CCIT Report #601
October 2006

Do Not Crawl in the DUST: Different URLs with Similar Text

Ziv Bar-Yossef* Idit Keidar Uri Schonfeld
Department of Electrical Engineering
Technion – Israel Institute of Technology
Haifa 32000, Israel.
Email: {zivby@ee,idish@ee,shuri@tx}.technion.ac.il.

October 2, 2006

Abstract

We consider the problem of DUST: Different URLs with Similar Text. Such duplicate URLs are prevalent in web sites, as web server software often uses aliases and redirections, translates URLs to some canonical form, and dynamically generates the same page from various different URL requests. We present a novel algorithm, *DustBuster*, for uncovering DUST; that is, for discovering rules for transforming a given URL to others that are likely to have similar content. DustBuster is able to mine DUST effectively from previous crawl logs or web server logs, *without* examining page contents. Verifying these rules via sampling requires fetching few actual web pages. Search engines can benefit from this information to increase the effectiveness of crawling, reduce indexing overhead as well as improve the quality of popularity statistics such as PageRank.

*Supported by the European Commission Marie Curie International Re-integration Grant.