Keynote Talk

Search Engine Support for Software Applications

Jamie Callan
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
callan@cs.cmu.edu

Abstract

Question-answering, computer-assisted language learning, text mining, and other software applications that use a full-search engine to find information in a large text corpus are becoming common. A software application may use metadata and text annotations to reduce the mismatch between the concept-based representations convenient for inference and the word-based representations typically used for text retrieval. Software applications may also be able to specify detailed requirements that retrieved passages must satisfy. This use of text search is very different than the ad-hoc, interactive search that information retrieval research typically studies.

Search engine developers are beginning to respond by extending indexing and retrieval models developed for structured (e.g., XML) documents to support multiple representations of document content, text annotations, metadata, and relationships. These new requirements force developers to reconsider basic assumptions about index data structures and ranked retrieval models.

How best to use these new capabilities is an open problem. Straightforward transformation of a detailed information need into a complex structured query can produce a query that is effective for exact-match retrieval, but a challenge for the retrieval model to use effectively for best-match retrieval. Bag-of-words retrieval is often disparaged, but its advantage is that it is robust: It works well even when desired documents do not exactly meet expectations.

This talk discusses some of the problems encountered when extending a search engine to support queries posed by other software applications and structured documents with derived annotations.

Categories & Subject Descriptors: H.3.3 [Information Search and Retrieval]: *Query Formulation; Retrieval Models; Search Process*

General Terms: Algorithms, Performance, Experimentation.

Keywords: Search engine architecture, Human language technologies, Question answering

Bio

Jamie Callan is a Professor of Computer Science at Carnegie Mellon's Language Technologies Institute and School of Information Systems & Management. Prior to joining CMU he was a Research Assistant Professor at the University of Massachusetts, where he also received his Ph.D. His research and teaching focus on text-based information retrieval, primarily search engine architectures, federated search of groups of search engines, adaptive information filtering, text mining, and information retrieval for educational applications. He has published more than 150 scientific papers. He is the Editor-in-Chief of ACM Transactions on Information Systems (TOIS) and was a founding Editor-in-Chief of Foundations and Trends in Information Retrieval (FnTIR). He has served as Chair of ACM SIGIR, and Program Chair of the SIGIR and CIKM conferences.

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.
ACM 978-1-4503-0099-5/10/10.