Beyond Hubs and Authorities: Spreading Out and Zooming In

Soumen Chakrabarti

Indian Institute of Technology, Bombay

Abstract

After crawling and keyword indexing, the next wave that has made a significant impact on Web search is topic distillation: analyzing properties of the hyperlink graph for enhanced ranking of Web pages in response to a query. Hyperlink induced topic search (HITS) and PageRank (used in Google) are two examples. The linear algebra involved in HITS and PageRank is standard, but selecting the relevant subgraph of the Web to which these algorithms should be applied is considerably less clear. PageRank was intended for the entire Web graph (or as much as a crawler can collect) whereas HITS used keyword match followed by a distance-one graph expansion to determine the relevant subgraph.

The clean graph model used in HITS and PageRank, where pages are nodes with no finer characteristics other than a few scalar popularity scores, is also in question. Pages have valuable markup structure and accompanying text. Moreover, the `hubs' or resource lists that make HITS so successful are often `mixed', meaning only specific regions in those pages are relevant to the query.

In this talk we will discuss two enhancements to the graph selection process. First we will describe a learning system called a ``focused crawler'' which discovers and collects large relevant graphs useful for enhanced topic distillation, starting with a few relevant examples and without crawling the Web at large. Second we will discuss a fine-grained model for `micro-hubs' and new algorithms based on the Minimum Description Length principle which let us identify regions in mixed hubs which are relevant to a query, which enhances both topic distillation as well as information extraction.

We will justify, using analyses and anecdotes, that as the Web evolves from static files to dynamically generated semi-structured content, these more complex models and algorithms will become crucial to the continued success of automatic resource discovery, extraction, and annotation.