Big Data Analytics Toolkit for Business Data

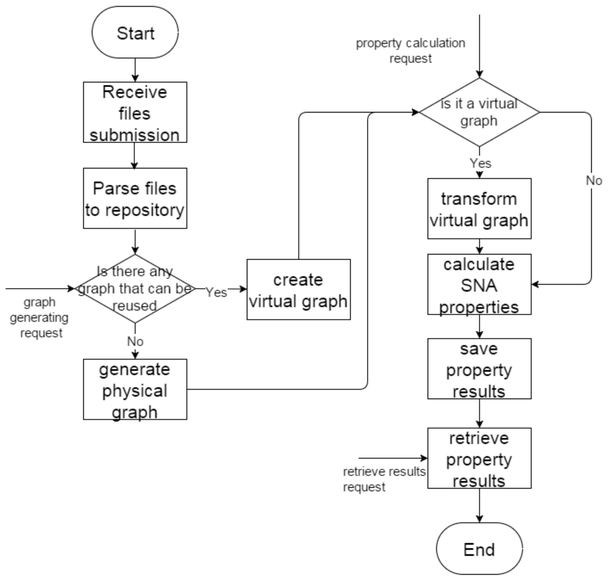
Fan Liang, Weichang Du

Faculty of Computer Science, University of New Brunswick

Introduction

Modeling data as networks is of great interest in business applications. Social network analysis (SNA) measures the relationships and structures using a set of metrics by building graphs.

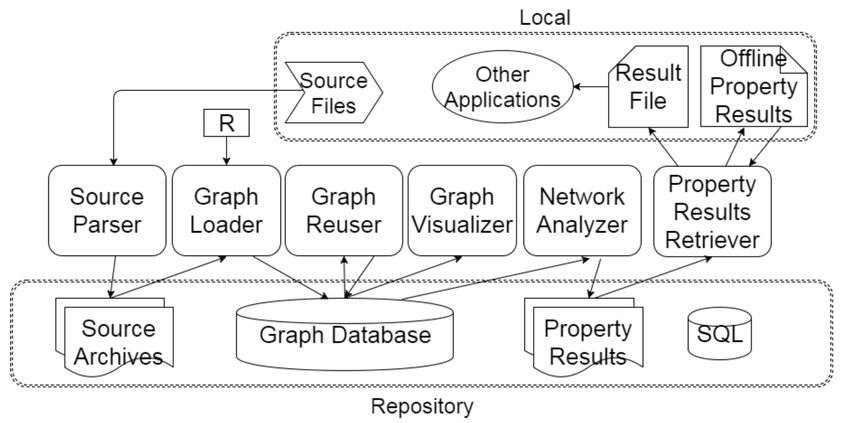
To analyze a large volume of business data, we develop a software system which combines the big data analytics and social network analysis techniques. Its workflow consists of data collection, graph generation, graph reuse, network property calculation, SNA result interpretation and application integration.



Architecture

The system supports ingestion, processing, and display of graphs on large datasets. It provides a flexible, component based, pipeline architecture for the integration and deployment of SNA property algorithms in the fields of business and social network analysis.

The system consists of six modules, which are source parser, graph loader, graph reuser, network analyzer, properties results handler and graph visualizer.



System Features

- Data collection: The system parses source files submitted by users to its repository. The system defines a set of certain formats of source files which the system can decipher.
- Graph generation: The system reads source data from the source and creates graphs in the graph database. Users give the specification which indicates how to construct networks. The system automates the process to select and create nodes and edges.
- Graph reuse: To save disk storage, we can reuse the existing graphs in the graph database. We store virtual graphs at the beginning. Whenever users want to do graph analysis, the virtual graphs are transformed from existing graphs. The process is automatic and is transparent to users.
- SNA network property calculation: The system implements a list of built-in property programs that users can count on. After the calculation is completed, the property results are saved to the repository for application interpretations.
- SNA property results interpretation and application integration: The system provides APIs for users to retrieve SNA property results and use in their own applications.

Application Example

We use the methods in [1] to build and identify key players in stock networks.

The system calls R to calculate correlations between stocks based on values of stock prices in time sequence. The user chooses four centrality properties to calculate: degree centrality, betweenness centrality, closeness centrality and eigenvector centrality. The overall scores are calculated by Principal Components Analysis (PCA) with combining with the four centralities scores.

Table 1: Top three highest overall scores stocks. (The time is from 2008-1-4 to 2013-1-3. The correlation is Pearson Correlation. The weight threshold is from 0.6 to 1.)

Name	Degree Centrality	Closeness Centrality	Betweenness Centrality	Eigenvector Centrality	Overall scores
Southwest Airlines Co.	0.001226	815.8534	0	6.98E-195	815.8534
Alaska Air Group	0.001226	815.8534	0	6.98E-195	815.8534
The Kroger Co.	0.00134	746.3575	0	4.60E-191	746.3575

Reference

[1] Gan S L, Djauhari M A. An Overall Centrality Measure: The Case of US Stock Market[J]. International Journal of Basic & Applied Sciences, 2012, 12(6): 99-103.