CS303 Artifitial Intelligence 2024F Project1 Report

Ben CHEN
chenb2022@mail.sustech.edu.cn

October 6, 2024

1 Introduction

1.1 Problem Description

In the era of information explosion, though the population of the world receives mountains of information from media every day, the Echo Chamber is formed thereby where a bunch of people only receive information from their groups. In order to break the Echo Chamber, we need to maximize the expected number of people who either receives from every groups or remains isolated. So in this project, we solve the Information Exposure Maximization problem (IEMP) by modeling the problem with a graph with two groups of nodes to simulate the information propagation from both campaigns.

The objective of this project is that, given a DAG with probabilities of passing information from one node to its neighbors as weights on edges, and two groups of nodes as the initial campaigns, we need to find two subsets of nodes for each campaign to maximize the expected number of nodes that either receives from both campaigns or remains oblivious.

1.2 Purpose

To solve the IEM problem, we divide the project into three steps:

1. Model the problem with a formal definition and a graph representation in Python.
2. Evaluate the size of the maximum number of nodes that either receives from both campaigns or remains oblivious from a given solution. It's not a trivial problem since a exact solution cannot be found in polynomial time.
3. Design a heuristic algorithm and an evolutionary algorithm to optimize the solution. Both algorithms should be efficient and effective.

In this report, we will formalize the problem in Section 2, introduce the pseudocode of our algorithm in Section 3, present the experiment design in Section 4, and analyze the results in Section Section 5. Since the source code might be hard to read, we will describe the algorithm in words in this report.

2 Preliminary

This section formally defines the problem with notations, models and formulates the result we want to achieve.

2.1 Notations

According to the problem description, we define the following terminologies:

Social Networks A DAG $G = (V, E)$ where V is the set of nodes and E is the set of edges. Each edge $(u, v) \in E$ has a weight $p_t(u, v) \in [0, 1]$ representing the probability of information propagated from node u to node v. Each edge has a tag $t = \{1, 2\}$ representing information from campaign 1 or 2.

Campaigns Two identical groups of nodes $C_1, C_2 \subseteq V$ representing the campaigns that would spread their opinion. Each campaign contains two sets $C_i = I_i \cup S_i$ where $i = 1, 2$.

Initial Seed Set Two subsets of nodes $I_1, I_2 \subseteq V$ representing the initial seed set of campaigns.

Balanced Seed Set Two subsets of nodes $S_1, S_2 \subseteq V$ representing the seed set of campaigns that we need to find to maximize the expected number of nodes that either receives from both campaigns or remains oblivious.

Budget The number of nodes that we can select. $|S_1| + |S_2| \leq k$.

Influence Result A subset of nodes $r(U) \subseteq V$ representing the nodes influenced by the seed set $U \subseteq V$. However, we need to find the mathematical expectation rather than the size of nodes.

2.2 Diffusion Model

We apply the Independent Cascade Model to simulate the information propagation in the social network. The model is defined as follows:

- Each nodes $v \in V$ has a state *active* or *inactive* representing whether the node has received the information.
- **Active** nodes receive information from their neighbors with and can activate their neighbors. Active nodes won't be changed to inactive.
- **Inactive** nodes have been attempted to be activated by their neighbors but failed. We don't consider the nodes that never be attempted to be reached, i.e., the nodes that have no active neighbors.

$r(U)$ is the set contains active and inactive nodes mentioned above.

2.3 Result

Given a social network $G = (V, E)$, two initial seed sets $I_1, I_2 \subseteq V$, and a budget k, we need to find two balanced seed sets $S_1, S_2 \subseteq V$ with $|S_1| + |S_2| \leq k$ to maximize the expected number of nodes that either receives from both campaigns or remains oblivious

$$\max \Phi\left(S_1, S_2\right) = \max \mathbb{E}\left[\left|V \backslash (r_1(I_1 \cup S_1) \Delta r_2(I_2 \cup S_2))\right|\right]$$

$$s.t. \quad |S_1| + |S_2| \leq k,\ S_1, S_2 \subseteq V$$

The expectation is calculated by the sum of probabilities times size of nodes calculated above. Since we need to decide the activation if the probability is neither 0 nor 1, the choices are $2^{|V|}$.

3 Methodology

3.1 Evaluation Algorithm

Since computing the balanced information exposure for a given solution is NP-hard, we cannot directly iterate all the possible nodes to caculate the expected number of nodes. Instead, we use a Monte Carlo simulation to estimate the expected number of nodes. The algorithm is described as follows:

3.2 Heuristic Algorithm

3.3 Evolutionary Algorithm

4 Experiments

4.1 Setup

4.2 Results

5 Analysis

6 Conclusion