Extracting knowledge trees from Wikipedia article network

Kaushik Mohan

Introduction

Network

- Wikipedia articles as nodes and hyperlinks as edges
- Structure is very different compared to traditional and historical encyclopedias

Objective

- To extract hierarchical structure of articles for a field (say, Mathematics, Physics..) using centrality based algorithms

Why?

- Individual learning is still hierarchical
- Can be used to create structured curriculum content with just resources available on the web
- Identify gaps in information on the web

Methodology

- Centrality measures to define hierarchy score
- Compare scores of neighbors to determine which is higher or lower in the hierarchy
- Attraction basin hierarchy algorithm*

*Lev Muchnik, Royi Itzhack, Sorin Solomon and Yoram Louzoun- *Self-emergence of knowledge trees: Extraction of the Wikipedia hierarchies*, Physical Review E (2007)

Challenges

- Validation of the results
- Computations of centrality measures are costly
  - Previously not been done for Wikipedia EN database due to size

Milestones

- Week 1-2
  - Load, subset and clean SNAP Wikipedia dataset (2011)
  - Create corresponding Wikipedia category network
  - Calculate and verify basic Network statistics
- Week 3-4
  - Implement betweenness centrality and Page Rank based hierarchy extraction algorithms
  - Implement HITS(Hyperlink Induced Topic Search) and Attraction basin hierarchy algorithm
- Week 5
  - Compare results against prior research and Wiki Category Network
  - Extract trees for different topics

Continuation

- Changing network structure of Wikipedia to content based vs. hyperlink based
- Using information retrieval methodologies or better content analysis methods on article text to create attributes
- Using knowledge of content based connections + preferential attachment as a potentially better model for Wiki/WWW
  - Current models underestimate the clustering within groups
  - Given topics within a group are highly likely to share common topics/keywords, it could better model the clustering or grouping seen in the network (hypothesis)