Summary

The World Wide Web can be represented as a graph, with web pages as nodes and links as edges. This idea and its related research is the impetus for many web technologies. We are working to advance the study of a more detailed web graph. That is, the creation and use of a system to classify links without imposed categorization. Thus, multi-relational subgraphs emerge naturally from the web and provide useful data hidden within the relationship between text and structure. This information can be further enhanced by temporal analysis and Eigendecomposition for spectral analysis.

Applications

Current web search technology is based upon analysis of the web graph. Subgraphs could be used either explicitly or implicitly to improve the relevance of results and provide more targeted ways to search for information.

Social and Political Influence

Classified subgraphs will have nodes with new centrality values, that in the original graph appear to have a low centrality, which could reveal clusters or single nodes with high influence in the traffic of information of a certain type.

Methods

Web page content is crawled in a breadth-first fashion, starting on a single page. A database record is created for the content of each page, as well as each link on the page. Additional identification is stored for each record to allow for a network data structure to be composed from the database. The crawler is currently written in Python and runs at a rate of 35,000 pages per day on a single machine.

Links are classified using Support Vector Machines. SVMs are a set of related supervised learning methods which analyze data and recognize patterns. An SVM is a classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM algorithm builds a model that predicts whether a new example falls into one category or the other. Currently a vector-space for the "Blogspot" graph has been built that classifies with 86.8% accuracy.

Timeline

The resulting classification system correctly labeled 434/500 links of an additional random 500.

References


Drexel Department of Electrical And Computer Engineering, Temple Department of Computer and Information Sciences August 2010