_{ui}}{\sum_{i=1}^{N_v} (e_{vi} + e_{ui} - e_{vi} * e_{ui})} \tag{8}$$

2) Attribute similarity: Apart from the structural information, attributes are utilized to infer the probability through attribute similarity calculation. Social network users are characterized by several dimensional attributes. $Y^i(u,v)$ represents the similarity between node *u* and *v* in *i*-th attribute. The overall attribute similarity is the average of that in each attribute.

$$S^a(u,v) = \frac{1}{n}\sum_{i=1}^{n} Y^i(u,v) \tag{9}$$

For each nominal attribute like sex, $Y^i(u,v)$ follows Bernoulli distribution.

$$Y^i(u,v) = \begin{cases} 1, a_u^i = a_v^i \\ 0, a_u^i \neq a_v^i \end{cases} \tag{10}$$

3) Update algorithm: Aside from the seed nodes in the network, $\boldsymbol{P}^i$ is updated for improved accuracy as the update sequence. In the process, the interests of the one-hop neighbors are unutilized to update the probability distribution. Similarity $S(u,v)$ between two user nodes u, v is regarded as weight to update $\boldsymbol{P}$. The $\boldsymbol{P}^i$ is related to not only the one-hop neighbors, but also the $\boldsymbol{P}^{i-1}$.

$$\boldsymbol{P}_v^i = \frac{\sum_{u\in V} S(u,v) * \boldsymbol{P}_v^{i-1}}{\sum_{u\in V} S(u,v)} \tag{11}$$

Algorithm 2 Process of updating

Input: Prior graph G, Attribute information A, Probability distribution matrix P^{i-1}, Update sequence U
Output: Probability distribution matrix P^i

```
Similarity set S ← ∅
for Each node v in U do
    for Each node u in v's one-hop neighbors do
        Search S(u,v) in S
        if S(u,v) is not found then
            Calculate the similarity S(u,v)
            S(v,u) ← S(u,v)
            Add S(u,v) and S(v,u) into S
        end if
    end for
    Update P_v^i for node v
end for
Construct and return P^i
```

D. User selection algorithm

The inference result is the probability of users towards each interest. To transform the problem, we introduce a complete weighted bipartite graph. Any link is a possible candidate match. *V, L*, and *B* are regarded as the source, transfer station, and destination separately. The supply amount of each source equals 1, while the demand of the destination equals q_i. The user selection problem can be eventually reduced to a transportation problem with $N_v + V_l + V_b$ sources and destinations, and thus, can be solved by the algorithm shown in Algorithm 3.

Algorithm 3 User selection algorithm

Input: $P^i, N_v, N_b, N_l, g_j, d_j^i$
Output: The decision set $\{v, l, b\}$

```
Construct demand, supply, unit cost model based on input
Finding an initial basic feasible solution using the least-cost method
while (The solution is not optimal) do
    Iterating the algorithm through determining the entering and leaving variable
    Checking the optimality
end while
Return the decision result {v, l, b}
```

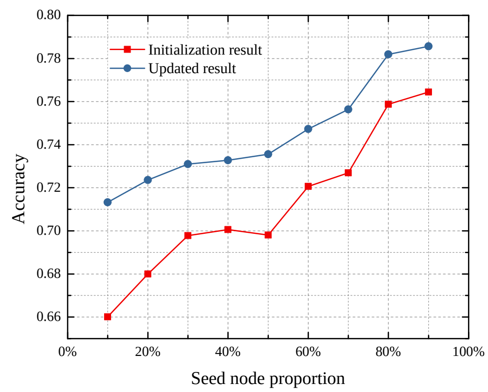

Fig. 2: The influence of seed node proportion in accuracy.

V. Experiment Evaluations

To evaluate the propose mechanism, experiments are conducted on two real world social network datasets. Then, we present a comprehensive evaluation result on our methods.

A. Methodology

We simulate our methods on two real world datasets, Facebook and Google+, both from the Stanford Network Analysis Project (SNAP) [11]. Facebook is one of the most popular social network websites and Google+ works as a social layer by Google Inc. for Google services. Rich social network data and profiles for users are contained in the two platforms. After preprocessing, there are 12 main attributes in Facebook and 6 attributes in Google+.

In the following, the evaluation and experiment results are given. First, we experiment with the real world datasets in attribute extraction and show the effectiveness of the proposed algorithm. Parts of nodes are selected as seed nodes, while other nodes' interests are removed and used as ground truth. We evaluate the influence of different seed node proportions on the interest inference accuracy. In addition, we also evaluate the influence of different *Q* on the total utility.

B. Evaluation on interest inference

The accuracy is used to measure the effectiveness of our proposed algorithm. Since the experiments on Facebook and Google+ are alike, we focus on the evaluation using the

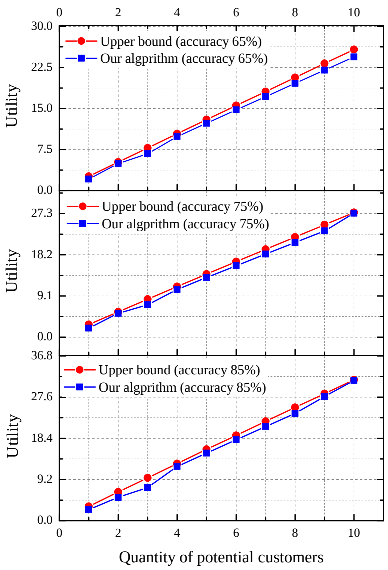

Fig. 3: The influence of different q on the total utility.

