Realtime and Big Data Analytics Project Proposal

Part 1. General Information

Team Members: Aditya Garg, Aashutosh Trivedi, Ansa Mary Ephraim
Project Title: Analysis of Citation Network
Project Description:
We access citation database of conferences over a period of time. Some analytics we hope to derive are
1) Potential Hot Topic : Paper with the least citation distance from all the published papers.
2) Find citation distance and collaboration distance between authors
3) Help suggest should two authors work together in future.
4) Suggest/predict conferences to submit entries to for authors

Data Sources - Use the table below to list and describe potential data sources.

Part 2. General Data Source Information

Data Sources	**Data Source Description**	**Data Size**
http://snap.stanford.edu/data/cit-HepTh.html	Paper citation network of Arxiv High Energy Physics Theory category	36.3 MB
http://www.cs.cornell.edu/projects/kddcup/datasets.html	29,000 hep-th papers with 1.7 gigs of data	36.37 MB
http://people.cs.umass.edu/~mccallum/data.html	Simulated/Real/Aviation/Auto UseNet data	1.02 MB
http://konect.uni-koblenz.de/networks/subelj_cora	Cora citation network datase	2.05 MB
Google Scholar	A freely accessible web search engine that indexes the full text of scholarly literature across an array of publishing formats and disciplines.	-
DBLP	This server provides bibliographic information on major **computer science journals and proceedings**. DBLP indexes more than 2.3 million articles and contains many links to home pages of computer scientists.	1.3 GB

Part 3. Detailed Data Source Information

Data Sources	**Data Characteristics**	**Data Frequency**
Google Scholar	Will require some scraping to access citation list of specific papers	Very frequent, almost up-to-date
DBLP	Metadata such as paper name, conference, authors available in a single XML file	Very frequent, almost up-to-date

http://snap.stanford.edu/data/cit-HepTh.html	This is a static historical data of which papers have cited which others in high energy theoretical physics	Historic, not updated
http://people.cs.umass.edu/~mccallum/data.html	Historical data of about 73000 articles from 4 discussion groups for simulated and real auto and aviation.	Historic, not updated
http://konect.uni-koblenz.de/networks/subelj_cora	Historic dataset of over 23000 cora citation papers network. The network is connected.	Historic, not updated
Data available online about conferences – specifically, the list of panel members	We will need to lookup selected journals and conferences and find out the editors or panel members of the same.	

Part 4. Technologies

We plan to write our code in Java, and do a lot of the parsing in Python. MapReduce programs will also be written to extract data and build graphs. Web scraping software will also be put to use in gathering more data off Google Scholar in particular.

Part 5. References

Analysis of SIGMOD's Co-Authorship Graph
http://sigmod.acm.org/publications/sigmod-record/0309/3.coauthorship.pdf

Some Analyses of Erdos″ Collaboration Graph
http://vlado.fmf.uni-lj.si/pub/networks/doc/erdos/erdos.pdf

The Chasms of CSCW: A Citation Graph Analysis of the CSCW Conference
http://www.tau.ac.il/~gailgf/pubs/Chasms.pdf

Understanding Importance of Collaborations in Co-authorship Networks: A Supportiveness Analysis Approach
http://www.cs.sfu.ca/~jpei/publications/SupportivenessAnalysis-sdm'09.pdf

The structure of scientific collaboration networks
http://www.pnas.org/content/98/2/404.full

Analysis of Papers from Twenty-Five Years of SIGIR Conferences: What Have We Been Doing for the Last Quarter of a Century ?
http://sigir.org/files/forum/F2002/smeaton.pdf