Explicit Web Search Result Diversification

Rodrygo L. T. Santos
School of Computing Science
University of Glasgow
G12 8QQ, Glasgow, UK
rodrygo@dcs.gla.ac.uk

February, 2013

Abstract

Queries submitted to a web search engine are typically short and often ambiguous. With the enormous size of the Web, a misunderstanding of the information need underlying an ambiguous query can misguide the search engine, ultimately leading the user to abandon the originally submitted query. In order to overcome this problem, a sensible approach is to diversify the documents retrieved for the user's query. As a result, the likelihood that at least one of these documents will satisfy the user's actual information need is increased.

In this thesis, we argue that an ambiguous query should be seen as representing not one, but multiple information needs. Based upon this premise, we propose xQuAD—E**x**plicit **Qu**ery **A**spect **D**iversification, a novel probabilistic framework for search result diversification. In particular, the xQuAD framework naturally models several dimensions of the search result diversification problem in a principled yet practical manner. To this end, the framework represents the possible information needs underlying a query as a set of keyword-based *sub-queries*. Moreover, xQuAD accounts for the overall *coverage* of each retrieved document with respect to the identified sub-queries, so as to rank highly diverse documents first. In addition, it accounts for how well each sub-query is covered by the other retrieved documents, so as to promote *novelty*—and hence penalise redundancy—in the ranking. The framework also models the *importance* of each of the identified sub-queries, so as to appropriately cater for the interests of the user population when diversifying the retrieved documents. Finally, since not all queries are equally ambiguous, the xQuAD framework caters for the ambiguity level of different queries, so as to appropriately trade-off *relevance* for *diversity* on a per-query basis.

The xQuAD framework is general and can be used to instantiate several diversification models, including the most prominent models described in the literature. In particular, within xQuAD, each of the aforementioned dimensions of the search result diversification problem can be tackled in a variety of ways. In this thesis, as additional contributions besides the xQuAD framework, we introduce novel machine learning approaches for addressing each of these dimensions. These include a learning to rank approach for identifying effective sub-queries as query suggestions mined from a query log, an intent-aware approach for choosing

the ranking models most likely to be effective for estimating the coverage and novelty of multiple documents with respect to a sub-query, and a selective approach for automatically predicting how much to diversify the documents retrieved for each individual query. In addition, we perform the first empirical analysis of the role of novelty as a diversification strategy for web search.

As demonstrated throughout this thesis, the principles underlying the xQuAD framework are general, sound, and effective. In particular, to validate the contributions of this thesis, we thoroughly assess the effectiveness of xQuAD under the standard experimentation paradigm provided by the diversity task of the TREC 2009, 2010, and 2011 Web tracks. The results of this investigation demonstrate the effectiveness of our proposed framework. Indeed, xQuAD attains consistent and significant improvements in comparison to the most effective diversification approaches in the literature, and across a range of experimental conditions, comprising multiple input rankings, multiple sub-query generation and coverage estimation mechanisms, as well as queries with multiple levels of ambiguity. Altogether, these results corroborate the state-of-the-art diversification performance of xQuAD.

Available online at `http://theses.gla.ac.uk/4106/`.