Complex Social Network Mining
—Theory, Methodologies, and Applications

Jie Tang

Department of Computer Science and Technology
Tsinghua University
Email: jietang@tsinghua.edu.cn
Social Networks

Web 1.0 (1989)
Pages, hyperlinks
Relevance search

Web 2.0 (2004)
social networks

Mobile Web (2008-20)
Connecting via mobiles…

Web-based (or mobile-based) social networks already become a bridge to connect our real daily life and the virtual web space
Web-based Social Network Mining
—Theory, Methodologies, and Applications

Social data integration

Search/query over networks

Attribute/link prediction

Social knowledge acquisition

Social influence analysis

Social dynamics

Trust and privacy

Collective learning | Learning from users | Social theory | Graphical models

Web 1.0: Web of Pages ➔ Web 2.0: Web of People ➔ Web 3.0: Web of Semantics

Theoretical layer
Outline

• ArnetMiner: Academic Social Network

• Core Techniques
  – Knowledge Acquisition
  – Semantic Integration
  – Heterogeneous Ranking
  – Social Influence Analysis

• Demo
ArnetMiner.org
- Academic research social network analysis and mining system

提供全面的研究者网络分析与挖掘功能

Papers published: ACM TKDD, KDD’08-10, SDM’09, ICDM’07-09, CIKM’07-09, DKE, JIS

http://arnetminer.org/
Why Arnetminer.org?

"Academic search is treated as document search, but ignore semantics"

"The information need is not only about publication..."
Examples – Expertise search

- When starting a work in a new research topic;
- Or brainstorming for novel ideas.

Who are experts in this field?

What are the top conferences in the field?

What are the best papers?

What are the top research labs?
Examples – Citation network analysis

Researcher B

- an in-depth understanding of the research field?

An Inverted Index Implementation

Introduction of Modern Information Retrieval

Self-Indexing Inverted Files for Fast Text Retrieval

Efficient Document Retrieval in Main Memory

Filtered Document Retrieval with Frequency-Sorted Indexes

Memory Efficient Ranking

A Document-centric Approach to Static Index Pruning in Text Retrieval Systems

Static Index Pruning for Information Retrieval Systems

Signature les: An access Method for Documents and its Analytical Performance Evaluation

Parameterised Compression for Sparse Bitmaps

Vector-space Ranking with Effective Early Termination

Topics

- Topic 1: Theory
- Topic 27: Information retrieval
- Topic 23: Index method
- Topic 21: Framework
- Topic 34: Parallel computing
- Other

Citation Relationship Type

- Basic theory
- Comparable work
- Other
Examples – Conference Suggestion

Latent Dirichlet Co-Clustering

M. Mahdi Shafiei and Evangelos E. Milios
Faculty of Computer Science, Dalhousie University
6050 University Ave., Halifax, Canada
shafiei@cs.dal.ca, eem@cs.dal.ca

Abstract

We present a generative model for simultaneously clustering documents and terms. Our model is a four-level hierarchical Bayesian model, in which each document is modeled as a mixture of document topics, where each topic is a distribution over some segments of the text. Each of these segments in the document can be modeled as a mixture of word topics where each topic is a distribution over words. We present efficient approximate inference techniques based on Markov Chain Monte Carlo method and a Moment-Matching algorithm for empirical Bayes parameter estimation. We report results in document modeling, document and term clustering, comparing to other topic models. Clustering and Co-Clustering algorithms including Latent Dirichlet Allocation (LDA), Model-based Overlapping Clustering (MOCC), Model-Based Overlapping Co-Clustering (MOCC), and Information-Theoretic Co-Clustering (ITCC).

1 Introduction

Finding the appropriate representation model for text data has been one of the main issues for the data mining community since it started to look at the problem of processing text automatically. The "bag-of-words" representation is the basic and most widely used representation method for textual data [19]. In this approach, the order of words at which they appear in documents is ignored and only the word frequencies are taken into account. But this approach has been criticized for several reasons. Among those, it provides a relatively high dimensional representation of data (equal to the dictionary size) which causes curse of dimensionality problem [19]. Furthermore, it does not consider synonymy and polysemy relations of words in natural language. It has been also criticized of losing information due to its ignorance of word order. Various preprocessing steps such as removing stop-words and stemming have been used to reduce dimensionality, create and select better features.

To overcome the high dimensionality issue of the bag-of-words representation, several dimension reduction methods have been proposed. Feature selection methods select a subset of words to reduce the dimensionality. Feature transformation methods try to tackle not only the high dimensionality problem of "bag-of-words" representation, but also consider synonymy and polysemy as well. Latent Semantic Indexing (LSI) [6] is one of these approaches which use singular value decomposition to identify a linear subspace in the original space of features. It is believed that the resulting new features also capture the two mentioned properties of natural language - polysemy and synonymy.

