© 2013 FIZ Karlsruhe

io-port 06106535

Liu, Hongyu; Milios, Evangelos

Probabilistic models for focused web crawling.

Comput. Intell. 28, No. 3, 289-328 (2012).

Summary: A focused crawler is an efficient tool used to traverse the web to gather documents on a specific topic. It can be used to build domain-specific web search portals and online personalized search tools. Focused crawlers can only use information obtained from previously crawled pages to estimate the relevance of a newly seen URL. Therefore, good performance depends on powerful modeling of context as well as the quality of the current observations. To address this challenge, we propose capturing sequential patterns along paths leading to targets based on probabilistic models. We model the process of crawling by a walk along an underlying chain of hidden states, defined by hop distance from target pages, from which the actual topics of the documents are observed. When a new document is seen, prediction amounts to estimating the distance of this document from a target. Within this framework, we propose two probabilistic models for focused crawling, maximum entropy Markov model (MEMM) and linear-chain conditional random field (CRF). With MEMM, we exploit multiple overlapping features, such as anchor text, to represent useful context and form a chain of local classifier models. With CRF, a form of undirected graphical models, we focus on obtaining global optimal solutions along the sequences by taking advantage not only of text content, but also of linkage relations. We conclude with an experimental validation and comparison with focused crawling based on best-first search (BFS), hidden Markov model (HMM), and context-graph search (CGS).

Keywords: topical crawlers; focused crawlers; web mining; machine learning; text and link analysis
doi:10.1111/j.1467-8640.2012.00411.x