Facebook dataset [12]. We run all programs on Windows 10 on a computer with an Intel(R) Core(TM) i7 CPU and 6 GB RAM. We choose the attribute "locale" as the interest. Different proportions of nodes are chosen as seed nodes whose value of "locale" is previously known, while other nodes' values are inferred according to the proposed algorithm. The inferred results are compared with the ground truth to get the accuracy.

Fig. 2 effectively shows the influence of the seed node proportion on the accuracy. The initial result is calculated based on the initial matrix $\boldsymbol{P}^0$. The inference accuracy is improved after the updating process. With more seed nodes in all nodes, the accuracy increases both in the initialization and the update process.

C. Evaluation on user selection method

To validate our user-selection method, we run the algorithm on different interest inference accuracies. The gain g_j from each customer and the cost d_j^i are generated randomly based on uniform distribution with averages of 1 and 0.2 respectively. N_b equals 4 and N_v equals 100. The upper bound is the sum of the top Q results regardless of the limitation of q_i for each product.

We evaluate the influence of a different q on the total utility as shown in Fig. 3. The utility is related to the inference accuracy and the quantity of users. The utility grows with the increase of the accuracy and the quantity of users. In each case, the result of our algorithm is gradually closed to the upper bound with the increase of q.

The constraint condition may change in different marketing cases. As a typical linear programming problem, traditional transportation problems provide solutions in the shown method. If the quantity of each product is flexible, the problem can be solved using the simplex method.

VI. Conclusion

In this paper, users are selected based on social network. The algorithm is more realistic with respect to real-world connections. The interest probability distribution for each customer, which achieves an improved accuracy. The user selection problem is eventually reduced to a transportation problem. Our evaluations on two real-world social network databases show the effectiveness and universality of our proposed method.

VII. Acknowledgments

This work is supported in part by the National Science Foundation of China (NSFC) grants no. 61571300, 61401273, 61562004, Shanghai Science and Technology Committee Grant No. 15PJ1433800, and NSF grants CNS 1629746, CNS 1564128, CNS 1449860, CNS 1461932, CNS 1460971, CNS 1439672, CNS 1301774, and ECCS 1231461.

References

[1] Z. Yu, D. Zhang, Z. Yu, and D. Yang, "Participant selection for offline event marketing leveraging location-based social networks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 6, pp. 853–864, 2015.

[2] H. Tran and M. Shcherbakov, "Detection and prediction of users attitude based on real-time and batch sentiment analysis of facebook comments," in *International Conference on Computational Social Networks.* Springer, 2016, pp. 273–284.

[3] X. Chen, X. Gong, L. Yang, and J. Zhang, "Exploiting social tie structure for cooperative wireless networking: A social group utility maximization framework," *IEEE Transactions on Networking (TON)*, vol. 24, no. 6, pp. 3593–3606, 2016.

[4] F. Zhou, R. J. Jiao, and B. Lei, "Bilevel game-theoretic optimization for product adoption maximization incorporating social network effects," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 46, no. 8, pp. 1047–1060, 2016.

[5] U. Brenner, "A faster polynomial algorithm for the unbalanced hitchcock transportation problem," *Operations Research Letters*, vol. 36, no. 4, pp. 408–413, 2008.

[6] L. Liao, Q. Ho, J. Jiang, and E.-P. Lim, "Slr: A scalable latent role model for attribute completion and tie prediction in social networks," in *Data Engineering (ICDE), 2016 IEEE 32nd International Conference on.* IEEE, 2016, pp. 1062–1073.

[7] W.-Y. Zhu, W.-C. Peng, L.-J. Chen, K. Zheng, and X. Zhou, "Exploiting viral marketing for location promotion in location-based social networks," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 11, no. 2, p. 25, 2016.

[8] T. N. Dinh, H. Zhang, D. T. Nguyen, and M. T. Thai, "Cost-effective viral marketing for time-critical campaigns in large-scale social networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 22, no. 6, pp. 2001–2011, 2014.

[9] J. Mahmud, J. Nichols, and C. Drews, "Home location identification of twitter users," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 3, p. 47, 2014.

[10] R. Dey, C. Tang, K. Ross, and N. Saxena, "Estimating age privacy leakage in online social networks," in *INFOCOM, 2012 Proceedings IEEE.* IEEE, 2012, pp. 2836–2840.

[11] J. Leskovec and A. Krevl, "Snap datasets: Stanford large network dataset collection, june 2014," *URL: http://snap. stanford. edu/data*, 2014.

[12] J. Qian, X.-Y. Li, C. Zhang, and L. Chen, "De-anonymizing social networks and inferring private attributes using knowledge graphs," in *Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on.* IEEE, 2016, pp. 1–9.