But the problem with most Cartesian space representation approaches for text like LSI is their inability to provide in terpretative components. Despite some work on interpreting the dimensions generated by these methods [5], these approaches are still far from providing a natural interpretation in the case of text. Topic models, on the other hand, are class of statistical models in which the semantic properties of words and documents are expressed in terms of probabilistic topics. Probabilistic topic modeling as a way of representing the content of words and documents has the distinct advantage that each topic is individually interpretable, providing a probability distribution over words that pick out a coherent cluster of correlated terms. The major difference between Cartesian space methods like LSI and statistical topic models is that LSI family methods claim the words and documents can be represented as points in the Euclidean space whereas for the topic models, this is not the case.

One common assumption among most statistical model for language is still the bag-of-words assumption. In these models, no assumption is made about the order of words. In other words, while this family of methods tries to deal with the two first issues of bag-of-words representation, high dimensionality and ignoring polysemy and synonymy properties, it still keeps the "bag-of-words" assumption intact. Recently, there has been increased research interest in models sensitive to this kind of information [18].
Examples – Reviewer Suggestion

KDD Committee

conference

Who are best matching reviewers for each paper?

Paper content
Our Social Network is Black White

Social network without role/relationship info, e.g. a company’s email network

Latent relationship graph

How to infer

CEO
Manager
Employee

Fortunately, user interactions form *implicit groups*
From BW to Color
Expertise Search

Finding experts, expertise conferences, and expertise papers for “information retrieval”
Course Search

Finding courses for “data mining”
Association Search

Finding associations between persons
- high efficiency
- Top-K associations

Usage:
- to find a partner
- to find a person with same interests
Sub-Graph Search

Sub graphs
200 topics have been discovered automatically from the academic network
Academic Performance Measurement

Academic Statistics

Top 3 by H-index:

1. Hector Garcia-Molina (Prof)
   - H-index: 95
   - Papers: 277
   - Highest cited papers: 72
   - Homepage: http://www
   - Expertise: Semantic web / Hypermedia Systems

2. Christos H. Papadimitriou
   - H-index: 90
   - Papers: 175
   - Highest cited papers: 63
   - Homepage: http://www
   - Expertise: Optimal Parallel Algorithms

3. Anil K. Jain (Distinguished Prof)
   - H-index: 89
   - Papers: 159
   - Highest cited papers: 32
   - Homepage: http://www
   - Expertise: Object Recognition / Ma

Top 3 by Citation:

1. Hector Garcia-Molina (Prof)
   - Citation: 2739
   - Top 10: 72
   - Homepage: http://www
   - Expertise: Semantic web / Hypermedia Systems

2. Christos H. Papadimitriou
   - Citation: 2037
   - Top 10: 63
   - Homepage: http://www
   - Expertise: Optimal Parallel Algorithms

3. Anil K. Jain (Distinguished Prof)
   - Citation: 1721
   - Top 10: 32
   - Homepage: http://www
   - Expertise: Object Recognition / Ma

Personal Statistics

- In Computer Science:
  - Top 3 by H-index:
    - Hector Garcia-Molina: H-index 95, 72 highest cited papers, 2739 citations
    - Christos H. Papadimitriou: H-index 90, 63 highest cited papers, 2037 citations
    - Anil K. Jain: H-index 89, 32 highest cited papers, 1721 citations
  - Top 3 by Citation:
    - Hector Garcia-Molina: 72 highest cited papers, 2739 citations
    - Christos H. Papadimitriou: 63 highest cited papers, 2037 citations
    - Anil K. Jain: 32 highest cited papers, 1721 citations

- AREA: Semantic web / Hypermedia Systems
  - Top 3 by H-index:
    - Hector Garcia-Molina: H-index 95, 72 highest cited papers, 2739 citations
    - Ravi Kumar: H-index 23, 14 highest cited papers, 1237 citations
    - Gene H. Golub: H-index 41, 10 highest cited papers, 1212 citations
  - Top 3 by Citation:
    - Hector Garcia-Molina: 72 highest cited papers, 2739 citations
    - Ravi Kumar: 14 highest cited papers, 1237 citations
    - Gene H. Golub: 10 highest cited papers, 1212 citations
Outline

• ArnetMiner: Academic Social Network

• Core Techniques
  – Knowledge Acquisition
  – Semantic Integration
  – Heterogeneous Ranking
  – Social Influence Analysis

• Demo
ArnetMiner: Overview

Modeling and Search Network

- Topic model
- Academic suggestion
- Expertise search

Social Network Analysis

- Social involution analysis
- Citation tracing analysis
- Social influence analysis

Social Network Storage

- Access interface
- Indexing
- RNKB

1. Social Network Extraction

- Integration
- Social Network Storage
- Name Disambiguation
- Profiling
- Extraction
- Homepage finding

2. Modeling and Search Network

3. Social Network Analysis

Dr. Tang

- Association
- SVM
- Publish

Lumin

- Write
- Publish

Prof. Wang

- Write
- Publish

Prof. Li

- Write
- Coauthor

Pc member

- Write
- Coauthor

Social involution analysis

Citation tracing analysis

Social influence analysis

ArnetMiner: Overview
Two questions:

