Chapter XXIII
Web Mining to Identify People of Similar Background

Quanzhi Li
Avaya, Inc., USA

Yi-fang Brook Wu
New Jersey Institute of Technology, USA

ABSTRACT

This chapter presents a new approach of mining the Web to identify people of similar background. To find similar people from the Web for a given person, two major research issues are person representation and matching persons. In this chapter, a person representation method which uses a person’s personal Web site to represent this person’s background is proposed. Based on this person representation method, the main proposed algorithm integrates textual content and hyperlink information of all the Web pages belonging to a personal Web site to represent a person and match persons. Other algorithms are also explored and compared to the main proposed algorithm. The evaluation methods and experimental results are presented.

INTRODUCTION

One of the popular Web search needs is person/people related search: users like to search information related to a specific person or people who are specialized in a subject; they also want to find other people possessing certain interest/expertise or sharing similar interests or background with them from the Web.

Previous studies on person/people related search mainly focus on two directions: 1. “person search” — searching Web pages authored by a specific person or containing information about this person, given
this person’s name as the query, and 2. “people search” — finding a list of people similar to the given one, in terms of their interests. In this study, the two concepts “person search” and “people search” are differentiated. The focus of this study is the latter.

On the Web, to find other people having similar interests, the simplest way is to browse through who’s who directories or other similar directories. The problems with this method are that such directories might not be updated regularly, and the scope may be limited to only certain popular domains. Many users utilize regular search engines to find people by sending keywords to search engines and then browsing through the results to see who authored the Web pages of their interest. However, regular search engines are not specialized for the task of finding similar people; users using this approach would find it laborious and ineffective. Other existing methods, such as the online dating systems and social matching systems (discussed in the Previous Studies section), also have various limitations. They usually require a lot of user involvement and efforts. For example, online dating systems need to build user profiles by getting users to answer a long list of questions on topics such as their religious beliefs, professions, physical appearance, etc. Some other systems need users’ browsing history to build their profiles. Furthermore, these systems are only available to the registered users and usually the searchable people in their database are limited to certain groups or domains.

This study attempts to find a people search solution that requires no manual user involvement in building searchable people profiles, is able to search people from various domains, and has access to a large body of people. The people search method proposed in this study is about specifying characteristics of a person automatically and finding other persons who share the similar characteristics with the given person. To design such a system, two major research issues need to be solved: how to represent a person and how to match persons on the Web. The first question is how to profile users - what type of information does a system use to represent its users, and how does it acquire this information? The second question is how to find matches - what is the system’s model of a good match? And how does the system compute matches (Terveen & McDonald, 2005)? To address these problems, in this study, an approach to representing a person online is first proposed. Based on the proposed person representation approach, a main algorithm, called People-Search, is proposed, and 13 other algorithms are also explored and compared to the main algorithm. Prototype systems were developed to test the feasibility of the proposed solution and to evaluate the effectiveness of these algorithms.

The remainder of this chapter is organized in the following manner. Related previous studies are first presented. Then the proposed person representation method and the 14 algorithms are described, followed by the evaluation method and experimental results. Finally, the future research directions are discussed.

PREVIOUS STUDIES

In this section, we introduce previous studies on person search and people search.

Previous studies of person search mainly focus on how to find Web pages related to a person, given this person’s name as the query. In their systems, the query, which is usually a person’s name, is sent to regular search engines, and the search results from the regular search engines are then refined to find Web pages related to this person.

Wan et al (2005) describes a person resolution system, called WebHawk. They claim that given a list of pages obtained by submitting a person’s name to a search engine, their system can cluster these