MINING HETEROGENEOUS INFORMATION NETWORKS

Yizhou Sun
College of Computer and Information Science
Northeastern University
yzsun@ccs.neu.edu

September 2, 2013
Dissertation Title:
Mining Heterogeneous Information Networks

• Computer Science, University of Illinois at Urbana-Champaign, 2012

• Advisor: Professor Jiawei Han

• Doctoral Committee:
  • Professor Jiawei Han, Chair & Director of Research, UIUC
  • Professor ChengXiang Zhai, UIUC
  • Professor Dan Roth, UIUC
  • Dr. Charu C. Aggarwal, IBM T.J. Watson Research Center
Information Networks Are Everywhere

They are all treated as Homogeneous Networks!

Social Networking Websites

Biological Network: Protein Interaction

Research Collaboration Network

Product Recommendation Network via Emails
Homogeneous Networks

• Single object type and single link type
  • Link analysis based applications

Ranking web pages [Brin and Page, 1998]

Clustering books about politics [Newman, 2006]

Link Prediction [Kleinberg, 2003]
Heterogeneous Networks

- Multiple object types and/or multiple link types

1. Homogeneous networks are **information loss** projection of heterogeneous networks!
2. New problems are emerging in heterogeneous networks!

**Directly Mining information richer heterogeneous networks**
Heterogeneous Networks Are Ubiquitous

- **Healthcare**
  - Doctor, patient, disease, treatment

- **Online source code repository**
  - Project, developer, programming language, project category

- **E-Commerce**
  - Seller, buyer, product, review

- **News**
  - Person, organization, location, text
Major Contributions

New Problems, Models & Efficient Algorithms for Mining Heterogeneous Information Networks

A research monograph published in 2012: “Mining Heterogeneous Information Networks: Principles and Methodologies”

Network Type

Dynamic Network

Link + Attribute

Link Only

Mining Functions

Clustering

Ranking

Classification

Similarity Search

Relationship Prediction

Relation Strength Learning

KDD’09

EDBT’09

ICDM’09

MLG’10

EDBT’10

WSDM’12

ASONAM’11

VLDB’12

ECML/PKDD’10

VLDB’11

CIKM’09

KDD’12
### What Can be Mined from Heterogeneous Networks?

#### DBLP: A Computer Science bibliographic database

A sample publication record in DBLP (>1.8 M papers, >0.7 M authors, >10 K venues)

<table>
<thead>
<tr>
<th>Knowledge hidden in DBLP Network</th>
<th>Mining Functions</th>
<th>Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>How are CS research areas <strong>structured</strong>?</td>
<td>Clustering</td>
<td>EDBT’09, KDD’09, ICDM’09</td>
</tr>
<tr>
<td>Who are the <strong>leading</strong> researchers on Web search?</td>
<td>Ranking</td>
<td>EDBT’09, KDD’09,</td>
</tr>
<tr>
<td>Who are the <strong>peer</strong> researchers of Jure Lescovec?</td>
<td>Similarity Search</td>
<td>VLDB’11</td>
</tr>
<tr>
<td>Whom <strong>will</strong> Christos Faloutsos <strong>collaborate with</strong> in the future?</td>
<td>Relationship Prediction</td>
<td>ASONAM’11</td>
</tr>
<tr>
<td>Whether <strong>will</strong> an author <strong>publish</strong> a paper in KDD, and <strong>when</strong>?</td>
<td>Relationship Prediction with Time</td>
<td>WSDM’12</td>
</tr>
<tr>
<td>Which types of <strong>relationships</strong> are most <strong>influential</strong> for an author to decide her topics?</td>
<td>Relation Strength Learning</td>
<td>VLDB’12, KDD’12</td>
</tr>
</tbody>
</table>
Outline

- Why Heterogeneous Information Networks?
- Dissertation Overview
  - Ranking-Based Clustering
  - Meta-Path-Based Similarity Search and Mining
  - User-Guided Relation Strength-Aware Mining
- Conclusion and Future Plan
Principles of Mining Heterogeneous Information Networks

• **Principle 1: Use Holistic Network Information**
  - Study information propagation across different types of objects and links

• **Principle 2: Explore Network Meta Structure**
  - Meta-path-based similarity search and mining

• **Principle 3: User-Guided Exploration**
  - Relation strength-aware mining with user guidance
Outline

• Why Heterogeneous Information Networks?

