Chapter 4
Expertise Learning and Identification with Information Retrieval Frameworks

Neil Rubens
University of Electro-Communications, Japan

Dain Kaplan
Tokyo Institute of Technology, Japan

Toshio Okamoto
University of Electro-Communications, Japan

ABSTRACT

In today’s knowledge-based economy, having proper expertise is crucial in resolving many tasks. Expertise Finding (EF) is the area of research concerned with matching available experts to given tasks. A standard approach is to input a task description/proposal/paper into an EF system and receive recommended experts as output. Mostly, EF systems operate either via a content-based approach, which uses the text of the input as well as the text of the available experts’ profiles to determine a match, and structure-based approaches, which use the inherent relationship between experts, affiliations, papers, etc. The underlying data representation is fundamentally different, which makes the methods mutually incompatible. However, previous work (Watanabe et al., 2005a) achieved good results by converting content-based data to a structure-representation and using a structure-based approach. The authors posit that the reverse may also hold merit, namely, a content-based approach leveraging structure-based data converted to a content-based representation. This paper compares the authors’ idea to a content only-based approach, demonstrating that their method yields substantially better performance, and thereby substantiating their claim.
INTRODUCTION

In today’s knowledge-based economy, having the proper expertise is crucial to resolving many tasks. In the pedagogical world, such tasks range from educating others, to solving difficult problems, assessing/guiding the research directions of others, and reviewing the quality of conference/journal papers. The most traditional approach to expertise finding has been a burdensome process, involving manual referrals and direct contact. Luckily computers have mitigated this burden to a considerable degree. Several excellent surveys exist concerning this (e.g., Yimam-Seid & Kobsa, 2003; Maybury, 2006). As a result of the aid of computers, expert finding systems (EFS) have started to gain acceptance and are being deployed in a variety of areas. The Taiwanese National Science Council utilizes EFS to find reviewers for grant proposals (Yang et al., 2009); Australia’s Department of Defense has deployed a prototype EFS to better utilize and manage its human resources (Prekop, 2007); ResearchScorecard Inc.’s EFS allows a user to find and rank scientists involved in biomedical research at Stanford University and at the University of California in San Francisco. There is also several expertise finding platforms that are applicable to wider domains and are utilized by an increasing number of companies (Maybury, 2006). Further, many methods have been developed to automate the task of expertise finding, including language and topic modeling (Yang et al., 2009), latent semantic indexing (LSI) (Lochbaum & Streeter, 1989), probabilistic modeling (Balog et al., 2006), and link analysis (Karimzadehgan et al., 2009; Halonen et al., 2010).

Essentially, the following two approaches are almost always employed in EFS. Typically, a description of a task is given, and the system aims at finding a person with the appropriate expertise to match it.

- Content-based approaches analyze the content (text) of both a given task’s description and of candidates’ papers (see Figure 1). For example, a candidate (researcher) may be knowledgeable about the given task (and therefore a good choice for the task) if the task description and the candidate’s works share many of the same terms (keyword extraction).
- Structure-based approaches analyze the topographical structure of the search space by treating papers as nodes in a graph, interlinked by citations, authorship, affiliations, etc. (see Figure 2). For example, a researcher may be considered a match for a given task if both the task and researcher cite the same papers (showing familiarity with the subject area).

There are many pros and cons to each approach. For one, while the underlying data of structure-based approaches is usually rather precise, content-based approaches must grapple with the inherent ambiguity and complexity of the data expressed in natural language. One way to circumnavigate this problem is to reduce the number of possible dimensions by extracting salient keywords (Laender et al., 2002); these methods can perform quite well, but are nevertheless hindered by their ability to extract appropriate and completely representative keywords. To get beyond this, intricate knowledge of natural language processing (NLP) is required a field which often lacks usable off-the-shelf tools. On the other hand, many graph analysis tools exist (ideal for structure-based approaches) (Wikipedia, 2009a). Availability of frameworks is also important. Many powerful information retrieval (IR) frameworks exist that provide easy deployment and scale well (luc, lem, ter); this is something that structure-based approaches often lack.

The bottom line is that the majority of methods to date use one approach (content-based, “C”) or the other (structure-based, “S”); though sometimes both approaches are used in tandem (C and S). Yet previous work (Watanabe et al., 2005b) has demon-