- How to accurately extract the researcher profile information from the Web?
- How to integrate the information from different sources?
70.60% of the researchers have at least one homepage or an introducing page

<table>
<thead>
<tr>
<th>Percentage</th>
<th>Source</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>85.6%</td>
<td>from universities</td>
<td>14.4% from companies</td>
</tr>
<tr>
<td>71.9%</td>
<td>homepages</td>
<td>28.1% are introducing pages</td>
</tr>
<tr>
<td>40%</td>
<td>in lists and tables</td>
<td>60% are natural language text</td>
</tr>
</tbody>
</table>

There are a large number of person names having the ambiguity problem

Even 3 “Yi Li” graduated from the author’s lab

70% moved at least one time
Our Approach Picture
– based on Markov Random Field

Markov Property:

\[ P(Y_i \mid \{Y_j \mid Y_j \neq Y_i\}) \]

\[ = P(Y_i \mid \{Y_j \mid Y_j \sim Y_i\}) \]

Special cases:
- Conditional Random Fields
- Hidden Markov Random Fields

Researcher Profiling
Name Disambiguation
CT2: Semantic Integration

(IEEE TKDE, SIGMOD’09, IJCAI’09, ISWC’09)
“I’m really surprised by the good results of these years RiMOM, you can compete with the top systems that make use of such background knowledge.”

RiMOM-A Tool for Semantic Integration
(OAEI’06-09)
CT3: Topic-based Heterogeneous Ranking
(Machine Learn. J, KDD’08, ICDM’08, CIKM’09, DKE)

Search with keyword

Data mining

Search with semantic modeling

Modeling using VSM

Return

Topics

Data mining
Association Rules
Database systems
Data management
Web databases
Information systems

Return

Principles of Data Mining,
DJ Hand - Drug Safety, 2007 - drugsafety.adisonline.com

Advances in Knowledge Discovery and Data Mining
UM Fayyad, G Piatetsky-Shapiro, P Smyth, R...

Data Mining: Concepts and Techniques
J Han, M Kamber - 2001...

Modeling using VSM

Query vector
Doc vector

Experts

Expertise conferences

Expertise papers
1. How to model the heterogeneous academic network?

2. How to capture the link information for ranking objects in the academic network?
Modeling the Academic Network

Author-Conference-Topic Model [Tang et al., 08]
Integrating Topic Model into Random Walk

Modeling academic network with topics

Random walk over the academic network

Author-Conference-Topic Model [Tang et al., 08]
Combination Method 1

Stage 1: Random walk

Stage 2: Topic-based relevance

ISWC
IJCAI
WWW
Tree CRF...
EOS...
Association...
Paper Graph G_p
Author Graph G_e
Prof. Wang
Prof. Tang
Jing Zhang
Conference Graph G_c
λ_{de}
λ_{ed}
λ_{cd}
λ_{dc}
λ_{dd}

Ranking score

Combination by multiplication

\[ R[d] = r[d] \times P(q|d) \]

Topic layer

Data mining
Query

Topic-based relevance score

\[ P_{LM}(q|d) = \prod_{w \in q} \frac{N_d}{N_d + \lambda} \cdot \frac{tf(w, d)}{N_d} + (1 - \frac{N_d}{N_d + \lambda}) \cdot \frac{tf(w, D)}{N_d} \]

\[ P_{ACT}(q|d, \theta, \phi) = \prod_{w \in q} \sum_{z=1}^T P(w|z, \phi_z) \sum_{a \in A_d} P(z|a, \theta_a) P(a|d) \]

\[ P(q|a) = P_{LM}(q|a) \times P_{ACT}(q|a) \]
Combination Method 2

Query: ontology alignment

ISWC
IJCAI
WWW
Tree CRF...
EOS...
Association...
pos
owl
Web 

service
Paper Graph Gp
Author Graph Gc
Conference Graph Gc

Prof. Tang
Jing Zhang
Prof. Wang

Hidden Theme
Graph Gt

λde
λdd
λed
λdc
λq
d
λq
t
λdt
λdq

λtd

pos
owl

Web service

Transition probability

\[
P(z_i | a_j) = \theta_{a_jz_i} \\
P(a_j | z_i) = \frac{P(z_i | a_j)P(a_j)}{P(z_i)} \\
P(z_i | d_j) = \frac{1}{\lambda_d} \sum_{x \in a_d} \theta_{x z_i} \\
P(d_j | z_i) = P(w_d | z_i) = \prod_{i=1}^{N_d} P(w_{di} | z_i) \\
P(c_j | z_i) = \psi_{z_i c_j} \\
P(z_i | c_j) = \frac{P(c_j | z_i)P(z_i)}{P(c_j)} \\
P(q | z_i) = \prod_{w \in q} P(w | z_i) \\
P(z_i | q) \propto P(q | z_i)P(z_i)
\]

