Dynamic Full Text Category Search

By

Patrick Wong

A Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the degree of

Master of Science

Major Subject: Computer Science

Approved:

_____

Mukkai Krishnamoorthy, Thesis Adviser

_____

Martin Hardwick, Committee Member

_____

David L. Spooner, Committee Member

Rensselaer Polytechnic Institute
Troy, New York

March, 2012
(For Graduation May 2012)

ABSTRACT

Finding a single or multiple well defined keywords in a document is a well-known problem with many solutions. But what happens if one wants to find a list of words that are related through some predefined categorization hierarchy? Furthermore, what if these hierarchies can change every week, day, or hour? In this paper, we implement a category searching system predicated on finding all related entries within a given document, whether that be a web-based HTML formatted file or a plain text file, of a user chosen predefined category. Users have the option of designing their own categories, using them to find matches within a document, and having their categories available for use by other users. Our system is based on part-of-speech tagging and uses a prebuilt part-of-speech regular expression grammar to locate all nouns and noun-phrases. For verifying our system, we built another system based on phrase matching. Results indicate that our phrase matching system actually has less false negatives and is almost three times as fast compared to our part-of-speech tagging system, but typically incurs more false positives. Overall, both systems on average contain little to no false negatives and are able to process an average HTML file size of 20k bytes in under 15 seconds.