Knowledge Acquisition for Web Search

Marius Paşca
Google Inc.
Mountain View, California 94043
mars@google.com

1 Introduction

The identification of textual items, or documents, that best match a user's information need, as expressed in search queries, forms the core functionality of information retrieval systems. Well-known challenges are associated with understanding the intent behind user queries; and, more importantly, with matching inherently-ambiguous queries to documents that may employ lexically different phrases to convey the same meaning. The conversion of semi-structured content from Wikipedia and other resources into structured data produces knowledge potentially more suitable to database-style queries and, ideally, to use in information retrieval. In parallel, the availability of textual documents on the Web enables an aggressive push towards the automatic acquisition of various types of knowledge from text. Methods developed under the umbrella of open-domain information extraction acquire open-domain classes of instances and relations from Web text. The methods operate over unstructured or semi-structured text available within collections of Web documents, or over relatively more intriguing streams of anonymized search queries. Some of the methods import the automatically-extracted data into human-generated resources, or otherwise exploit existing human-generated resources. In both cases, the goal is to expand the coverage of the initial resources, thus providing information about more of the topics that people in general, and Web search users in particular, may be interested in.

2 Content Overview

The tutorial examines the role of knowledge resources in information retrieval in general, and Web search in particular. A better understanding of the structure and meaning of queries enables a better match of queries against documents, and better ranking of search results. The generation of alternative search results of finer granularity (e.g., quote for a stock symbol, weather forecast for a location, contact info for a business), which more directly answer the user's query, can increase the search effectiveness and the time to result. Similarly, suggesting relevant query completions reduce the time needed to type the entire query, and therefore the time to result. In general, enhancements of the search experience, in the form of spell checking the queries or offering alternative query refinements, represent valuable aids to users of information retrieval systems.

3 Outline

. Introduction
. - Open-domain knowledge
. - Impact of knowledge in information retrieval
. Human-curated knowledge resources
. - Expert resources
. - Collaborative, non-expert resources
. - Hybrid resources
. Automatically-extracted knowledge
. - Web-based textual data sources
. - Open-domain information extraction
. Role of knowledge in information retrieval
. - Retrieval and ranking
. - Search aids
. Discussion
. - Limitations
. - Implications

4 Instructor

Marius Paşca is a research scientist at Google. Current research interests include the acquisition of factual information from unstructured text within documents and queries, and its applications to Web search.