Ranking score

\[
r(d_i) = \lambda_{dd} \sum_{(d_j, d_k) \in E_{dd}} p(d_i | d_j) \cdot r(d_j) + \lambda_{cd} \sum_{(e_i, d_j) \in E_{cd}} P(d_i | e_i) \cdot r(e_i) \\
+ \lambda_{ed} \sum_{(e_i, d_j) \in E_{ed}} P(d_i | e_i) \cdot r(e_i) \\
+ \lambda_{dc} \sum_{(e_i, d_j) \in E_{dc}} P(d_i | e_i) \cdot r(e_i) \\
+ \lambda_{td} \sum_{(e_i, d_j) \in E_{td}} P(d_i | e_i) \cdot r(e_i) \\
+ \lambda_{td} \sum_{(e_i, d_j) \in E_{td}} P(d_i | e_i) \cdot r(e_i)
\]
Learning to Rank Experts

• Combining more information

\[
\min_{w_T} \left\{ \sum_{i=1}^{n_2} \left[ 1 - z_{T_i} \left< w_T, x_{T_i}^a - x_{T_i}^b \right> \right] + \lambda \| w_T \|^2 \right\}
\]

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1-L10</td>
<td>Low-level language model features</td>
</tr>
<tr>
<td>H1-H3</td>
<td>High-level language model features</td>
</tr>
<tr>
<td>S1</td>
<td>The year he/she published his/her first paper</td>
</tr>
<tr>
<td>S2</td>
<td>The number of papers of an expert</td>
</tr>
<tr>
<td>S3</td>
<td>The number of papers in recent 2 years</td>
</tr>
<tr>
<td>S4</td>
<td>The number of papers in recent 5 years</td>
</tr>
<tr>
<td>S5</td>
<td>The number of citations of all his/her papers</td>
</tr>
<tr>
<td>S6</td>
<td>The number of papers cited more than 5 times</td>
</tr>
<tr>
<td>S7</td>
<td>The number of papers cited more than 10 times</td>
</tr>
<tr>
<td>S8</td>
<td>PageRank score</td>
</tr>
<tr>
<td>SumCo1-8</td>
<td>The sum of coauthors' S1-S8 scores</td>
</tr>
<tr>
<td>AvgCo1-8</td>
<td>The average of coauthors' S1-S8 scores</td>
</tr>
<tr>
<td>SumStu1-8</td>
<td>The sum of his/her advisees' S1-S8 scores</td>
</tr>
<tr>
<td>AvgStu1-8</td>
<td>The average of his/her advisees' S1-S8 scores</td>
</tr>
</tbody>
</table>

Language model, BM25, tf*idf
Heterogeneous Cross-domain Ranking

Query: “data mining”

Conferences
- KDD
- SDM
- ICDM
- PAKDD

Papers
- Principles of Data Mining
- Data Mining: Concepts and Techniques

Authors
- P. Yu
- ?
- ?

Loss in one domain
\[
\min_{w_S, w_T} \left\{ \sum_{i=1}^{n_1} \left[ 1 - z_{S_i} \langle w_S, x_{S_i}^a - x_{S_i}^b \rangle \right] + C \sum_{i=1}^{n_2} \left[ 1 - z_{T_i} \langle w_T, x_{T_i}^a - x_{T_i}^b \rangle \right] + \lambda \|W\|_{2,1}^2 \right\}
\]

Loss in another domain

\[
\min_{w_S, w_T, U} \left\{ \sum_{i=1}^{n_1} \left[ 1 - z_{S_i} \langle w_S, U^T (x_{S_i}^a - x_{S_i}^b) \rangle \right] + C \sum_{i=1}^{n_2} \left[ 1 - z_{T_i} \langle w_T, U^T (x_{T_i}^a - x_{T_i}^b) \rangle \right] + \lambda \|W\|_{2,1}^2 \right\}
\]

Common feature space

P. Yu

??
Learning Algorithm

- Equivalent objective function:

\[
\min_{D,M} \left\{ \sum_{i=1}^{n} \left[ 1 - z_{S_i} \langle \alpha_1, x^a_{S_i} - x^b_{S_i} \rangle \right]_+ + C \sum_{i=1}^{n} \left[ 1 - z_{T_i} \langle a_2, x^a_{T_i} - x^b_{T_i} \rangle \right]_+ + \lambda \sum_{i=1}^{2} \langle \alpha_i, D^+ \alpha_i \rangle \right\}
\]

