A Miner.org
—Deep Analysis and Mining for Academic Social Networks
(http://aminer.org)

Jie Tang
Knowledge Engineering Group
Department of Computer Science and Technology
Tsinghua University
The era of big and complex data...

• Web trend
  – Data centric $\rightarrow$ User centric
  – Offline sparse social $\rightarrow$ Online dense social
  – Large-scale data mining $\rightarrow$ Big data deep analytic

• Tech trend
  – Formal text $\rightarrow$ Informal text
  – User modeling $\rightarrow$ Collective intelligence
  – Keyword-based modeling $\rightarrow$ Semantic-based modeling
  – Macro-level analysis $\rightarrow$ Micro-level analysis
  – …
Why AMiner.org (Arnetminer)?

"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?
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 textual 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 higher 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 to 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 interpretable 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 it 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

Who are best matching reviewers for each paper?

Who will accept the invitation to review the paper

KDD Committee
Conference/Journal

Paper content
Outline

• ArnetMiner: Academic Social Network

• Core Techniques
  – Knowledge Acquisition
  – Modeling and Heterogeneous Ranking
  – Social Network Analysis
AMiner.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, IEEE TKDE, J. Informetrics, KDD’08-12, WWW’12, SIGMOD’09, IJCAI’09

http://aminer.org/
- Researcher profile extraction
- Expert finding
- Social network search
- Topic browser
- Conference analysis
- ArnetApp platform
Expertise Search

Finding experts, expertise conferences, and expertise papers for “data mining”
Organization Search

Ranking organizations on "machine learning"
Geographic search

Finding the most hot regions on “data mining”
Conference Analysis

Which year is the most successful in the KDD's history?

Who are the most cited authors?

What is author distribution for the highly cited KDD papers in the past years?
Reviewer Suggestion

Interest matching
COI avoiding
Load balancing
Forecast review quality
Cross-domain Collaboration Recommendation

What kinds of topics should you collaborate with the target domain?

Who are the best collaborators on each topic?
200 topics have been discovered automatically from the academic network.
Academic Performance Measure

Academic Statistics

New Stars

Top 3 by H-index:
1. Anil K. Jain (Distinguished Professor, Michigan State University)
   - H-index: 112, Papers: 368, Citation: 65611
   - Homepage: http://www.cs.msu.edu/~jain/
   - Expertise: Object Recognition / Two-View Motion Estimation, Face recognition / Image analysis
2. H. Garcia (Professor, Departments of Computer Science and Electrical Engineering, Stanford University)
   - H-index: 107, Papers: 412, Citation: 45542
   - Homepage: http://www.stanford.edu/people/hector.html
   - Expertise: Real-Time Systems / Automated Software Test Data, XML Data, Database Systems, Dynamic Networks / Extended Abstract
3. D. Papadias (Professor, Computer Science Division University of California at Berkeley)
   - H-index: 100, Papers: 356, Citation: 54668
   - Homepage: http://www.cs.berkeley.edu/~christos/
   - Expertise: Mechanism design / Learning Stochastic Finite Automata, Communication Complexity / Lower Bounds, Approximation Algorithms / Perfect Graphs, Finite Languages / Database Queries

Top 3 by Citation:
1. L. Zadeh (Professor Emeritus, Graduate School, Computer Science Division Department of Electrical Engineering and Computer Sciences University of California)
   - H-index: 34, Papers: 109, Citation: 86834
   - Homepage: http://www.cs.berkeley.edu/~zadeh/
   - Expertise: Fuzzy system, Rough Set, Intelligent Information Systems, Neural Network Learning
2. Anil K. Jain (Distinguished Professor, Michigan State University)
   - H-index: 112, Papers: 368, Citation: 65611
   - Homepage: http://www.cs.msu.edu/~jain/
   - Expertise: Object Recognition / Two-View Motion Estimation, Face recognition / Image analysis
3. Han Daekishan (Professor, Department of EECS Massachusetts Institute of Technology)
   - H-index: 89, Papers: 142, Citation: 60965
   - Homepage: http://www.csl.uiuc.edu/~han/
AMiner (ArnetMiner)

