Data Mining: Dynamic Past and Promising Future

Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign
www.cs.uiuc.edu/homes/hanj
April 30, 2010
Retrospective: Impressive Developments

- In-depth Investigations on **pattern mining**
  - Frequent and discriminative patterns, sequential & graph patterns, colossal patterns, approximate patterns, interestingness & pattern quality, constraint-based, pattern usage (for classification and clustering)
- New methodologies for **classification and model construction**
  - Efficiency, effectiveness and high-dimensionality, SVM, graph regularization, pattern-based classification, graph/network classification
- New methodologies for **data clustering**
  - Subspace clustering, density-based clustering, pattern-based clustering, constraint-based clustering, various outlier analysis methods
- New kinds of **data challenges**
  - Streams, time-series, sequence data, graphs, text and web, social and information networks, massive scaled/parallel data mining
- **Publicity, education, and social impacts**
  - Wide acceptance + wide concern, privacy and security in data mining, discipline confluence, conferences & books, industry (Google \rightarrow Netflix)
What Things Should Have Been Done Better?

- Data mining in science and engineering
  - Direction influenced much by the “same” industry?
  - Lots of real challenging data and problems in science, engineering, ...
  - Work with domain experts and broad applications
- Information networks: Re-examine links and new mining methodology
  - Google and Flickr lessons: Links and massive collaboration
- Integration of data obtained from new technologies
  - Besides Flash/PCM storage, tera-core systems, clouding computing, ...
  - Work on spatiotemporal data, GPS, moving objects, multimedia (video, image and text data) and their connections and integrations
- Discipline confluence
  - Statistics, machine learning, database and data warehousing, high-performance computing, Web, IR, NLP, ...
  - Multi-disciplinary education
- Applications, applications, and applications!
Integrating Clustering with Ranking for Heterogeneous Information Network Analysis (EDBT’09/KDD’09/VLDB’09 demo)

Global Ranking? VS.

Clustering in heterogeneous network?

What feature can I use?

Cluster and rank people or events
Find highly suspicious groups/events

DBLP

<table>
<thead>
<tr>
<th>DB</th>
<th>Network</th>
<th>AI</th>
<th>Theory</th>
<th>IR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>VLDB</td>
<td>INFOCOM</td>
<td>AIAMAS</td>
<td>SODA</td>
</tr>
<tr>
<td>2</td>
<td>ICDE</td>
<td>SIGMETRICS</td>
<td>AAIAI</td>
<td>STOC</td>
</tr>
<tr>
<td>3</td>
<td>SIGMOD</td>
<td>ICNP</td>
<td>Agents</td>
<td>FOCS</td>
</tr>
<tr>
<td>4</td>
<td>KDD</td>
<td>SIGCOMM</td>
<td>AAAI/AAAI</td>
<td>ICALP</td>
</tr>
<tr>
<td>5</td>
<td>ICDM</td>
<td>MOBICOM</td>
<td>ECAI</td>
<td>CCC</td>
</tr>
<tr>
<td>6</td>
<td>EDBT</td>
<td>ICDCS</td>
<td>RoboCup</td>
<td>SPA</td>
</tr>
<tr>
<td>7</td>
<td>DASFAA</td>
<td>NETWORKING</td>
<td>IAT</td>
<td>PODC</td>
</tr>
<tr>
<td>8</td>
<td>PODS</td>
<td>MobiHoc</td>
<td>ICMAS</td>
<td>APPROX.RANDOM</td>
</tr>
<tr>
<td>9</td>
<td>SSDBM</td>
<td>ISCC</td>
<td>CP</td>
<td>EUROCRYPT</td>
</tr>
<tr>
<td>10</td>
<td>SDM</td>
<td>SenSys</td>
<td></td>
<td>CIVR</td>
</tr>
</tbody>
</table>
RankCompete: A Competing Random Walk Model for Information Network Analysis [WWW’10]

- Allow multiple random walkers in the same network to compete for integrated clustering and ranking
- A faster and more intuitive way to group network nodes
- Effectively refine image retrieval results and summarize personal photo collections in a photo network

Automatically cluster a set of photos and find the most representative (i.e., highly ranked) photo in each cluster!
Outlook: Major Challenges

- Integrating, understanding and mining data, anywhere, anytime
  - Data extraction, validation, integration by data mining
- Mining Information networks
  - Exploring inter-related data, heterogeneous information networks
  - Connected data are much smarter than isolated ones: Google hints!
- Mining cyber-physical networks
  - Sensors, video/audio devices (with dynamic, spatiotemporal data) connected with information networks
  - Real-time, dynamic exploration and action (robots)
- Taming the web by data mining
  - Mining web structures, automatic construction of integrated information repositories
- Data mining: ubiquitous and invisible functions in everyday life!