**Algorithm Procedure**

1. **Input:** Training set: \( L_S \cup L_T \); Test set: \( S \)
2. **Output:** Ranking function \( f_T = \langle w_T^*, x \rangle \) and the predicted preferences over test data: \( \{y_t\}_{i=1}^n \)
3. **Initialization:** \( D = \frac{I_{dxd}}{d} \)
4. **Step 1: Latent Space Finding**
   1. while not reach maximal iteration number \( T \) do
   2. \( \alpha_1 = \text{argmin} \left\{ \sum_{i=1}^{n} \left[ 1 - z_{S_i} \langle \alpha, x^a_{S_i} - x^b_{S_i} \rangle \right]_+ + \lambda \langle \alpha, D^+ \alpha \rangle \right\} \)
   3. \( \alpha_2 = \text{argmin} \left\{ \sum_{i=1}^{n} \left[ 1 - z_{T_i} \langle \alpha, x^a_{T_i} - x^b_{T_i} \rangle \right]_+ + \lambda \langle \alpha, D^+ \alpha \rangle \right\} \)
   4. \( M = [\alpha_1, \alpha_2] \)
   5. set \( D = \frac{(MM^T)^{\frac{1}{2}}}{\text{trace}(MM^T)^{\frac{1}{2}}} \)
   6. end while
7. **Step 2: Learning in Latent Space**
   1. Apply SVD decomposition on \( D \), \( D = USV^T \)
   2. Construct \( U \) by the eigenvectors corresponding to the first and second biggest eigenvalues of \( D \)
   3. \( w_T^* = \text{argmin} \left\{ \sum_{i=1}^{n} \left[ 1 - z_{S_i} \langle w, U^T(x^a_{S_i} - x^b_{S_i}) \rangle \right]_+ + C \sum_{i=1}^{n} \left[ 1 - z_{T_i} \langle w, U^T(x^a_{T_i} - x^b_{T_i}) \rangle \right]_+ + \lambda \| w \|_2^2 \right\} \)
   4. for \( i = 1 \) to \( n \) do
   5. \( y_t = \langle w_T^*, U^T x_t \rangle \)
   6. end for

- Optimize the loss function for each domain
- Common space discovery
- Optimize the weight via the common space
Experimental Results

• Data sets
  – Homogeneous Data
    • LETOR 2.0: TREC2003, TREC2004, and OHSUMED
  – Heterogeneous Data
    • Academic network consisting of 14,134 authors, 10,716 papers, and 1,434 conferences.
  – Heterogeneous Tasks
    • Expert finding vs. Bole search
• Baselines
  – RSVM
  – Language model
Results on Homogeneous Data

(a) TREC2003_TR
KL divergence = 2.4

(b) MAP and Recall for various datasets

(c) OHSUMED_TR
KL divergence = 2.20
Results on Heterogeneous Data

Table Performance of different approaches for expert finding.

<table>
<thead>
<tr>
<th>Approach</th>
<th>MAP</th>
<th>N@1</th>
<th>N@3</th>
<th>N@5</th>
<th>N@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Libra</td>
<td>.5104</td>
<td>.4800</td>
<td>.4634</td>
<td>.4467</td>
<td>.4978</td>
</tr>
<tr>
<td>Rexa</td>
<td>.4621</td>
<td>.4512</td>
<td>.4332</td>
<td>.4236</td>
<td>.4798</td>
</tr>
<tr>
<td>PSVM</td>
<td>.6024</td>
<td>.6071</td>
<td>.5920</td>
<td>.5844</td>
<td>.5985</td>
</tr>
</tbody>
</table>

Features | Weights
---|------
S1  | 2.7201
L10 | -2.5080
H2  | 2.5018
H3  | 1.9956
H1  | -1.5827
L2  | 1.5122
S4  | 1.1284
L9  | 1.0525
S2  | -0.9863
L6  | 0.6276
Results on Heterogeneous Tasks

• Expert finding verse Bole search (finding best supervisor)
• To obtain ground truth of bole for each query
  – We sent emails to 50 senior researchers and 50 junior researchers (91.6% are post doc or graduates)
  – Average their feedbacks

<table>
<thead>
<tr>
<th>Table</th>
<th>Results on heterogeneous tasks.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P@5</td>
</tr>
<tr>
<td>Approach</td>
<td></td>
</tr>
<tr>
<td>RSVM</td>
<td>.7714</td>
</tr>
<tr>
<td>RSVMt</td>
<td>.8000</td>
</tr>
<tr>
<td>MTRSVM</td>
<td>.8000</td>
</tr>
<tr>
<td>HCDRank</td>
<td>.8285</td>
</tr>
<tr>
<td>Language model</td>
<td>.6250</td>
</tr>
</tbody>
</table>
CT4: Social Influence Analysis
(KDD’10, KDD’10, KDD’09, ICDM’09, JIS)

- How to quantify the influence between users?
- What is the relationship between users?
- How to discover topic distribution over links?
- Can we predict the user’s actions?
Topic-based Social Influence Analysis

- Social network -> Topical influence network
Social Influence Sub-graph on “Data mining”
## Influential nodes on different topics

