Abstract

The rapid growth of the World-Wide Web poses unprecedented scaling challenges for general-purpose crawlers and search engines. In the personalized search domain, an alternative to general purpose crawler called focused crawlers are receiving increasing attention. The goal of these crawlers is to selectively seek out pages that are relevant to a pre-defined set of topics or theme. Rather than collecting and indexing all accessible Web documents to be able to answer all possible ad-hoc queries, these crawlers analyzes their crawl boundary to find the links that are likely to be most relevant for the crawl, and avoids irrelevant regions of the Web. This leads to significant savings in hardware and network resources, and helps keep the crawl more up-to-date. This paper presents and compares two focused crawlers called traditional focused crawler and accelerated focused crawler. Accelerated focused
crawler takes offline lessons from traditional focused crawler. It emulates human surfer by trying to predict the relevance of a HREF target page based on words around the link on the source page. The topics are specified using exemplary documents in these experiments. Naive Bayesian classifier is used to guide the crawlers. The crawlers were evaluated for different number of pages crawled, for different number of features gathered from different distances from the link and with different feature selection methods.

References


Index Terms

Computer Science
Web Services

Keywords
Focused Crawler  World Wide Web  and Accelerated focused crawlers.