Towards an application-specific expert finding repository on the Social Web

Tom Tourwé and Elena Tsiporkova

Software Engineering & ICT Group, Sirris
{tom.tourwe,elena.tsiporkova}@sirris.be

1 Introduction

In this position paper, we advocate the need for developing an application-specific expert finding platform which extracts and unifies user-related information from a variety of online sources and subsequently builds a repository of expert user profiles in an incremental fashion. Such profiles can be used in many different applications, e.g. the identification of experts in a particular technological domain (for the purpose of technology scouting), the matching of partners for research proposals, or the visualisation of research activities and experts within geographical regions (technology brokerage).

With the rise of the social web and the advent of linked data initiatives a growing amount of data is becoming publicly available: people communicate on social networks, blogs and discussion forums, research events publish their online programmes including article abstracts and authors, websites of technological conferences and exhibitions advertise new products and technologies, governments and commercial organisations publish data, and the linked open data cloud keeps on growing. An enormous potential exists for exploiting this data by combining it and extracting intelligence.

As an illustration of such potential, consider a pharmaceutical company that needs to make important decisions related to the planing of large-scale clinical trials world-wide. Considering the enormous cost of such trials and the impact their results could have on the future and profitability of the whole enterprise, the importance of performing optimal planning is essential. In order to achieve this, thorough knowledge of the different clinical researchers world-wide active in the disease targeted by the planned clinical trials is essential. This knowledge includes current research results in the field, present affiliation, size of the lab, number of patients with the disease in question treated per year, etc.

Although a large pool of the data required for applications as described above is available on the web, it is often still gathered manually, a time-intensive, tedious and error-prone process due to the fact that the data is not centralised, is available in different formats, can be outdated or contradictory, etc. Most applications that automatically gather user data from the web serve personalisation or recommendation purposes. These differ significantly from the clinical trial planning application described above, which imposes more strict requirements: 1) the data needs to be up-to-date at all times; 2) high accuracy/reliability of the data needs to be guaranteed; 3) very high (if not complete) level of coverage over the domain should be attained; 4) it should be possible to somehow rank the experts in terms of impact and relevance.

2 Proposed Approach

The problem of finding experts is not new, but is usually tackled top-down by mining vast online data sources as e.g. Wikipedia and CiteSeer, in order to gather a sufficiently large data repository containing information about persons, articles, social links, etc. This has the advantage of achieving high coverage of the experts active in a certain domain and a relatively complete expert profile in space and time. However, there are certain shortcomings associated with this approach as for example such large data collections contain a substantial proportion of noisy data (contradictions, duplicates, ...).

We propose to approach the problem of finding experts bottom-up by first identifying online sources targeted to the application domain in question. These serve as seeds for the further incremental growth of our expert repository. Specifically, we believe the following types of information will be most relevant to consider:

- **Actors in the field**: leading researchers and experts, as well as the companies, research institutes, or universities they are associated to, ...
- **Technology-related publications**: patents, scientific and popularised publications, presentations and keynotes, press releases, technology blogs, ...
- **Research activities**: past and ongoing research projects, academic and scientific conferences, reviewing boards, ...
- **Career trends and evolution**: new research directions, job transitions, ...

This information can be gathered from a multitude of sources, both domain-dependent and -independent, such as dedicated conference websites, digital libraries such as ACM, IEEE, DBLP and PubliMed, professional social networks such as LinkedIn and Xing, or other sources gathering scientific output such as Google Scholar, Scientific Commons and Microsoft Academic Search. It is also important to identify and make explicit the relationships between the different data entries, e.g.

- **Formal and informal expert networks**, identified through joint publications, joint organisation of events, social networks, professional and educational history, ...
- **Strategic alliances between organisations**, identified through joint participation in research projects, preferred partnerships advocated on company website or brochures, exchange of personnel.

3 Scientific and Technological Challenges

The proposed approach induces several scientific and technological challenges. In order to guarantee high coverage one needs to consider **information extraction from multiple heterogeneous data sources**: structured (LinkedIn, Twitter), semi-structured (DBLP, ACM DL) and unstructured (web pages) online sources. High accuracy and reliability requires 1) the **development and application of advanced disambiguation techniques**, because the gathered data is noisy, e.g. it contains different spellings of author names and affiliations in different papers and different nick names used throughout different social platforms; 2) to **qualify the different sources in terms of reliability and trustworthiness** of the information they offer, e.g. distinguish between doubtful sources (such as weblogs) and reputable sources (such as digital libraries). Keeping the data up-to-date requires a **data streaming pipeline that continuously presents newly gathered data** to approve new, or revoke previously taken decisions (e.g. for disambiguation or source qualification). It is also crucial to identify adequate criteria and metrics, which most probably will be application- and problem-dependent, allowing to perform some multi-criteria decision analysis in order to **arrive at some representative expert ranking in terms of relevance and impact**.

These topics are currently considered by the research community, but are still far from reaching the required level of maturity for deployment in critical decision support applications such as described above.