- Academic Social Network Analysis and Mining system—Aminer (http://arnetminer.org)
  - Online since 2006
  - >1 million researcher profiles
  - >131 million requests
  - >2.35 Terabyte data
  - 100K IP access from 170 countries per month
  - 10% increase of visits per month
- Deep SN analysis, mining, and search
User Distribution

4.04 million IP from 220 countries/regions
User Distribution

4.04 million IP from 220 countries/regions

Top 10 countries

1. USA
2. China
3. Germany
4. India
5. UK
6. Canada
7. Japan
8. Spain
9. France
10. Italy
Outline

• ArnetMiner: Academic Social Network
• Core Techniques
  – Knowledge Acquisition
  – Modeling and Heterogeneous Ranking
  – Social Network Analysis
Two questions:

• How to accurately extract the researcher profile information from the Web?
• How to integrate the information from different sources?
Researcher Profile Database

Extracted more than 1,000,000 researcher profiles from the Web

RiMOM[1-3]-A Tool for Semantic Integration (OAEI’06-09)

Benchmark Results

Chair Message:
“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.”

http://keg.cs.tsinghua.edu.cn/project/RiMOM/


### Name Disambiguation\textsuperscript{[1,2]}

<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jing Zhang (26)</td>
<td>Shanghai Jiao Tong Univ.</td>
</tr>
<tr>
<td></td>
<td>Yunnan Univ.</td>
</tr>
<tr>
<td></td>
<td>Tsinghua Univ.</td>
</tr>
<tr>
<td></td>
<td>Alabama Univ.</td>
</tr>
<tr>
<td></td>
<td>Univ. of California, Davis</td>
</tr>
<tr>
<td></td>
<td>Carnegie Mellon University</td>
</tr>
<tr>
<td></td>
<td>Henan Institute of Education</td>
</tr>
</tbody>
</table>

#### Jing Zhang

List of publications from the DBLP Bibliography Server - FAQ

- How to perform the assignment automatically?
- How to estimate the person number?


Disambiguation Performance
Outline

• ArnetMiner: Academic Social Network
• Core Techniques
  – Knowledge Acquisition
  – Modeling and Heterogeneous Ranking
  – Social Network Analysis
CT2: Topic-based Heterogeneous Ranking (KDD’08, KDD’12, Machine Learning)

Search with keyword

Modeling using VSM

Return

Search with semantic modeling

Modeling using semantic topics

Return

Data mining

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…
1. How to **model** the heterogeneous academic network?

2. How to **capture** the link information for ranking objects in the academic network?
(1) Modeling the Academic Network\cite{1}

Author-Conference-Topic Model

We present a generative model for clustering documents and terms. Our model is a four hierarchical Bayesian model. We present efficient inference techniques based on Markov Chain Monte Carlo. We report results in document modeling, document and terms clustering …
(2) Integrating with Random Walk[1]

Modeling academic network with topics

Random walk over the academic network

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

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 Heterogeneous Data

Table: Performances 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>.9084</td>
<td>.6071</td>
<td>.5920</td>
<td>.5844</td>
<td>.5885</td>
</tr>
</tbody>
</table>

![Graph showing weight values vs feature IDs]

