CCSU

DEPARTMENT OF MATHEMATICAL SCIENCES

COLLOQUIUM

Thursday, October 11
12:10 – 1:30 PM
Willard Hall, Room 204

TOPICAL DISCOVERY OF WEB CONTENT

GIANCARLO CROCETTI

(Data Mining MS Thesis Presentation)

CENTRAL CONNECTICUT STATE UNIVERSITY

Abstract: Unstructured data sources, related to research topics, are identified and indexed by search engines that will augment the textual content with annotations and transform it into a form suited for information retrieval. Even though this process is proven successful, it contains a weak link represented by the manual selection of new unstructured data sources to be added to the system over time. These data sources represent a corpus of relevant documents related to a specific research area and are in the form of web pages. To this end, this thesis describes the theory and the implementation of the author's new software tool, the "Web Topical Discovery System" (WTDS), which provides an approach to the automatic discovery and selection of new web pages relevant to specific analytical needs. We will see how it is possible to specify the research context with search keywords related to the area of interest and consider the important problem of removing extraneous data from a web page containing an article in order to reduce, to a minimum, false positives represented by a match on a keyword that is showing up on the latest news box of the same page. The removal of duplicates, the analysis of richness of information contained in the article and lexical diversity are all taken into consideration in order to provide the optimum set of recommendations to the end user or system.

For further information:

gotchevi@ccsu.edu 860-832-2839

http://www.math.ccsu.edu/gotchev/colloquium/