Web Mining ed Analisi delle Reti Sociali

Mining on Complex (Social) Network

Dino Pedreschi
Dipartimento di Informatica
Università di Pisa
www.di.unipi.it/~pedre

Social Network Analysis

- Social Network Introduction
- Statistics and Probability Theory
- Models of Social Network Generation
- Mining on Social Network
- Summary

Information on the Social Network

- Heterogeneous, multi-relational data represented as a graph or network
  - Nodes are objects
    - May have different kinds of objects
    - Objects have attributes
    - Objects may have labels or classes
  - Edges are links
    - May have different kinds of links
    - Links may have attributes
    - Links may be directed, are not required to be binary
- Links represent relationships and interactions between objects - rich content for mining

What is New for Link Mining Here

- Traditional machine learning and data mining approaches assume:
  - A random sample of homogeneous objects from single relation
- Real world data sets:
  - Multi-relational, heterogeneous and semi-structured
- Link Mining
  - Newly emerging research area at the intersection of research in social network and link analysis, hypertext and web mining, graph mining, relational learning and inductive logic programming

A Taxonomy of Common Link Mining Tasks

- Object-Related Tasks
  - Link-based object ranking
  - Link-based object classification
  - Object clustering (group detection)
  - Object identification (entity resolution)
- Link-Related Tasks
  - Link prediction
- Graph-Related Tasks
  - Subgraph discovery
  - Graph classification
  - Generative model for graphs

What Is a Link in Link Mining?

- Link: relationship among data
- Two kinds of linked networks
  - homogeneous vs. heterogeneous
- Homogeneous networks
  - Single object type and single link type
  - Single model social networks (e.g., friends)
  - WWW: a collection of linked Web pages
- Heterogeneous networks
  - Multiple object and link types
  - Medical network: patients, doctors, disease, contacts, treatments
  - Bibliographic network: publications, authors, venues

Link-Based Object Ranking (LBR)

- LBR: Exploit the link structure of a graph to order or prioritize the set of objects within the graph
  - Focused on graphs with single object type and single link type
- This is a primary focus of link analysis community
- Web information analysis
  - PageRank and Hits are typical LBR approaches
- In social network analysis (SNA), LBR is a core analysis task
  - Objective: rank individuals in terms of "centrality"
  - Degree centrality vs. eigen vector/power centrality
  - Rank objects relative to one or more relevant objects in the graph vs. ranks object over time in dynamic graphs

PageRank: Capturing Page Popularity (Brin & Page'98)

- Intuitions
  - Links are like citations in literature
  - A page that is cited often can be expected to be more useful in general
- PageRank is essentially "citation counting", but improves over simple counting
  - Consider "indirect citations" (being cited by a highly cited paper counts a lot…)
  - Smoothing of citations (every page is assumed to have a non-zero citation count)
- PageRank can also be interpreted as random surfing (thus capturing popularity)

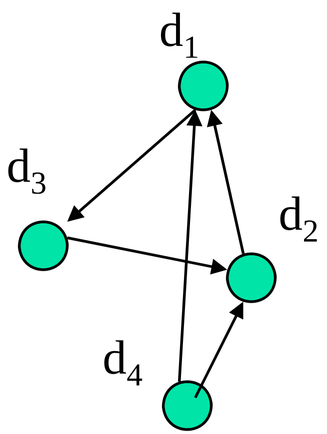
The PageRank Algorithm (Brin & Page'98)

Random surfing model:

At any page,

With prob. α, randomly jumping to a page

With prob. $(1 - \alpha)$, randomly picking a link to follow



$$M = \begin{bmatrix} 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{bmatrix}$$

"Transition matrix"

Same as α/N (why?)

$$p_{t+1}(d_i) = (1-\alpha) \sum_{d_j \in IN(d_i)} m_{ji} p_t(d_j) + \alpha \sum_k \frac{1}{N} p_t(d_k)$$

$$p(d_i) = \sum_k [\frac{1}{N}\alpha + (1-\alpha) m_{ki}] p(d_k)$$

Stationary ("stable") distribution, so we ignore time

$$\vec{p} = (\alpha I + (1-\alpha) M)^T \vec{p}$$

$I_{ij} = 1/N$

Initial value p(d)=1/N

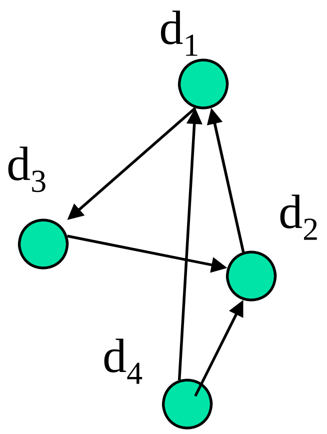
Iterate until converge

Essentially an eigenvector problem….