<table>
<thead>
<tr>
<th>Author</th>
<th>Topic</th>
<th>Representative Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Mining</td>
<td>Heikki Mannila, Philip S. Yu, Dimitrios Gunopulos, Jiawei Han, Christos Faloutsos, Bing Liu, Vipin Kumar, Tom M. Mitchell, Wei Wang, Qiang Yang, Xindong Wu, Jeffrey Xu Yu, Osmar R. Zaiane</td>
<td></td>
</tr>
<tr>
<td>Machine Learning</td>
<td>Pat Langley, Alex Warbel, Trevor Darrell, C. Lee Giles, Terrence J. Sejnowski, Samy Bengio, Daphne Koller, Luc De Raedt, Vasant Honavar, Floriana Esposito, Bernhard Scholzkef</td>
<td></td>
</tr>
<tr>
<td>Database System</td>
<td>Gerhard Weikum, John Mylopoulos, Michael Stonebraker, Barbara Pernici, Philip S. Yu, Sharad Mehrotra, Wei Sun, V. S. Subramaniam, Alejandro P. Buchmann, Kian-Lee Tan, Jiawei Han</td>
<td></td>
</tr>
<tr>
<td>Information Retrieval</td>
<td>Gerard Salton, W. Bruce Croft, Ricardo A. Baeza-Yates, James Allan, Yi Zhang, Mounia Lalmas, Zheng Chen, Ophir Fkeder, Alan F. Smeaton, Kong Jin</td>
<td></td>
</tr>
<tr>
<td>Web Services</td>
<td>Yan Wang, Liang-je Zhang, Shahrban Dustin, Jian Yang, Fabio Casati, Wei Xu, Zakaria Maamar, Ying Li, Xin Zhang, Boualem Benatallah, Boualem Benatallah</td>
<td></td>
</tr>
<tr>
<td>Semantic Web</td>
<td>Wolfgang Nejdl, Daniel Schwabe, Steffen Staab, Mark A. Musen, Andrew Tomkins, Juliana Freire, Carole A. Goble, James A. Hendler, Rudi Studer, Enrico Motta</td>
<td></td>
</tr>
<tr>
<td>Bayesian Network</td>
<td>Daphne Koller, Paul R. Cohen, Floriana Esposito, Henri Prade, Michael I. Jordan, Didier Dubois, David Heckerman, Philippe Smets</td>
<td></td>
</tr>
<tr>
<td>Citation</td>
<td>Fast Algorithms for Mining Association Rules in Large Databases, Using Segmented Right-Deep Trees for the Execution of Pipelined Hash Joins, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, Discovery of Multiple-Level Association Rules from Large Databases, Interleaving a Join Sequence with Semijoins in Distributed Query Processing</td>
<td></td>
</tr>
<tr>
<td>Web Services</td>
<td>The Web Service Modeling Framework WSMF, Interval Timed Coloured Petri Nets and their Analysis, The design and implementation of real-time schedulers in RED-linux, The Self-Serv Environment for Web Services Composition</td>
<td></td>
</tr>
<tr>
<td>Semantic Web</td>
<td>FaCT and iFaCT, The GRAIL concept modelling language for medical terminology, Semantic Integration of Semistructured and Structured Data Sources, Description of the RACER System and its Applications, DL-Lite: Practical Reasoning for Rich DIs</td>
<td></td>
</tr>
</tbody>
</table>
CT4: Social Influence Analysis
(KDD’10, KDD’10, KDD’09, ICDM’09, JIS)

- How to quantify the influence between users?
- What is the relationship between users?
- How to discover topic distribution over links?
- Can we predict the user’s actions?
Mining Advisor-Advisee Relationship from Research Publication Networks
Results
Application: visualization

RULE

TPFG
An example on a real system: Arnetminer

Performance improvement
Results (cont.)

### Table 1: Accuracy of prediction by $P@2, \theta$: $\frac{T}{T+F}$

<table>
<thead>
<tr>
<th>data set</th>
<th>RULE</th>
<th>SVM</th>
<th>IndMAX</th>
<th>TPGF</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEST1</td>
<td>69.9%</td>
<td>73.4%</td>
<td>75.2%</td>
<td>78.9%</td>
</tr>
<tr>
<td>TEST2</td>
<td>69.8%</td>
<td>74.6%</td>
<td>74.6%</td>
<td>79.0%</td>
</tr>
<tr>
<td>TEST3</td>
<td>80.6%</td>
<td>86.7%</td>
<td>83.1%</td>
<td>90.9%</td>
</tr>
</tbody>
</table>

### Advisee List

<table>
<thead>
<tr>
<th>Advisee</th>
<th>Top Ranked Advisor</th>
<th>Time</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2. Jiawei Han</td>
<td>2004-2008</td>
<td>PhD advisor, 2008</td>
</tr>
</tbody>
</table>

3 cited from a blog of Sergey Brin, who left Stanford to found Google around 1998.
CT4: Social Influence Analysis
(KDD’10, KDD’10, KDD’09, ICDM’09, JIS)

• How to quantify the influence between users?
• What is the relationship between users?
• How to discover topic distribution over links?
• Can we predict the user’s actions?
Examples – Topic distribution analysis over citations

Researcher A

• an in-depth understanding of the research field?

Topics

- Basic theory
- Comparable work
- Comparable work
- Other

Citation Relationship Type

- Basic theory
- Comparable work
- Comparable work
- Other
Problem: Link Semantic Analysis

Citation context words

Introduction
... Our method increases query throughput by 15% over the method proposed by Anh and Moffat [3] while still remaining rank safe results...

