People Search in the Enterprise

Krisztian Balog
ISLA, University of Amsterdam
K.Balog@uva.nl
http://staff.science.uva.nl/∼kbalog/phd-thesis/

The enormous increase in recent years in the amount of information available online has led to a renewed interest in a broad range of IR-related areas that go beyond plain document retrieval. Some of this new attention has fallen on a subset of IR tasks, in particular on *entity retrieval* tasks. This emerging area differs from traditional document retrieval in a number of ways. Entities are not represented directly (as retrievable units such as documents), and we need to identify them "indirectly" through occurrences in documents. This brings new, exciting challenges to the fields of Information Retrieval and Information Extraction. In this thesis we focus on one particular type of entity: *people*.

In an enterprise setting, a key criterion by which people are selected and characterized is their level of expertise with respect to some topic. Finding the right person in an organization with the appropriate skills and knowledge is often crucial to the success of projects being undertaken.

The work described in this thesis focuses exclusively on core algorithms for two information access tasks: expert finding and expert profiling. The goal of *expert finding* is to identify a list of people who are knowledgeable about a given topic (*"Who are the experts on topic X?"*). This task is usually addressed by uncovering associations between people and topics; commonly, a co-occurrence of the name of a person with topics is assumed to be evidence of expertise of the person on the topic. An alternative task, using the same idea of people-topic associations, is *expert profiling*, where the task is to return a list of topics that a person is knowledgeable about (*"What topics does person Y know about?"*).

The main contribution of the thesis is a generative probabilistic modeling framework for capturing the expert finding and profiling tasks in a uniform way. On top of this general framework two main families of models are introduced, by adapting generative language modeling techniques for document retrieval in a transparent and theoretically sound way.

Throughout the thesis we extensively evaluate and compare these baseline models across different organizational settings, and perform a systematic exploration and analysis of the experimental results obtained. We show that our baseline models are robust and deliver very competitive performance.

Through a series of examples we demonstrate that our generic models are able to incorporate and exploit special characteristics and features of test collections and/or the organizational settings that they represent.

We provide further examples that illustrate the generic nature of our baseline models and apply them to finding associations between topics and entities other than people.