University of Texas at Dallas
CS 6322 : Information Retrieval
Fall 2011
Instructor: Dr. Sanda Harabagiu
Grader: Bryan Rink

**Project 3: Citation Analysis**

In this project, an architecture involving a citation analysis system and a graphical interface has to be built. Citation analysis is the examination of the frequency, patterns and graphs of references (citations) in scientific articles and books. It uses citations in scholarly works to establish links to other works or other researchers. It is one of the most widely used methods of the science of bibliometrics. For example, bibliographic coupling and co-citation are association measures based on citation analysis (shared citations or shared references). A paper that describes an implementation of citation analysis is:

- C.L. Giles, K. Bollacker, S. Lawrence, "CiteSeer: An Automatic Citation Indexing System," DL'98 Digital Libraries, 3rd ACM Conference on Digital Libraries, 89-98, 1998

Automated citation indexing[] has changed the nature of citation analysis research, allowing millions of citations to be analyzed for large scale patterns and knowledge discovery. The first example of automated citation indexing was CiteSeer, later to be followed by Google Scholar.

The architecture of this project involves :

**Module 1***:* *Paper Management Module* which downloads and saves a large set of published papers in Computational linguistics. The papers are available at:
http://www.aclweb.org/anthology-new/
It is recommended to download only papers published after 2000.
**Module 2:** *Index of citations* – enabling the processing the pdf files of the papers and organizing an index of citations for each paper as well as for each citation.
**Module 3***:* *Citation query processing* – allowing users to find:

1. which papers are cited by which papers in the paper collection
2. which authors cite which paper authors
3. which words are mostly used in papers that cite each other, as an implementation of the topics that they describe and the labels of those topics.

**Module 4**: *Graphical visualization* – of your choice. Displays the resulting networks of citations;

Possible visualisation :