<table>
<thead>
<tr>
<th>Features</th>
<th>Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>2.7201</td>
</tr>
<tr>
<td>L10</td>
<td>-2.5080</td>
</tr>
<tr>
<td>H2</td>
<td>2.5018</td>
</tr>
<tr>
<td>H3</td>
<td>1.9956</td>
</tr>
<tr>
<td>H1</td>
<td>-1.5827</td>
</tr>
<tr>
<td>L2</td>
<td>1.5122</td>
</tr>
<tr>
<td>S4</td>
<td>1.1284</td>
</tr>
<tr>
<td>L9</td>
<td>1.0525</td>
</tr>
<tr>
<td>S2</td>
<td>-0.9863</td>
</tr>
<tr>
<td>L6</td>
<td>0.6276</td>
</tr>
</tbody>
</table>
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>Approach</th>
<th>P@5</th>
<th>P@10</th>
<th>P@15</th>
<th>MAP</th>
<th>N@5</th>
<th>N@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSVM</td>
<td>.7714</td>
<td>.8429</td>
<td>.8285</td>
<td>.7756</td>
<td>.5545</td>
<td>.5947</td>
</tr>
<tr>
<td>RSVMt</td>
<td>.8000</td>
<td>.8286</td>
<td>.8476</td>
<td>.7837</td>
<td>.5923</td>
<td>.5999</td>
</tr>
<tr>
<td>MTRSVVM</td>
<td>.8000</td>
<td>.8286</td>
<td>.8476</td>
<td>.7875</td>
<td>.6140</td>
<td>.6075</td>
</tr>
<tr>
<td>HCDRank</td>
<td>.8285</td>
<td>.7857</td>
<td>.8571</td>
<td>.7971</td>
<td>.6189</td>
<td>.6112</td>
</tr>
<tr>
<td>Language model</td>
<td>.6250</td>
<td>.6875</td>
<td>.6500</td>
<td>.6726</td>
<td>.3343</td>
<td>.3809</td>
</tr>
</tbody>
</table>
Outline

• ArnetMiner: Academic Social Network
• Core Techniques
  – Knowledge Acquisition
  – Modeling and Heterogeneous Ranking
  – Social Network Analysis
CT3: Social Network Analysis (KDD’09, ’10, ’11, DMKD’12, JIS)

- **User modeling**: can we model and forecast users’ behaviors?
- **Influence**: how to quantify the influence between users?
- **Tie**: what is the relationship between users?
- **Community**: which (core) communities determine the evolution of the network structure?
(1) Social Action Modeling and Prediction[1]

**Action Prediction:**
Will John submit a paper to KDD’12?

**Personal attributes:**
1. Interested in data mining
2. Collaborated with Jiawei Han
3. ....

---

(2) Social Influence Analysis \[1\]

**Input:** coauthor network

**Social influence analysis**

**Output:** topic-based social influences

Several key challenges:

- How to differentiate the social influences from different topics?
- How to incorporate both structure and content into a unified model?
- How to estimate the model on real-large networks?

<table>
<thead>
<tr>
<th>Author</th>
<th>Dataset</th>
<th>Topic</th>
<th>Representative Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<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>
</tr>
<tr>
<td></td>
<td></td>
<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 Scholkmef</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Database System</td>
<td>Gerard Weikum, John Mylopoulos, Michael Stonebraker, Barbara Pernici, Philip S. Yu, Sharad Mehrotra, Wei Sun, V. S. Subrahmanian, Alejandro P. Buchmann, Kian-Lee Tan, Jiawei Han</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Information Retrieval</td>
<td>Gerard Salton, W. Bruce Croft, Ricardo A. Baeza-Yates, James Allan, Yi Zhang, Mounia Lalmas, Zheng Chen, Ophir Frieder, Alan F. Smeaton, Hong Jin</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Web Services</td>
<td>Yan Wang, Liang-je Zhang, Schahram Dustdar, Jian Yang, Fabio Casati, Wei Xu, Zakaria Maamar, Ying Li, Xin Zhang, Boualem Benatallah, Boualem Benatallah</td>
</tr>
<tr>
<td></td>
<td></td>
<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>
</tr>
<tr>
<td></td>
<td></td>
<td>Bayesian Network</td>
<td>Daphne Koller, Paul R. Cohen, Floriana Esposito, Henri Prade, Michael I. Jordan, Didier Dubois, David Heckerman, Philippe Smets</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data Mining</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>
</tr>
<tr>
<td></td>
<td></td>
<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>
</tr>
<tr>
<td></td>
<td></td>
<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>
</tr>
</tbody>
</table>
(3) Mining Advisor-Advisee Relationship\cite{Wang2010}

**Input: Temporal collaboration network**

**Output: Relationship analysis**

Output: potential types of relationships and their probabilities:

\[(\text{type, prob, } [\text{s\_time, e\_time}])\]

---

Results

- DBLP data: 654, 628 authors, 1,076,946 publications, years provided.
- Ground truth: MathGenealogy Project; AI Genealogy Project; Faculty Homepage

