

Discovery of ranking functions based on Genetic Programming

Marcos Goncalves
Univ. Federal de Minas Gerais
Brazil

Abstract: The effectiveness of information retrieval systems depends fundamentally on the quality of their ranking functions. Literally, thousands of ranking functions have already been proposed and studied in the IR literature. Standard ranking functions, like BM25 and TF-IDF, can, though, present inconsistent behavior, according to the context (e.g., collections, queries) in which they are applied. Thus, some approaches that are able to learn characteristics of that context to generate specific, better tuned, ranking functions have been suggested. One of these approaches is Genetic Programming (GP). Most of the related works are based on the use and combination of statistical evidence of the collection, documents, and queries. Our work, differently from previous approaches, use meaningful and structured evidence extracted from well-known ranking functions (CCA approach), as well as from some probabilities of occurrence of terms and documents in a collection (PROB approach). Our best results using this set of evidence on TREC-8 demonstrate improvements in mean average precision (MAP) of about 41% over BM25, and of almost 18% over a GP approach based on statistical evidences. Other similar gains were also obtained in the WBR99, a web-based collection.