Related Work
... Impact-sorted indexes are described in a series of papers by Anh and Moffat [1, 2, 3]... Fagin et al. proposed a class of algorithms known as threshold algorithms [12]. These algorithms, like the ones shown in this paper, ...

Static pruning methods remove information from the index that is unlikely to affect query effectiveness. Carmel et al. considered this process [9]. More recently...

Approach
... A similar technique has been used previously by Buettcher and Clarke [7]...

... Moffat and Zobel [14] suggest a method for determining this parameter. We use a slightly different formulation in this paper...

Link semantics
Pairwise Restricted Boltzmann Machines (PRBMs)

Example
## Accuracy of Link Categorization

<table>
<thead>
<tr>
<th>Approach</th>
<th>Type</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-measure</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>drill down</td>
<td>64.09</td>
<td>53.60</td>
<td>58.24</td>
<td>88.50</td>
</tr>
<tr>
<td></td>
<td>similar</td>
<td>77.66</td>
<td>83.34</td>
<td>80.33</td>
<td>76.59</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>64.63</td>
<td>59.58</td>
<td>61.77</td>
<td>79.71</td>
</tr>
<tr>
<td></td>
<td>Avg.</td>
<td>68.79</td>
<td>65.51</td>
<td>66.78</td>
<td>81.60</td>
</tr>
<tr>
<td>SVM+LDA</td>
<td>drill down</td>
<td>66.31</td>
<td>58.56</td>
<td>61.71</td>
<td>89.24</td>
</tr>
<tr>
<td></td>
<td>similar</td>
<td>80.69</td>
<td>85.47</td>
<td>82.93</td>
<td>79.71</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>69.58</td>
<td>64.36</td>
<td>66.75</td>
<td>82.42</td>
</tr>
<tr>
<td></td>
<td>Avg.</td>
<td>72.19</td>
<td>69.46</td>
<td>70.46</td>
<td>83.79</td>
</tr>
<tr>
<td>SVM+RBM</td>
<td>drill down</td>
<td>68.63</td>
<td>53.0</td>
<td>59.66</td>
<td>89.40</td>
</tr>
<tr>
<td></td>
<td>similar</td>
<td>76.47</td>
<td>91.74</td>
<td>83.36</td>
<td>78.89</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>76.65</td>
<td>53.28</td>
<td>62.54</td>
<td>82.58</td>
</tr>
<tr>
<td></td>
<td>Avg.</td>
<td>73.92</td>
<td>66.02</td>
<td>68.52</td>
<td>83.62</td>
</tr>
<tr>
<td>gPRBM</td>
<td>drill down</td>
<td>69.77</td>
<td>65.22</td>
<td>67.42</td>
<td>88.07</td>
</tr>
<tr>
<td></td>
<td>similar</td>
<td>77.33</td>
<td>87.22</td>
<td>81.98</td>
<td>79.01</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>78.00</td>
<td>60.94</td>
<td>68.42</td>
<td>85.19</td>
</tr>
<tr>
<td></td>
<td>Avg.</td>
<td>75.03</td>
<td>71.12</td>
<td>72.61</td>
<td>84.09</td>
</tr>
<tr>
<td>dPRBM</td>
<td>drill down</td>
<td>96.00</td>
<td>44.44</td>
<td>61.54</td>
<td>95.88</td>
</tr>
<tr>
<td></td>
<td>similar</td>
<td>76.85</td>
<td>96.89</td>
<td>85.71</td>
<td>78.60</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>75.00</td>
<td>37.50</td>
<td>50.00</td>
<td>80.25</td>
</tr>
<tr>
<td></td>
<td>Avg.</td>
<td><strong>82.62</strong></td>
<td>59.61</td>
<td>65.75</td>
<td>84.91</td>
</tr>
<tr>
<td>hPRBM</td>
<td>drill down</td>
<td>65.38</td>
<td>68.00</td>
<td>66.67</td>
<td>93.00</td>
</tr>
<tr>
<td></td>
<td>similar</td>
<td>83.54</td>
<td>92.31</td>
<td>87.71</td>
<td>84.77</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>88.14</td>
<td>69.33</td>
<td>77.61</td>
<td>87.65</td>
</tr>
<tr>
<td></td>
<td>Avg.</td>
<td><strong>79.02</strong></td>
<td><strong>76.55</strong></td>
<td><strong>77.33</strong></td>
<td><strong>88.48</strong></td>
</tr>
</tbody>
</table>

Tested on Arnetminer citation data

**gPRBM**: our approach with generative learning

**dPRBM**: our approach with discriminative learning

**hPRBM**: our approach with hybrid learning
CT4: Social Influence Analysis (KDD’10, KDD’10, KDD’09, ICDM’09, JIS)

- How to quantify the influence between users?
- What is the relationship between users?
- How to discover topic distribution over links?
- Can we predict the user’s actions?
What can we do in SNS?
Social Action

Twitter

Flickr

KDD

Add favorites

Comment on Haiti Earthquake

Publish in KDD Conference
1. Always watch news
2. Enjoy sports
3. …
## Results

Table 1: Performance of action prediction with different approaches (%).

