The Role of Wikipedia in Text Analysis and Retrieval

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

Abstract

This tutorial examines the characteristics, advantages and limitations of Wikipedia relative to other existing, human-curated resources of knowledge; derivative resources, created by converting semi-structured content in Wikipedia into structured data; the role of Wikipedia and its derivatives in text analysis; and the role of Wikipedia and its derivatives in enhancing information retrieval.

1 Overview

High expectations of quality and consistency in expert-created knowledge resources reduce the number of their potential contributors. In turn, this makes it difficult to maintain the resources; refresh or add knowledge of the same type, as it becomes relevant over time; or incorporate knowledge of a new type. Especially in the context of Web search, where queries in the long tail reflect different backgrounds and interests of millions of users, resources that are more likely to be stale or incomplete are less likely to consistently provide value. As a counterpart to expert-created resources, non-expert users may collaboratively create large resources of unstructured or semi-structured knowledge, a leading representative of which is Wikipedia. The decentralized construction leads to the inherent lack of any guarantees of quality or reliability, and cannot rule out attempts at adversarial content editing. Nevertheless, articles within Wikipedia are incrementally edited and improved. Collectively, they form an easily-editable collection, reflecting an ever-growing number of topics of interest to people, in general, and Web users, in particular. Furthermore, the conversion of semi-structured content from Wikipedia into structured data makes knowledge from Wikipedia or from one of its derivatives potentially even more suitable for use in text processing or information retrieval.

This tutorial examines the role of Wikipedia in tasks related to text analysis and retrieval. Text analysis tasks, which take advantage of Wikipedia, include coreference resolution, word sense and entity disambiguation, to name only a few. More prominently, they include information extraction. In information retrieval, a better understanding of the structure and meaning of queries enables a better match of queries against documents, and retrieval of knowledge panels for queries asking about popular entities. Concretely, the tutorial teaches the audience about characteristics, advantages and limitations of Wikipedia relative to other existing, human-curated resources of knowledge; derivative resources, created by converting semi-structured content in Wikipedia into structured data; the role of Wikipedia and its derivatives in text analysis; and the role of Wikipedia and its derivatives in enhancing information retrieval.

2 Structure

- Introduction
  - Open-domain knowledge
  - Impact in text analysis
  - Impact in information retrieval
- Human-curated resources
  - Expert resources
  - Collaborative, non-expert resources
  - Hybrid resources
- Knowledge within Wikipedia
  - Articles, infoboxes, links, categories
  - Resources derived from Wikipedia
- Role in text analysis
  - Information extraction
  - Beyond information extraction
- Role in information retrieval
  - Query and document analysis
  - Retrieval and ranking

3 Presenter

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.