• Dissertation Overview
  • Ranking-Based Clustering
  • Meta-Path-Based Similarity Search and Mining
  • User-Guided Relation Strength-Aware Mining

• Conclusion and Future Plan
RankClus [EDBT’09]: Ranking-based Clustering on Bi-Typed Networks

RankClus Algorithm Illustration

<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>AAMAS</td>
<td>SODA</td>
</tr>
<tr>
<td>2</td>
<td>ICDE</td>
<td>SIGMETRICS</td>
<td>IJCAI</td>
<td>STOC</td>
</tr>
<tr>
<td>3</td>
<td>SIGMOD</td>
<td>ICNP</td>
<td>AAAI</td>
<td>FOCS</td>
</tr>
<tr>
<td>4</td>
<td>KDD</td>
<td>SIGCOMM</td>
<td>Agents</td>
<td>ICALP</td>
</tr>
<tr>
<td>5</td>
<td>ICDM</td>
<td>MOBICOM</td>
<td>AAAI/IAAI</td>
<td>CCC</td>
</tr>
<tr>
<td>6</td>
<td>EDBT</td>
<td>ICDCS</td>
<td>ECAI</td>
<td>SPAA</td>
</tr>
<tr>
<td>7</td>
<td>DASFAA</td>
<td>NETWORKING</td>
<td>RoboCup</td>
<td>PODC</td>
</tr>
<tr>
<td>8</td>
<td>PODS</td>
<td>MobiHoc</td>
<td>IAT</td>
<td>CRYPTO</td>
</tr>
<tr>
<td>9</td>
<td>SSDBM</td>
<td>ISCC</td>
<td>ICMAS</td>
<td>APPROX-RANDOM</td>
</tr>
<tr>
<td>10</td>
<td>SDM</td>
<td>SenSys</td>
<td>CP</td>
<td>EUROCRYPT</td>
</tr>
</tbody>
</table>

RankClus Result on DBLP
NetClus [KDD’09]: Ranking-based Clustering on Star Networks

NetClus Idea Illustration

A Net-Cluster of Database Area
Outline

• Why Heterogeneous Information Networks?

• Dissertation Overview
  • Ranking-Based Clustering
  • Meta-Path-Based Similarity Search and Mining
  • User-Guided Relation Strength-Aware Mining

• Conclusion and Future Plan
Similarity Search: Find Similar Objects in Networks
[VLDB’11]

- **DBLP**
  - Who are most similar to “Christos Faloutsos”?

- **IMDB**
  - Which movies are most similar to “Little Miss Sunshine”?

- **E-Commerce**
  - Which products are most similar to “Kindle”?

How to systematically answer these questions in heterogeneous information networks?
Network Schema and Meta-Path

Objects are connected together via different types of relationships!

“Jim-P1-Ann”
“Mike-P2-Ann”
“Mike-P3-Bob”

“Jim-P1-SIGMOD-P2-Ann”
“Mike-P3-SIGMOD-P2-Ann”
“Mike-P4-KDD-P5-Bob”

Author-Paper-Author

Author-Paper-Venue-Paper-Author

• Network schema
  • Meta-level description of a network

• Meta-Path
  • Meta-level description of a path between two objects
  • A path on network schema
  • Denote an existing or concatenated relation between two object types
Different Meta-Paths Tell Different Semantics

• Who are most similar to Christos Faloutsos?

