

ISMB 2003 Tutorial

Mining the Biomedical Literature using Semantic Analysis and Natural Language Processing Techniques, a Link Analysis Approach

Lecturer: Ronen Feldman

Contact Information:

Ronen Feldman, Ph.D.
ClearForest Corporation
15 East 26th Street, Suite 1711
New York, NY, 10010
ronen@clearforest.com
Tel: 212-432-1515 (x203)
Fax: 212-432-1929

Abstract

The information age has made it easy to store large amounts of data. The proliferation of the Biomedical literature available on the Web, on corporate intranets, on news wires, and elsewhere is overwhelming. However, while the amount of data available to us is constantly increasing, our ability to absorb and process this information remains constant. Search engines only exacerbate the problem by making more and more documents available in a matter of a few key strokes.

Link Analysis is a new and exciting research area that tries to solve the information overload problem by combining techniques from data mining, machine learning, Information Extraction, Text Categorization, Visualization and Knowledge Management. Link Analysis is the process of building up networks of interconnected objects through various relationships in order to discover patterns and trends. The main tasks of link analysis are to extract, discover, and link together sparse evidence from vast amounts of data sources, to represent and evaluate the significance of the related evidence, and to learn patterns to guide the extraction, discovery, and linkage of entities. The discovered relationships could be transactional, geographical, social, or temporal.

Link Analysis for BioInformatics involves the preprocessing of biomedical document collections (by using text categorization, term extraction, and information extraction), integration with structured information sources, the storage of the intermediate representations, the techniques to analyze these intermediate representations (distribution analysis, clustering, trend analysis, association rules, etc.) and visualization of the results. In this tutorial we will present the general theory of Link Analysis for BioInformatics, survey recent work, and will demonstrate several systems that use these principles to enable interactive exploration of a combination of structured and unstructured collections.

We will present a general architecture of link analysis systems and outline the algorithms and data structures behind the systems. The Tutorial will cover the state of the art in this rapidly growing area of research. Several real world applications of link analysis will be presented.

Audience

The tutorial should be of interest to practitioners from BioInformatics, and users of the general biomedical literature that are interested in this fast-growing research area. In particular we will provide examples from genomics, proteomics and other bio-information domains.

Prerequisites

The tutorial is suitable to the general audience. No special knowledge is needed as the tutorial is self-contained. We will cover information extraction and link analysis techniques with reasonable depth, and provide complete scenarios for link analysis from the biomedical literature.

Coverage

The tutorial will cover the state of the art in Link Analysis. The tutorial will be broad in nature, but we will get deeper on several topics. These topics will include Link Analysis System's architecture, preprocessing techniques such as Information extraction techniques, analytic techniques, visualization methods, and applications of link analysis. The collections that we will use for the examples include: Medline, Flybase and YPD.

Lecturers' Biography

Ronen Feldman is a senior lecturer at the Mathematics and Computer Science Department of Bar-Ilan University in Israel, and the Director of the Data Mining Laboratory. He received his B.Sc. in Math, Physics and Computer Science from the Hebrew University, M.Sc. in Computer Science from Bar-Ilan University, and his Ph.D. in Computer Science from Cornell University in NY. He is the founder and president of ClearForest Corporation, a NY based company specializing in development of text mining tools and applications. He is also an Adjunct Professor at NYU Stern Business School.

Other Tutorials given by Ronen Feldman:

Mining Unstructured Data: KDD'97, KDD'99, IJCAI'99, EACL'99, ASIS'99, AAI'2000, CIKM'2001, KDD'2002.

This tutorial is based on the "Advanced Topics" course that I have taught in the NYU Stern Business School. This tutorial is intended as a purely educational session and is NOT by any means intended to promote the business of ClearForest Corporation, or any other commercial entity. It is based on many years of experience in Text Mining, Link Analysis and Information Extraction.

Outline:

1. Introduction to Link Analysis

- a. What is Link Analysis?
 - b. Architecture of Link Analysis Systems
- 2. Pre Processing Techniques
 - a. Information Extraction
 - i. Types of IE Tasks
 - 1. Entity Extraction
 - 2. Fact Extraction
 - 3. Event Extraction
 - ii. Architecture of IE systems
 - 1. Rule Based Systems
 - 2. Machine Learning Based Systems
 - iii. Environments for Creating IE Systems
 - iv. Evaluation of IE Systems
 - 1. MUC
 - 2. ACE
 - b. Information Extraction for BioInformatics
 - i. Extraction of Genes, Gene Products, Proteins from Scientific Articles.
 - ii. Extracting Experimental Evidence about Genes, Proteins and their relationships from Medline Articles.
 - c. Document Classification
 - i. Probabilistic Algorithms
 - ii. Instance Based Algorithms
 - iii. Neural Network Based Algorithms
 - iv. Support Vector Machines
 - v. Committee Based Algorithms
 - d. Using Categorization for Biomedical Research
- 3. Knowledge Representation
 - a. Handling Uncertainty
 - b. Binary Relationships vs. N-ary Relationships
 - c. Temporal Relationships
 - d. Using Ontological Information
 - i. Biomedical Ontologies
 - 1. UMLS
 - 2. MeSH
 - 3. SNOMED
 - 4. GeneOntology
 - e. Document Representation Standards
 - i. XML
 - ii. DAML
 - iii. RKF
- 4. Link Analysis Systems
 - a. Knowledge Based Systems
 - i. Ontology Based Systems
 - CYC
 - Using CYC with Biomedical Applications

- ii. Inference Mechanisms
 - Logic Based Systems
 - iii. Pattern Based Languages
 - KIF
 - b. Statistical Systems
 - i. Clustering Based Systems
 - c. Social Based Systems
- 5. Link Analysis Operations
 - a. Integration of Unstructured Data and Structured Data
 - i. Virtual Taxonomies
 - ii. Connection with Biomedical Databases
 - b. Algorithms for Identifying Links between Entities
 - c. Pattern Matching between Network Structures
 - i. Soft Matching between Entities
 - ii. Graph Matching Algorithms
 - iii. Model Based Matching
 - d. Identifying “Interesting” Patterns
 - e. Pruning Search Spaces
 - f. Visual Query Languages
 - g. Detecting Changes in Link Structures
 - h. Link Exploration Techniques
 - i. Using Agents to fill Information Gaps
 - j. User Profiling
 - k. Alert Generation
- 6. Visualization Techniques
 - a. Circle Graphs
 - b. Hierarchical Graphs
 - c. Symmetric Graphs
 - d. “Fisheye” Diagrams
 - e. Critical Path Analysis
 - f. Dynamic Evolution of relationships
- 7. Commercial Systems
 - a. NetMap
 - b. Analysts’ Notebook
 - c. Visual Analytics
- 8. Bio Information applications
 - i. Genomics and Proteomics
 - ii. Live Demo: Link Analysis based on Medline abstracts
 - iii. Live Demo: Link Analysis based on Flybase