Context & Semantics in News & Web Search

Daan Odijk
University of Amsterdam
Amsterdam, The Netherlands
daan@odijk.me

Abstract

This thesis presents research towards a core aim of information retrieval (IR): providing users with easy access to information. Three research themes guide the research presented in this thesis, contributing to three aspects of IR research: the domain in which an IR system is used, the users interacting with the system, and the different access scenarios in which these users engage with an IR system. Central to these research themes is the aim to gain insights into the behavior of searchers and develop algorithms to support them in their quest, whether it is a researcher exploring or studying a large collection, a web searcher struggling to find something, or a television viewer searching for related content.

The first research theme is motivated by the information seeking tasks of researchers exploring and studying large collections. To enable their search on a larger scale, we propose computational methods to connect collections and to infer the perspective offered in a news story. Motivated by how historians select documents for close reading, we propose novel methods for connecting collections using automatically extracted temporal references. To illustrate how these algorithms can be used to automatically create connections between collections, we introduce a novel search interface to explore and analyze the connected collections. The interface highlights different perspectives and requires little domain knowledge. Based on how communication scientists study framing in news, we propose an automatic thematic content analysis approach.

The second research theme is addressed in a mixed-methods study on how web searchers behave when they cannot find what they are looking for. Based on large-scale log analysis, crowd-sourced labeling, and predictive modeling we show behavioral differences given task success and failure. Based on these findings we propose ways in which systems can reduce struggling in search. To support searchers, we propose and evaluate algorithms that accurately predict the nature of future actions and their anticipated impact on search outcomes. Our findings have implications for the design of search systems that help searchers struggle less and succeed more.

In the third and final research theme, we consider a pro-active search scenario, specifically in a live television setting. We propose algorithms that leverage contextual information to retrieve diverse related content for a leaned-back TV viewer. While watching television, people increasingly consume additional content related to what they are watching. Two methods to automatically retrieve content based on subtitles are introduced, one using entity linking, and one that uses reinforcement learning to generate effective queries for finding

related content. Both methods are highly efficient and are currently used in a live television setting in near real time.

Each research chapter in this thesis provides insights and algorithms that help searchers when using IR applications. For varying domains, users, and access scenarios, the research presented in this thesis improves the ease of access to information.

Supervisor:	Prof. dr. M. de Rijke, University of Amsterdam
Co-supervisor:	Dr. E.J. Meij, Yahoo Labs
Available online at:	http://daan.odijk.me/thesis