Meta-Path: **Author-Paper-Author**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Author</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Christos Faloutsos</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Spiros Papadimitriou</td>
<td>0.127</td>
</tr>
<tr>
<td>3</td>
<td>Jimeng Sun</td>
<td>0.12</td>
</tr>
<tr>
<td>4</td>
<td>Jia-Yu Pan</td>
<td>0.114</td>
</tr>
<tr>
<td>5</td>
<td>Agma J. M. Traina</td>
<td>0.110</td>
</tr>
<tr>
<td>6</td>
<td>Jure Leskovec</td>
<td>0.096</td>
</tr>
<tr>
<td>7</td>
<td>Caetano Traina Jr.</td>
<td>0.096</td>
</tr>
<tr>
<td>8</td>
<td>Hanghang Tong</td>
<td>0.091</td>
</tr>
<tr>
<td>9</td>
<td>Deepayan Chakrabarti</td>
<td>0.083</td>
</tr>
<tr>
<td>10</td>
<td>Flip Korn</td>
<td>0.053</td>
</tr>
</tbody>
</table>

Christos’s students or close collaborators

Meta-Path: **Author-Paper-Venue-Paper-Author**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Author</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Christos Faloutsos</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Jiawei Han</td>
<td>0.842</td>
</tr>
<tr>
<td>3</td>
<td>Rakesh Agrawal</td>
<td>0.838</td>
</tr>
<tr>
<td>4</td>
<td>Jian Pei</td>
<td>0.8</td>
</tr>
<tr>
<td>5</td>
<td>Charu C. Aggarwal</td>
<td>0.739</td>
</tr>
<tr>
<td>6</td>
<td>H. V. Jagadish</td>
<td>0.705</td>
</tr>
<tr>
<td>7</td>
<td>Raghu Ramakrishnan</td>
<td>0.697</td>
</tr>
<tr>
<td>8</td>
<td>Nick Koudas</td>
<td>0.689</td>
</tr>
<tr>
<td>9</td>
<td>Surajit Chaudhuri</td>
<td>0.677</td>
</tr>
<tr>
<td>10</td>
<td>Divesh Srivastava</td>
<td>0.661</td>
</tr>
</tbody>
</table>

Work on similar topics and have similar reputation
Some Meta-Path Is “Better” Than Others

• Which pictures are most similar to the given image?

Evaluate the similarity between images according to their linked tags

Meta-Path: Image-Tag-Image

Evaluate the similarity between images according to tags and groups

Meta-Path: Image-Tag-Image-Group-Image-Tag-Image
PathSim: Similarity in Terms of “Peers”

• Why peers?
  • Strongly connected, while similar visibility

• In addition to meta-path
  • Need to consider similarity measures
Only PathSim Can Find Peers

**PathSim**

- Normalized path count between x and y following meta-path \( \mathcal{P} \)

\[
s(x, y) = \frac{2 \times |\{p_{x \rightarrow y} : p_{x \rightarrow y} \in \mathcal{P}\}|}{|\{p_{x \rightarrow x} : p_{x \rightarrow x} \in \mathcal{P}\}| + |\{p_{y \rightarrow y} : p_{y \rightarrow y} \in \mathcal{P}\}|}
\]

- Favor “peers”:
  - objects with strong connectivity and similar visibility under the given meta-path

- Calculation
  - For \( \mathcal{P} \): \( A_1 \rightarrow A_2 \rightarrow \cdots \rightarrow A_l \rightarrow A_{l-1} \rightarrow \cdots \rightarrow A_1 \)
    - \( M = W_{A_1 A_2} W_{A_2 A_3} \cdots W_{A_{l-1} A_l} W_{A_l A_{l-1}} \cdots W_{A_3 A_2} W_{A_2 A_1} \)
  - \( s(x, y) = \frac{2M_{xy}}{M_{xx} + M_{yy}} \)

- A co-clustering based pruning algorithm is provided
  - 18.23% - 68.04% efficiency improvement over the baseline
Find Academic Peers by PathSim

