A PageRank-Based Heuristic Algorithm for Influence Maximization in the Social Network

Zhi-Lin Luo, Wan-Dong Cai, Yong-Jun Li, and Dong Peng

Abstract. The influence maximization is the problem of how to find a small subset of nodes (seed nodes) that could maximize the spread of influence in social network.However,it proved to be NP-hard.We propose a new heuristic algorithm, the High-PageRank greedy algorithm (HPR_Greedy),which searches the seed nodes in a small portion containing only the high-PageRank nodes, based on the power-law influence distribution in non-uniform networks. The experimental results showed that, compared with classical algorithms, the HPR_Greedy algorithm reduced search time and achieved better scalability without losing influence.

1 Introduction

With the development of WEB2.0, online social networking has become a popular way to share and disseminate information over Internet. According to the Nielsen Company's reports[1], up to two-thirds of the global online population visited social networks and blogs. Social network has become a fundamental part of the online society. Some large-scale social net-working sites are very popular, e.g. Facebook has become the largest website in the US, surpassing Google. Social networking sites provide huge potential business opportunities for companies to market their products. How to efficiently find a small subset of nodes (seed nodes) with the greatest influence in large-scale social networks is the focus of this study.

How to find the seed nodes with the greatest influence is the influence maximization problem, which is of great interest to many companies or people who want to market their products or service through Internet . When studying the viral-style marketing, Domingos and Richardson et al. [3,4]are the first to discuss it.Kempe et al. [5] formally proposed the influence maximization problem as the discrete optimization and proved that it is NP-hard. They presented the original greedy approximation algorithm.Leskovec et al.[6] presented the greedy with CELF

Zhi-Lin Luo · Wan-Dong Cai · Yong-Jun Li · Dong Peng
School of Compute Science Northwestern Polytechnic University, Xi'an China
e-mail: {lzluo007,justastriver}@gmail.com,
{caiwd,lyj}@nwpu.edu.cn

F.L. Gaol (Ed.): Recent Progress in DEIT, Vol. 2, LNEE 157, pp. 485–490.
springerlink.com

(Cost-Effective Lazy Forward selection) optimization algorithm based on the submodularity function theory. Their results showed that the influence of the Greedy with CELF was nearly close to those of the original Greedy , while achieved as much as 700 times speed.Chen Wei et al.[8] proposed the NewGreedy and the MixGreedy algorithms. MixGreedy improves the efficiency.But both algorithms' scalability is poor in Linear threshold model.

2 Information Dissemination Models

There are three basic information dissemination models.

Independent cascade model (IC)

The network is modeled as an undirected graph.Each node has two states: active or inactive.nodes are active if they have accepted information, inactive otherwise. Nodes only can switch from being inactive to being active. Active nodes activate inactive nodes with a constant influence factor p ($0<p<1$).

Weighted cascade model (WC)

WC is very similar to IC. The difference is that a social network is modeled as an directed graph in WC model, with asymmetric influence factors, The influence factor of a node v is $1/dv$, where dv is the degree of the node v.

Linear threshold model (LT)

LT is different from the two models above. A node v is randomly assigned a threshold and influenced by each neighbor w with a weight.The condition for the node v to be activated is that the total weight of its active neighbors is greater than threshold

3 The Algorithms for Influence Maximization

In this section, we will first introduce the classical greedy with CELF optimization, and then propose a new heuristic algorithm. Before describing the algorithms, let's define $\sigma(\bullet)$,S, and U as the influence function, the set of the seed nodes and the set of all nodes, respectively.

Theorem 1. *THEOREM 1 (Nemhauser et al.1978)*

$$\forall A \subseteq B \subseteq N, \forall j \in N \backslash B, if\ \mathrm{f}(A+j) - \mathrm{f}(A) \geq \mathrm{f}(B+j) - \mathrm{f}(B) \tag{1}$$

f *is submodular function.*

So submodular function is non-negative and monotone.

Theorem 2. *(Kempe et al.2005)For an arbitrary instance of the IC or WC or LC model, the resulting influence function* $\sigma(\bullet)$ *is submodular.*

Based on theorems 2, It can be deduced that the influence function $\sigma(\bullet)$ of each node in the set U gradually weakens with the number of the set S increasing.

The classical greedy with CELF can be divided into two different rounds: in the first round it compute the influence of each node in the set U and select the node