Proposal of CS54701
Information Retrieval Project

Wei-Yen Day
wday@purdue.edu
Department of Computer Science
Purdue University

September 24, 2012

1 Introduction

asd

asd asdf

Expertise search has been a special and famous topic in information retrieval related field. In modern IR tasks, many researchers utilize the auxiliary information, such as personal webpage data or profiles in social network. In this project, $S_n = \sum_{i=1}^{n} i^2$ we look at the expertise search in academic field with the auxiliary information from social network. More specifically, we will tackle a task: given a query which contains several keywords, can we find the most suitable and influential person to be your academic collaborators or advisors? This task is a practical problem, which involves several points of view such as conventional IR task and the information extracted from social network. We'll look at this task in detail in the following paragraphs.

Recent research works on expertise search aim at finding the most important objects by analyzing social network information. For example, IBM proposed a expertise search system called 'SmallBlue' by analyzing explicit and implicit data between people [3]. These explicit sources are public, such as co-authored documents, users' blogs, or social tagging websites. For implicit sources, such as email communication logs, can hardly be acquired because they are viewed as personal private data. On the other hand, in [2] Li et al. proposed a temporal random walk model to help search a specific person. In this project, we will focus on the social network part: how the network structures help academic expertise search and retrieval, and how to utilize them to develop a simple and efficient academic expertise search framework.

As mentioned above, the goal is given a set of queries, we want to find the most suitable and influential person for these queries. The result would be a ranking list for each query. To develop this system, there will be three main stages in this project:

- Conventional IR task: use indexing and retrieval technique to organize author information from data
- PageRank [4]: utilize the node information from author-author network
- More quality PageRank: utilize the node information from weighted author-author network

We'll discuss the details in the following sections.

2 Algorithms, Models, and Proposed Plan

2.1 Conventional IR task

The first stage is to develop an IR system. The Lemur toolkit[1], as we have seen in class, will be used for the main indexing and retrieval task in this stage. Furthermore, we'll investigate how different models, such as language model or okapi, influence the result of the academic IR system.

2.2 PageRank and Co-author Network

We plan to perform PageRank on co-author network. PageRank [4] proposed by Page and Brin is a useful and effective algorithm to compute the weights of the nodes in a graph. It is a random walk model based on Markov process. To simply describe, the PageRank algorithm is solving a linear system as below:

$$(\mathbf{I} - \alpha\mathbf{P})\mathbf{x} = \mathbf{e_i}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{P}$ is the stochastic transition matrix computed from adjacency matrix, $\mathbf{x}$ is the initial state (probability) of each node, and $\mathbf{e_i}$ is a vector with all elements 1. α is the probability to follow the link started from current node. By solving such linear system, the PageRank score of each node can be computed very fast and used to estimate the weight of such node in the network. There is another similar node weight estimation method called Katz index [1]. We can also investigate this method as a comparison of generating node weights.

2.3 Weighted co-author network

To further utilize the information from social network, we plan to investigate how to construct a weighted co-author network. There are several ways to construct a weighted graph. The heuristic method is to define a similarity score between each node pair and compute its similarity as a weight. In this project, we'll investigate how to generate a useful weighted co-author network, rather than the binary weight of original network.

References

[1] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, March 1953.

[2] Y. Li and J. Tang. Expertise search in a time-varying social network. In *WAIM*, pages 293–300, 2008.

[3] C.-Y. Lin, N. Cao, S. Liu, S. Papadimitriou, J. Sun, and X. Yan. Smallblue: Social network analysis for expertise search and collective intelligence. In *ICDE*, pages 1483–1486, 2009.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[1] The Lemur project, http://www.lemurproject.org/