- **Anhai Doan**
  - CS, Wisconsin
  - Database area
  - PhD: 2002

- **Jignesh Patel**
  - CS, Wisconsin
  - Database area
  - PhD: 1998

- **Amol Deshpande**
  - CS, Maryland
  - Database area
  - PhD: 2004

- **Jun Yang**
  - CS, Duke
  - Database area
  - PhD: 2001

Meta-Path: Author-Paper-Venue-Paper-Author

<table>
<thead>
<tr>
<th>Rank</th>
<th>P-PageRank</th>
<th>SimRank</th>
<th>PathSim</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AnHai Doan</td>
<td>AnHai Doan</td>
<td>AnHai Doan</td>
</tr>
<tr>
<td>2</td>
<td>Philip S. Yu</td>
<td>Douglas W. Cornell</td>
<td>Jignesh M. Patel</td>
</tr>
<tr>
<td>3</td>
<td>Jiawei Han</td>
<td>Adam Silberstein</td>
<td>Amol Deshpande</td>
</tr>
<tr>
<td>4</td>
<td>Hector Garcia-Molina</td>
<td>Samuel DeFazio</td>
<td>Jun Yang</td>
</tr>
<tr>
<td>5</td>
<td>Gerhard Weikum</td>
<td>Curt Ellmann</td>
<td>Renée J. Miller</td>
</tr>
</tbody>
</table>
PathPredict: Meta-Path-Based Co-authorship Prediction in DBLP [ASONAM’11]

- **Co-authorship prediction problem**
  - Whether two authors are going to collaborate for the first time

- **Co-authorship encoded in meta-path**
  - Author-Paper-Author

- **Topological features encoded in meta-paths**

<table>
<thead>
<tr>
<th>Meta-Path</th>
<th>Semantic Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A - P \to P - A$</td>
<td>$a_i$ cites $a_j$</td>
</tr>
<tr>
<td>$A - P \leftrightarrow P - A$</td>
<td>$a_i$ is cited by $a_j$</td>
</tr>
<tr>
<td>$A - P - V - P - A$</td>
<td>$a_i$ and $a_j$ publish in the same venues</td>
</tr>
<tr>
<td>$A - P - A - P - A$</td>
<td>$a_i$ and $a_j$ are co-authors of the same authors</td>
</tr>
<tr>
<td>$A - P - T - P - A$</td>
<td>$a_i$ and $a_j$ write the same topics</td>
</tr>
<tr>
<td>$A - P \to P \to P - A$</td>
<td>$a_i$ cites papers that cite $a_j$</td>
</tr>
<tr>
<td>$A - P \leftrightarrow P \leftrightarrow P - A$</td>
<td>$a_i$ is cited by papers that are cited by $a_j$</td>
</tr>
<tr>
<td>$A - P \rightarrow P \leftrightarrow P - A$</td>
<td>$a_i$ and $a_j$ cite the same papers</td>
</tr>
<tr>
<td>$A - P \leftarrow P \rightarrow P - A$</td>
<td>$a_i$ and $a_j$ are cited by the same papers</td>
</tr>
</tbody>
</table>

Meta-paths between authors under length 4
The Power of PathPredict

- Explain the prediction power of each meta-path
  - Wald Test for logistic regression
- Higher prediction accuracy than using projected homogeneous network
  - 11% higher in prediction accuracy