HITS: Capturing Authorities & Hubs (Kleinberg'98)

- Intuitions
  - Pages that are widely cited are good authorities
  - Pages that cite many other pages are good hubs
- The key idea of HITS
  - Good authorities are cited by good hubs
  - Good hubs point to good authorities
  - Iterative reinforcement …

The HITS Algorithm (Kleinberg 98)



$$A = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

"Adjacency matrix"

Initial values: a=h=1

$$\left.\begin{aligned} h(d_i) &= \sum_{d_j \in OUT(d_i)} a(d_j) \\ a(d_i) &= \sum_{d_j \in IN(d_i)} h(d_j) \end{aligned}\right\} \text{Iterate}$$

Normalize:

$$\sum_i a(d_i)^2 = \sum_i h(d_i)^2 = 1$$

$$\vec{h} = A\vec{a}; \quad \vec{a} = A^T \vec{h}$$

$$\vec{h} = AA^T \vec{h}; \quad \vec{a} = A^T A\vec{a}$$

Again eigenvector problems…

Block-level Link Analysis (Cai et al. 04)

- Most of the existing link analysis algorithms, e.g. PageRank and HITS, treat a web page as a single node in the web graph
- However, in most cases, a web page contains multiple semantics and hence it might not be considered as an atomic and homogeneous node
- Web page is partitioned into blocks using the vision-based page segmentation algorithm
- extract page-to-block, block-to-page relationships
- Block-level PageRank and Block-level HITS

Link-Based Object Classification (LBC)

- Predicting the category of an object based on its attributes, its links and the attributes of linked objects
- **Web**: Predict the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags, etc.
- **Citation**: Predict the topic of a paper, based on word occurrence, citations, co-citations
- **Epidemics**: Predict disease type based on characteristics of the patients infected by the disease
- **Communication**: Predict whether a communication contact is by email, phone call or mail

Challenges in Link-Based Classification

- Labels of related objects tend to be correlated
- Collective classification: Explore such correlations and jointly infer the categorical values associated with the objects in the graph
- Ex: Classify related news items in Reuter data sets (Chak'98)
  - Simply incorp. words from neighboring documents: not helpful
- Multi-relational classification is another solution for link-based classification

Group Detection

- Cluster the nodes in the graph into groups that share common characteristics
  - **Web:** identifying communities
  - **Citation:** identifying research communities
- Methods
  - Hierarchical clustering
  - Blockmodeling of SNA
  - Spectral graph partitioning
  - Stochastic blockmodeling
  - Multi-relational clustering

Entity Resolution

- Predicting when two objects are the same, based on their attributes and their links
- Also known as: deduplication, reference reconciliation, co-reference resolution, object consolidation
- Applications
  - **Web**: predict when two sites are mirrors of each other
  - **Citation**: predicting when two citations are referring to the same paper
  - **Epidemics**: predicting when two disease strains are the same
  - **Biology:** learning when two names refer to the same protein

Entity Resolution Methods

- Earlier viewed as pair-wise resolution problem: resolved based on the similarity of their attributes
- Importance at considering links
  - Coauthor links in bib data, hierarchical links between spatial references, co-occurrence links between name references in documents
- Use of links in resolution
  - Collective entity resolution: one resolution decision affects another if they are linked
    - Propagating evidence over links in a depen. graph
  - Probabilistic models interact with different entity recognition decisions

Link Prediction

- Predict whether a link exists between two entities, based on attributes and other observed links
- Applications
  - **Web**: predict if there will be a link between two pages
  - **Citation**: predicting if a paper will cite another paper
  - **Epidemics**: predicting who a patient's contacts are
- Methods
  - Often viewed as a binary classification problem
  - Local conditional probability model, based on structural and attribute features
  - Difficulty: sparseness of existing links
  - Collective prediction, e.g., Markov random field model

