ECIR 2008 Tutorial

Advanced language modeling approaches (Case study: Expert search)

Djoerd Hiemstra
University of Twente
http://www.cs.utwente.nl/~hiemstra

March 30, 2008

1 Introduction

This year, it is about 10 years ago that statistical language models were first successfully applied to Information Retrieval (IR). Today, language models are used as standard tools for developing new applications of IR. One of the reasons for their success is that on the one hand, the models can be surprisingly simple generative models, similar to well-known urn models. (i.e., as simple as calculating the probability of drawing a colored ball from an urn.) On the other hand, however, they come with a powerful set of modeling tools: such as smoothing, graphical models, estimation of (document) priors, and maximum likelihood training of unobserved variables. These and other advanced language modeling approaches will be explained in detail in this half day tutorial, taking *expert search* as a case study. Expert search is a novel, relatively easy to understand IR problem that is well-suited for explaining language modeling assumptions. Instead of finding documents, the primary goal of an expert search system is to find individuals in an organization that possess certain expertise and skills. There are three entities to model in expert search: 1) the experts, 2) the documents, and 3) the terms, and therefore several possible (conditional) independence assumptions to make: For instance, we might assume that experts and terms occur independently given a document, we might assume non-uniform expert priors, we might assume that documents are actually mixtures of (unknown) expert language models, we might add additional expert–expert dependencies because two experts are in the same department, etc.

2 Goal and outcome

The tutorial's main goal is to give the participants a clear and detailed overview of advanced language modeling approaches and tools. At the end of the tutorial, they can apply these approaches and tools to solve new IR problems in general, and specifically problems related to expert search. By attending the tutorial, attendants will:

- aquire intuitive understanding of the difficulty of IR, and the need for models of IR to solve IR problems;
- be able to name differences between language models and other probabilistic models (the Robertson/Spärck-Jones model and, Google's Pagerank model) and to choose a model that is well-suited for a particular task;
- be able to use graphical models as a tool to visualize complex conditional independence assumptions;
- be able to apply advanced IR language modeling approaches such as document priors, expectation maximisation training, translation models, relevance models and parsimonious models to novel application areas of IR, and specifically to expert search.

3 Course content

The tutorial consists of the following content.

1a. **Motivation**: In this part I will answer basic questions such as: What is a model? Why is IR hard? Why is modeling needed? What kind of applications are we modeling? I will address a broad spectrum of applications, from ad hoc navigational queries, to for instance email spam filtering and expert search.

1b. **Probability theory and probabilistic modeling**: In this part I will quickly introduce some notations (however, basic knowledge of probability theory is a prerequisite for the course). The basics of three well-known probabilistic models for IR will be explained: 1) The Robertson/Spärck-Jones probabilistic model, 2) Google's Pagerank model, and 3) Language models.

2. **Extensions of the basic language modeling approach**: In this part I will discuss in detail the use of document priors, translation models, and relevance models, using graphical modeling as a tool to visualize differences between models.

3. **Language modeling approaches that need model training**: In this part I will discuss parsimonious language models and probabilistic latent semantic indexing. Expectation maximization training will be explained in detail.

4. **Case study, expert search**: In the case study I will discuss the difference between profile-centric and document-centric approaches to expert search, and its consequences for language modeling. I will start with the most simple approaches to expert search that could possibly work, and then gradually add and/or change modeling assumptions based on the tools and models introduced in Part 2 and Part 3 described above. Finally, I will show that a language modeling approach can be combined with pagerank's probabilistic random walk to find the important experts in an expert–document graph. The tutorial will contain evaluation results of the approaches on the TREC enterprise search data.

4 Course material

Handouts of slides, and a detailed bibliography will be available for the participants of the tutorial. If needed, for instance based on discussions on site, additional information will be available on the web.

5 Biography

Djoerd Hiemstra wrote his Ph.D. thesis on language models for information retrieval. He contributed to over 80 research papers in the field of IR. Djoerd is assistant professor at the University of Twente for which he teaches several master courses, including *Information Retrieval* and *XML and Databases.* Djoerd gave lectures on *Formal models of IR* at two editions of the European Summer School on Information Retrieval (ESSIR). He is involved in several advanced SIKS courses for Dutch Ph.D. students (SIKS is an interuniversity research school that comprises 12 research groups in which currently nearly 400 researchers are active, including over 190 Ph.D. students.) Djoerd was involved in the local organization of SIGIR 2007 in Amsterdam, and in the organization of several workshops including editions of the Dutch-Belgian Information Retrieval Workshop series.