<table>
<thead>
<tr>
<th>Datasets</th>
<th>RULE</th>
<th>SVM</th>
<th>IndMAX</th>
<th>Model 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEST1</td>
<td>69.9%</td>
<td>73.4%</td>
<td>75.2%</td>
<td>80.2%</td>
</tr>
<tr>
<td>TEST2</td>
<td>69.8%</td>
<td>74.6%</td>
<td>74.6%</td>
<td>81.5%</td>
</tr>
<tr>
<td>TEST3</td>
<td>80.6%</td>
<td>86.7%</td>
<td>83.1%</td>
<td>88.8%</td>
</tr>
</tbody>
</table>

- heuristics
- Supervised learning
- Empirical optimized parameter
Results
Pareto Principle: Less than 1% of the Twitter users (e.g. entertainers, politicians, writers) produce 50% of its content, while the others (e.g. fans, followers, readers) have much less influence and completely different social behavior.

Kernel users and ordinary users exhibit very different behaviors.

Approach (Greedy & WeBA)

- Challenges
  - How to identify kernel members, and
  - How to determine the structural of community kernels.

- Formalize the problem into an optimization problem.
  - Proposed two new algorithms
    - Greedy
      - Maximum cardinality search (MCS)
      - Runs in linear time
    - Global Relaxation
      - Random walk (Annealing)
      - Theoretical error bound
      - Almost linear time.
## Results on Coauthor & Wikipedia

- On average, \textsc{WeBA} improves Precision by \textbf{340\% (wiki)} and \textbf{70\% (coauthor)}, and improves Recall by \textbf{130\% (wiki)} and \textbf{41\% (coauthor)}.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>wiki</td>
<td>coauthor</td>
</tr>
<tr>
<td>\textsc{Talk}</td>
<td>0.061</td>
<td>0.502</td>
</tr>
<tr>
<td>\textsc{User}</td>
<td>0.085</td>
<td>0.342</td>
</tr>
<tr>
<td>\textsc{AI}</td>
<td>0.573</td>
<td>0.573</td>
</tr>
<tr>
<td>\textsc{NC}</td>
<td>0.573</td>
<td>0.573</td>
</tr>
<tr>
<td>\textsc{Average}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>\textsc{LSP}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>\textsc{d-LSP}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>\textsc{p-LSP}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>\textsc{METIS+MQI}</td>
<td>0.847</td>
<td>0.055</td>
</tr>
<tr>
<td>\textsc{LOUVAIN}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>\textsc{NEWMAN1}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>\textsc{NEWMAN2}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>\textsc{α-β}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>\textsc{WEBA}</td>
<td>0.456</td>
<td>0.852</td>
</tr>
<tr>
<td>\textsc{GREEDY}</td>
<td>0.334</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
EFFICIENCY — TWITTER & COAUTHOR

465,023 nodes, 833,590 edges

822,415 nodes, 2,928,360 edges
Arnetminer Today
— A brief summary
## ArnetMiner’s History

<table>
<thead>
<tr>
<th>Date</th>
<th>Version</th>
<th>New Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006/5</td>
<td>V0.1</td>
<td>Profile extraction, person/paper/conf. search</td>
</tr>
<tr>
<td>2006/8</td>
<td>V1.0</td>
<td>Rewritten all codes in Java.</td>
</tr>
<tr>
<td>2007/7</td>
<td>V2.0</td>
<td>Survey search, research interest, association search</td>
</tr>
<tr>
<td>2008/4</td>
<td>V3.0</td>
<td>Query understanding, New search GUI, log analysis</td>
</tr>
<tr>
<td>2008/11</td>
<td>V4.0</td>
<td>Graph search, topic mining, NSFC/NSF</td>
</tr>
<tr>
<td>2009/4</td>
<td>V5.0</td>
<td>Bole/course search, profile editing, open resources, #citation</td>
</tr>
<tr>
<td>2009/12</td>
<td>V6.0</td>
<td>Academic statistics, user feedbacks, refined ranking</td>
</tr>
<tr>
<td>2010/5</td>
<td>V7.0</td>
<td>Name disambiguation, reviewer assignment, supervisor suggestion, open API</td>
</tr>
<tr>
<td>2010/7</td>
<td>V8.0</td>
<td>ArnetApp Platform</td>
</tr>
<tr>
<td>2011/7</td>
<td>V II</td>
<td>AMiner, location search, conference analysis</td>
</tr>
<tr>
<td>2012</td>
<td>V 1.0</td>
<td>New UI, cross-domain collaboration</td>
</tr>
</tbody>
</table>
Widely used..