<table>
<thead>
<tr>
<th>Meta Path</th>
<th>p-value</th>
<th>significance level</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A \rightarrow P \rightarrow P \rightarrow A$</td>
<td>0.0378</td>
<td>**</td>
</tr>
<tr>
<td>$A \rightarrow P \leftarrow P \rightarrow A$</td>
<td>0.0077</td>
<td>***</td>
</tr>
<tr>
<td>$A \rightarrow P \rightarrow V \rightarrow P \rightarrow A$</td>
<td>1.2974e-174</td>
<td>****</td>
</tr>
<tr>
<td>$A \rightarrow P \rightarrow A \rightarrow P \rightarrow A$</td>
<td>1.1484e-126</td>
<td>****</td>
</tr>
<tr>
<td>$A \rightarrow P \rightarrow T \rightarrow P \rightarrow A$</td>
<td>3.4867e-51</td>
<td>****</td>
</tr>
<tr>
<td>$A \rightarrow P \rightarrow P \rightarrow P \rightarrow A$</td>
<td>0.7459</td>
<td></td>
</tr>
<tr>
<td>$A \rightarrow P \leftarrow P \leftarrow P \rightarrow A$</td>
<td>0.0647</td>
<td>*</td>
</tr>
<tr>
<td>$A \rightarrow P \rightarrow P \leftarrow P \rightarrow A$</td>
<td>9.7641e-11</td>
<td>****</td>
</tr>
<tr>
<td>$A \rightarrow P \leftarrow P \rightarrow P \rightarrow A$</td>
<td>0.0966</td>
<td>*</td>
</tr>
</tbody>
</table>

$^1$*: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$, ****: $p < 0.001$

Co-author prediction for Jian Pei: Only 42 among 4809 candidates are true first-time co-authors!
(Feature collected in [1996, 2002]; Test period in [2003, 2009])
When Will It Happen? [WSDM’12]

• From “whether” to “when”
  • “Whether”: Will Jim rent the movie “Avatar” in Netflix?
    - Output: P(X=1)=?
  • “When”: When will Jim rent the movie “Avatar”?

  - What is the probability Jim will rent “Avatar” within 2 months?
    - \( P(Y \leq 2) \)
  - By when Jim will rent “Avatar” with 90% probability?
    - \( t: P(Y \leq t) = 0.9 \)
  - What is the expected time it will take for Jim to rent “Avatar”?
    - \( E(Y) \)

Output: A distribution of time!

May provide useful information to supply chain management
The Relationship Building Time Prediction Model

• **Solution**
  - Directly **model relationship building time**: $P(Y=t)$
    - Geometric distribution, Exponential distribution, Weibull distribution
  - Use **generalized linear model**
    - Deal with censoring (relationship builds beyond the observed time interval)

$T$: Right Censoring

**Generalized Linear Model under Weibull Distribution Assumption**

$$LLW(\beta, \lambda) = \sum_{i=1}^{n} I_{\{y_i < T\}} \log \frac{\lambda y_i^{\lambda - 1}}{e^{-\lambda x_i \beta}} - \sum_{i=1}^{n} \left( \frac{y_i}{e^{-x_i \beta}} \right)^\lambda$$

$$\log L = \sum_{i=1}^{n} \left( f_Y(y_i | \theta_i, \lambda) I_{\{y_i < T\}} + P(y_i \geq T | \theta_i, \lambda) I_{\{y_i \geq T\}} \right)$$
Author Citation Time Prediction in DBLP

• Top-4 meta-paths for author citation time prediction

- \[ A - P - T - P - A \]
- \[ A - P \leftrightarrow P \rightarrow P - A \]
- \[ A - P - A - P \rightarrow P - A \]
- \[ A - P - T - P - A - P \rightarrow P - A \]

Social relations are less important in author citation prediction than in co-author prediction.

