Introduction to Probabilistic Models in IR

Victor P. Lavrenko
University of Edinburgh, School of Informatics, Edinburgh, U.K., v.lavrenko@gmail.com

Abstract:

Most of today's state-of-the-art retrieval models, including BM25 and language modeling, are grounded in probabilistic principles. Having a working understanding of these principles can help researchers understand existing retrieval models better and also provide industrial practitioners with an understanding of how such models can be applied to real world problems.

This half-day tutorial will cover the fundamentals of two dominant probabilistic frameworks for Information Retrieval: the classical probabilistic model and the language modeling approach. The elements of the classical framework will include the probability ranking principle, the binary independence model, the 2-Poisson model, and the widely used BM25 model. Within language modeling framework, we will discuss various distributional assumptions and smoothing techniques. Special attention will be devoted to the event spaces and independence assumptions underlying each approach. The tutorial will outline several techniques for modeling term dependence and addressing vocabulary mismatch. We will also survey applications of probabilistic models in the domains of cross-language and multimedia retrieval. The tutorial will conclude by suggesting a set of open problems in probabilistic models of IR.

Attendees should have a basic familiarity with probability and statistics. A brief refresher of basic concepts, including random variables, event spaces, conditional probabilities, and independence will be given at the beginning of the tutorial. In addition to slides, some hands on exercises and examples will be used throughout the tutorial.

ACM Categories & Descriptors:

H.3.3 Information Search and Retrieval; Retrieval Models

General Terms:

Algorithms, Experimentation, Theory

Keywords:

Probability Ranking Principle, Language Modeling, Estimation, Term Dependence, Cross-language Retrieval

Bios

Victor Lavrenko is a Lecturer in Informatics at the University of Edinburgh. He received his Ph.D. in Computer Science from the University of Massachusetts Amherst in 2004, and worked as a language technology consultant for the Credit Suisse Group prior to his appointment at Edinburgh. Victor presented tutorials on language modeling and probabilistic approaches to IR at SIGIR 2003 and SIGIR 2009. He has published research papers in and has reviewed for the SIGIR, CIKM, NAACL/HLT, KDD and NIPS conferences. Victor's research interests include formal models for searching text in multiple languages, annotating and retrieving images, and detecting and tracking novel events in the news. More information: http://homepages.inf.ed.ac.uk/vlavrenk

Copyright is held by the author/owner(s).
SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.