- The largest publisher: Elsevier
- Conferences
  - KDD 2010
  - KDD 2011
  - KDD 2012
  - WSDM 2011
  - ICDM 2011
  - ICDM 2012
  - SocInfo 2011
  - ICMLA 2011
  - WAIM 2011
  - etc.
ArnetApp Platform
---to deploy your apps on Arnetminer.org

To customize the search on Arnetminer.
Arnetminer as a platform…

<table>
<thead>
<tr>
<th>Arnetminer</th>
</tr>
</thead>
</table>

- Mining knowledge from patents:
  - competitor analysis
  - company search
  - patent summarization

<table>
<thead>
<tr>
<th>PatentMiner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mining knowledge from patents:</td>
</tr>
<tr>
<td>• competitor analysis</td>
</tr>
<tr>
<td>• company search</td>
</tr>
<tr>
<td>• patent summarization</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>QQMiner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mining “QQ”</td>
</tr>
<tr>
<td>• Association search</td>
</tr>
<tr>
<td>• Influence analysis</td>
</tr>
<tr>
<td>• Hot topic detection</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PubmedMiner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mining Pubmed data</td>
</tr>
<tr>
<td>• Expert finding</td>
</tr>
<tr>
<td>• Ranking subgraphs</td>
</tr>
<tr>
<td>• Novel search</td>
</tr>
<tr>
<td>• Instant search</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mining more data…</td>
</tr>
</tbody>
</table>
PatentMiner
Patent Search

Patents on “data mining”

Related sub topics

Inventor

Company

Summary of "data mining":

Result Categories (U.S. Class)

Taps: left click to focus, right click to return

All issues

<table>
<thead>
<tr>
<th>Mining or site disintegration of hard material</th>
<th>Hydraulic and earth engineering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard material disintegrating machines</td>
<td>Earth treatment or control</td>
</tr>
<tr>
<td>Processes</td>
<td></td>
</tr>
<tr>
<td>Cutter tooth or tooth head</td>
<td>In situ conversion of solid to fluid</td>
</tr>
</tbody>
</table>

Related sub topics

Inventor

Company
PatentMiner Today

* Patent data:
  > 3.8M patents
  > 2.4M inventors
  > 400K companies
  > 10M citation relationships

* Journal data:
  > 2k journal papers
  > 3.7k authors

The crawled data is increasing to >300 Gigabytes.
Opportunity: exploiting social network and semantic web in the real-world

Web, relational data, ontological data, social data

Scientific Literature
- Users cover >180 countries
- >600K researcher
- >3M papers
- Arnetminer.org (NSFC, 863)

Social search & mining
- Social extraction
- Social mining
- IBM US, Tencent
- IBM CRL

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

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

- Juanzi Li, Jie Tang, Yi Li, Qiong Luo. RiMOM: A Dynamic Multi-Strategy Ontology Alignment Framework. IEEE *TKDE*, 2009. (Top 6 cited papers among TKDE 2009’s papers)
- Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. Cross-domain Collaboration Recommendation. *KDD’12* (Full Presentation & Best Poster Award)
- 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, Jie Tang, Duo Zhang, Yintao Yu, 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*. (Top 4 cited papers among KDD 2009's papers)
- Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. *KDD’08*. (Top 6 cited papers among KDD 2008’s papers)
- Jie Tang, Ho-fung Leung, Qiong Luo, Dewei Chen, and Jibin Gong. Towards Ontology Learning from Folksonomies. *IJCAI’09*.
Thanks!

Demo: [http://arnetminer.org](http://arnetminer.org)