Special University Oral Examination

Applications of Web Link Analysis

Zoltán Gyöngyi
Department of Computer Science
Stanford University

3:30 PM, Wednesday, June 6, 2007
David Packard Electrical Engineering Building, Room 101
(Dessert and refreshments served at 3:15 PM)

Abstract

Web search engines augment traditional text-based information retrieval techniques with hyperlink analysis to increase the relevance of search results. The most thoroughly studied application of link analysis is the authority-based ranking of web pages. We look beyond assessing page authority and use link analysis to solve other problems that search engines encounter in their quest for better search results.

First, we focus on combating search engine spamming: actions intended to mislead search engines into ranking some web pages higher than they deserve. Over the last five years, the amount of search engine spam has increased dramatically, leading to a degradation of web content quality. To set the stage, we survey current spamming techniques and organize them into a comprehensive taxonomy. Then we delve into detecting a particular technique called link spamming. We introduce the concept of spam mass, a measure of link spamming's impact on the ranking of a page. We discuss how to estimate spam mass and how the estimates can help identifying pages that benefit significantly from link spamming. In our experiments we use spam mass estimates to successfully identify tens of thousands of instances of heavy-weight link spamming.

Second, we turn our attention to web page categorization. Even though web pages are hyperlinked, most proposed efficient classification techniques take little advantage of the link structure and rely primarily on text features. We introduce a link-based approach to classification, centered on summarizing relevant link information about a page into a numeric vector. Our approach can be used in isolation or in conjunction with text-based classification. Large-scale experimental results indicate that link-based classification is on par with text-based classification and the combination of the two offers the best of both worlds.