<table>
<thead>
<tr>
<th>Data set</th>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>SVM</td>
<td>10.41</td>
<td>16.71</td>
<td>13.85</td>
</tr>
<tr>
<td></td>
<td>wvRN</td>
<td>0.45</td>
<td>7.89</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>NTT-FGM</td>
<td>26.40</td>
<td>21.14</td>
<td>23.47</td>
</tr>
<tr>
<td>Flickr</td>
<td>SVM</td>
<td>34.48</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>wvRN</td>
<td>60.02</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NTT-FGM</td>
<td>56.18</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>ArnetMiner</td>
<td>SVM</td>
<td>10.19</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>wvRN</td>
<td>14.83</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NTT-FGM</td>
<td>31.14</td>
<td>44</td>
<td></td>
</tr>
</tbody>
</table>

Figure 8: Example correlation analysis between researchers. The strength represents the correlation score between two researchers.
Arnetminer Today
— A brief summary
ArnetMiner’s History

- 2006/5, V0.1 Perl-based CGI version
  - Profile extraction, person/paper/conf. search
- 2006/8, V1.0 Java (Demo @ ASWC)
  - Rewrite the above functions
- 2007/7, V2.0 (Demo @ KDD, ISWC)
  - New: survey search, research interest, association search
- 2008/4, V3.0 (Demo @ WWW)
  - Query understanding, New search GUI, log analysis
- 2008/11, V4.0 (Demo @ KDD, ICDM)
  - Graph search, topic mining, NSFC/NSF
- 2009/4, V5.0 (Demo @ KDD)
  - Bole/course search, profile editing, open resources, #citation
- 2009/12, V6.0
  - Academic statistics, user feedbacks, refined ranking
- V7.0, coming soon
  - Name disambiguation, reviewer assignment, supervisor suggestion, open API
ArnetMiner Today

* ArnetMiner data:
  > 0.6 M researcher profiles
  > 3M papers
  > 17M citation relationships
  > 5K conferences
  > 50M logs

* Visits come from more than 190 countries
* Continuously +20% increase of visits per month
* >100,000 page views per day

**Messages from Users**

... I’ve happened to visit your Arnetminer, and shocked. It was really impressive, its usefulness and your works!!! ... [from ...@selab.snu.ac.kr]

...I would first of all congratulate you on the excellent work you have done in ArnetMiner and I am much inspired... [from ...@nu.edu.pk]

**a survey by UK Southampton**

Title: Semantic Technologies for Learning and Teaching in Web 2.0.
— Thanassis Tiropanis, Hugh Davis, Dave Millard, Mark Weal

...Exposing the expertise of the institution to the outside world in order to attract funding and students. **ArnetMiner is the most representative** example of such tools at the moment...

Contextualised queries and searches, searches across repositories potentially in different departments or institutions, and matching of people for collaborative activities. **Best example** of the surveyed technologies to this end is ArnetMiner.
Opportunity: exploiting semantic web and social network in the real-world

Data Mining and Social Network techniques

Scientific Literature
Users cover >180 countries >600K researcher >3M papers
Arnetminer.org

Social search & mining
Social network Extraction Social network Mining
IBM

Advertisement
Advertisement Recommendation
Sohu

Mobile Context
Mobile search & recommendation
Nokia

Energy trend analysis
Energy product Evolution Techniques Trend
Oil Company

Large-scale Mining
Scalable algorithms for message tagging and community Discovery
Google

科技信息资源内容监测与分析服务平台 (中国科技部信息情报研究所)

Search, browsing, complex query, integration, collaboration, trustable analysis, decision support, intelligent services,
Arnetminer

PatentMiner  ScopusMiner  PubmedMiner  CheMiner
Representative Publications

- Chenhao Tan, Jie Tang, Jimeng Sun, Quan Lin, and Fengjiao Wang. Social Action Tracking via Noise Tolerant Time-varying Factor Graphs. *KDD’10*.
- Chi Wang, Jiawei Han, Yuntao Jia, Duo Zhang, Yintao Yu, Jie Tang, Jingyi Guo. Mining Advisor-Advisee Relationships from Research Publication Networks. *KDD’10*.
- Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social Influence Analysis in Large-scale Networks. *KDD’09*.
- Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. *KDD’08*.
- Jie Tang, Hang Li, Yunbo Cao, and Zhaohui Tang. Email Data Cleaning. *KDD’05*.
- Jie Tang, Ho-fung Leung, Qiong Luo, Dewei Chen, and Jibin Gong. Towards Ontology Learning from Folksonomies. *IJCAI’09*.
- Chonghui Zhu, Jie Tang, Hang Li, Hwee Tou Ng, and Tiejun Zhao. A Unified Tagging Approach to Text Normalization. *ACL’07*.

Others: ICDM’07-09, CIKM’07-09, SDM’09, ISWC’06, DKE, JIS, etc.
Thanks!

Demo: http://arnetminer.org
HP: http://keg.cs.tsinghua.edu.cn/persons/tj/