Dissertations

Discovering and Using Implicit Data for Information Retrieval

Xing Yi, *University of Massachusetts - Amherst*

Date of Award
9-2011

Document Type
Open Access Dissertation

Degree Name
Doctor of Philosophy (PhD)

Degree Program
Computer Science

First Advisor
James Allan

Second Advisor
W. Bruce Croft

Third Advisor
David D. Jensen

Keywords
Contextual Translation, Implicit Data, Information Retrieval, Language Model, Missing information, Relevance Model

Subject Categories
Computer Sciences

Abstract
In real-world information retrieval (IR) tasks, the searched items and/or the users' queries often have implicit information associated with them -- information that describes unspecified aspects of the items or queries. For example, in web search tasks, web pages are often pointed to by hyperlinks (known as anchors) from other pages, and thus have human-generated succinct descriptions of their content (anchor text) associated with them. This indirectly available information has been shown to improve search effectiveness for different retrieval tasks. However, in many real-world IR challenges this information is sparse in the data; i.e., it is incomplete or missing in a large portion of the data. In this work, we explore how to discover and use implicit information in large amounts of data in the context of IR. We present a general perspective for discovering implicit information and demonstrate how to use the discovered data in four specific IR challenges: (1) finding relevant records in semi-structured databases where many records contain incomplete or

empty fields; (2) searching web pages that have little or no associated anchor text; (3) using click-through records in web query logs to help search pages that have no or very few clicks; and (4) discovering plausible geographic locations for web queries that contain no explicit geographic information. The intuition behind our approach is that data similar in some aspects are often similar in other aspects. Thus we can (a) use the observed information of queries/documents to find similar queries/documents, and then (b) utilize those similar queries/documents to reconstruct plausible implicit information for the original queries/documents. We develop language modeling based techniques to effectively use content similarity among data for our work. Using the four different search tasks on large-scale noisy datasets, we empirically demonstrate the effectiveness of our approach. We further discuss the advantages and weaknesses of two complementary approaches within our general perspective of handling implicit information for retrieval purpose. Taken together, we describe a general perspective that uses contextual similarity among data to discover implicit information for IR challenges. Using this general perspective, we formally present two language modeling based information discovery approaches. We empirically evaluate our approaches using different IR challenges. Our research shows that supporting information discovery tailored to different search tasks can enhance IR systems' search performance and improve users' search experience.