Link Cardinality Estimation

- Predicting the number of links to an object
  - **Web**: predict the authority of a page based on the number of in-links; identifying hubs based on the number of out-links
  - **Citation**: predicting the impact of a paper based on the number of citations
  - **Epidemics**: predicting the number of people that will be infected based on the infectiousness of a disease
- Predicting the number of objects reached along a path from an object
  - **Web**: predicting number of pages retrieved by crawling a site
  - **Citation**: predicting the number of citations of a particular author in a specific journal

Subgraph Discovery

- Find characteristic subgraphs
  - Focus of graph-based data mining
- Applications
  - **Biology:** protein structure discovery
  - **Communications:** legitimate vs. illegitimate groups
  - **Chemistry:** chemical substructure discovery
- Methods
  - Subgraph pattern mining
- Graph classification
  - Classification based on subgraph pattern analysis

Metadata Mining

- Schema mapping, schema discovery, schema reformulation
- **cite** – matching between two bibliographic sources
- **web** - discovering schema from unstructured or semi-structured data
- **bio** – mapping between two medical ontologies

Link Mining Challenges

- Logical vs. statistical dependencies
- Feature construction
- Instances vs. classes
- Collective classification
- Collective consolidation
- Effective use of labeled & unlabeled data
- Link prediction
- Closed vs. open world

Challenges common to any link-based statistical model (Bayesian Logic Programs, Conditional Random Fields, Probabilistic Relational Models, Relational Markov Networks, Relational Probability Trees, Stochastic Logic Programming to name a few)

Logical vs. Statistical Dependence

- Coherently handling two types of dependence structures:
  - Link structure - the logical relationships between objects
  - Probabilistic dependence - statistical relationships between attributes
- Challenge: statistical models that support rich logical relationships
- Model search complicated by the fact that attributes can depend on arbitrarily linked attributes -- issue: how to search this huge space

Feature Construction

- In many cases, objects are linked to a **set** of objects. To construct a single feature from this set of objects, we may either use:
  - Aggregation
  - Selection

Individuals vs. Classes

- Does model refer
  - explicitly to individuals
  - classes or generic categories of individuals
- On one hand, we'd like to be able to model that a connection to a particular individual may be highly predictive
- On the other hand, we'd like our models to generalize to new situations, with different individuals

Collective Classification

- Using a link-based statistical model for classification
- Inference using learned model is complicated by the fact that there is correlation between the object labels

Collective Consolidation

- Using a link-based statistical model for object consolidation
- Consolidation decisions should not be made independently

Labeled & Unlabeled Data

- In link-based domains, unlabeled data provide three sources of information:
  - Helps us infer object attribute distribution
  - Links between unlabeled data allow us to make use of attributes of linked objects
  - Links between labeled data and unlabeled data (training data and test data) help us make more accurate inferences

Link Prior Probability

- The prior probability of any particular link is typically extraordinarily low
- For medium-sized data sets, we have had success with building explicit models of link existence
- It may be more effective to model links at higher level--required for large data sets

Closed World vs. Open World

- The majority of SRL approaches make a closed world assumption, which assumes that we know all the potential entities in the domain
- In many cases, this is unrealistic
- Work by Milch, Marti, Russell on BLOG

Social Network Analysis

- Social Network Introduction
- Statistics and Probability Theory
- Models of Social Network Generation
- Networks in Biological System
- Mining on Social Network
- Summary

Ref: Mining on Social Networks

- D. Liben-Nowell and J. Kleinberg. The Link Prediction Problem for Social Networks. CIKM'03
- P. Domingos and M. Richardson, Mining the Network Value of Customers. KDD'01
- M. Richardson and P. Domingos, Mining Knowledge-Sharing Sites for Viral Marketing. KDD'02
- D. Kempe, J. Kleinberg, and E. Tardos, Maximizing the Spread of Influence through a Social Network. KDD'03.
- P. Domingos, Mining Social Networks for Viral Marketing. IEEE Intelligent Systems, 20(1), 80-82, 2005.
- S. Brin and L. Page, The anatomy of a large scale hypertextual Web search engine. WWW7.
- S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, Mining the link structure of the World Wide Web. IEEE Computer'99
- D. Cai, X. He, J. Wen, and W. Ma, Block-level Link Analysis. SIGIR'2004.

Other References

- Lecture notes from Professor Lise Getoor's website.
  http://www.cs.umd.edu/~getoor/
- Lecture notes from Professor ChengXiang Zhai's website.
  http://www-faculty.cs.uiuc.edu/~czhai/