Theoretical Evaluation of XML Retrieval

Tobias Blanke
King's College London
26-29 Drury Lane
King's College London, London WC2B 5RL
tobias.blanke@kcl.ac.uk

March 26, 2012

Abstract

This thesis has developed a theoretical framework to evaluate XML retrieval. XML retrieval deals with retrieving those document parts that specifically answer a query. It is concerned with using the document structure to improve the retrieval of information from documents by only delivering those parts of a document an information need is about. We define a theoretical evaluation methodology based on the idea of 'aboutness' and apply it to XML retrieval models.

Situation Theory is used to express the aboutness proprieties of XML retrieval models. We propose a Situation Theory framework to evaluate XML retrieval, which is based on the basic and most general information retrieval question how a document (or in our case an XML element) can be about a query. This framework allows us to compare and analyze the reasoning behaviour of XML retrieval models experimented within INEX evaluation campaigns. We develop a dedicated methodology for the evaluation of XML retrieval and apply this methodology to five XML retrieval models from INEX. For each model we derive functional and qualitative properties that qualify its formal behaviour. We compare this behaviour with the underlying flat document retrieval model as well as with a model we specially design to determine how much an XML retrieval model includes XML structure in its reasoning behaviour.

More INEX specific, this thesis further investigates the use of our theoretical evaluation methodology to describe the INEX evaluation methodology. We exemplify theoretical models of user agents and assessment procedures in INEX and derive reasoning assumptions that are included in the specific XML retrieval experimental evaluation, its scales and the ways assessments are done. We point to potential inconsistencies and make suggestions for alternative views on the experimental evaluation dimensions for XML retrieval. Further INEX specifics are discussed when we theoretically analyse filters, as they are used in INEX to deliver only specific answers to an information need. We introduce our theoretical methodology to analyse filters as special aboutness decisions, before applying it to the XML retrieval filtering models.

We finally use the theoretical properties of XML retrieval models and their filters to explain experimental results obtained with some of the XML retrieval models within INEX and draw upon all our previous results to demonstrate how theoretical evaluation insights can

be used to explain results from mainstream experimental evaluations. We relate our theoretical evaluation results with the experimental ones for XML retrieval to find out how the adjustment of existing flat document retrieval models compares to the creation of completely new ones, especially designed to meet the requirements of XML retrieval. For each of the XML retrieval experimental evaluation tasks, we shall determine the reasoning properties that support a good performance and discuss on this basis the experimental performance of the XML retrieval models.

An updated version of the thesis will appear with IOS-Press: Tobias Blanke, Using Situation Theory to evaluate XML Retrieval, IOS-Press: Amsterdam, 2012.