• Predict when Philip S. Yu will cite a new author

<table>
<thead>
<tr>
<th>( \alpha_i )</th>
<th>( \alpha_j )</th>
<th>Ground Truth</th>
<th>Median</th>
<th>Mean</th>
<th>25% quantile</th>
<th>75% quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philip S. Yu</td>
<td>Ling Liu</td>
<td>1</td>
<td>2.2386</td>
<td>3.4511</td>
<td>0.8549</td>
<td>4.7370</td>
</tr>
<tr>
<td>Philip S. Yu</td>
<td>Christian S. Jensen</td>
<td>3</td>
<td>2.7840</td>
<td>4.2919</td>
<td>1.0757</td>
<td>5.8911</td>
</tr>
<tr>
<td>Philip S. Yu</td>
<td>C. Lee Giles</td>
<td>0</td>
<td>8.3985</td>
<td>12.9474</td>
<td>3.2450</td>
<td>17.7717</td>
</tr>
<tr>
<td>Philip S. Yu</td>
<td>Stefano Ceri</td>
<td>0</td>
<td>0.5729</td>
<td>0.8833</td>
<td>0.2214</td>
<td>1.2124</td>
</tr>
<tr>
<td>Philip S. Yu</td>
<td>David Maier</td>
<td>9+</td>
<td>2.5675</td>
<td>3.9581</td>
<td>0.9920</td>
<td>5.4329</td>
</tr>
<tr>
<td>Philip S. Yu</td>
<td>Tong Zhang</td>
<td>9+</td>
<td>9.5371</td>
<td>14.7028</td>
<td>3.6849</td>
<td>20.1811</td>
</tr>
<tr>
<td>Philip S. Yu</td>
<td>Rudi Studer</td>
<td>9+</td>
<td>9.7752</td>
<td>15.0698</td>
<td>3.7769</td>
<td>20.6849</td>
</tr>
</tbody>
</table>

Under Weibull distribution assumption
Outline

• Why Heterogeneous Information Networks?

• Dissertation Overview
  • Ranking-Based Clustering
  • Meta-Path-Based Similarity Search and Mining
  • User-Guided Relation Strength-Aware Mining

• Conclusion and Future Plan
Content-Rich Heterogeneous information networks become increasingly popular

- Heterogeneous links + (incomplete) attributes
- Examples
  - Social media
  - E-Commerce
  - Cyber-physical system

Soft clustering objects using both link information and attribute information

- E-Commerce: customers, products, comments, ...
- Social websites: people, groups, books, posts, ...

Understanding the strengths for different relations in determining object’s cluster
### Incomplete Attributes

<table>
<thead>
<tr>
<th>Age</th>
<th>Salary</th>
<th>Interests</th>
<th>Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>10K</td>
<td>Sports, Music</td>
<td>Champaign, Boston</td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>50</td>
<td>N/A</td>
<td>Shopping, Books</td>
<td>N/A</td>
</tr>
<tr>
<td>52</td>
<td>120K</td>
<td>N/A</td>
<td>Boston</td>
</tr>
<tr>
<td>N/A</td>
<td>100K</td>
<td>Cooking, Books</td>
<td>Chicago, Seattle</td>
</tr>
</tbody>
</table>

#### Customer Segmentation According to Customer Profiles

<table>
<thead>
<tr>
<th>Temperature (F)</th>
<th>Precipitation (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>5</td>
</tr>
<tr>
<td>N/A</td>
<td>15</td>
</tr>
<tr>
<td>N/A</td>
<td>20</td>
</tr>
<tr>
<td>80</td>
<td>N/A</td>
</tr>
<tr>
<td>85</td>
<td>N/A</td>
</tr>
</tbody>
</table>

#### Weather Pattern Clustering According to Weather Sensor Records

- **Precip. Sensor Type**
  - N/A
  - 5
  - 15
  - 20
  - N/A

- **Temp. Sensor Type**
  - N/A
  - 80
  - 85
  - N/A

Object level: Missing data obs.

Schema level: Some type of objects only contains partial attribute types
### The Links Help!

#### Customer Segmentation According to Customer Profiles

<table>
<thead>
<tr>
<th>Age</th>
<th>Salary</th>
<th>Interests</th>
<th>Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>10K</td>
<td>Sports, Music</td>
<td>Champaign, Boston</td>
</tr>
<tr>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>50</td>
<td>N/A</td>
<td>Shopping, Books</td>
<td>N/A</td>
</tr>
<tr>
<td>52</td>
<td>120K</td>
<td>N/A</td>
<td>Boston</td>
</tr>
<tr>
<td>N/A</td>
<td>100K</td>
<td>Cooking, Books</td>
<td>Chicago, Seattle</td>
</tr>
</tbody>
</table>

#### Weather Pattern Clustering According to Weather Sensor Records

<table>
<thead>
<tr>
<th>Temperature (F)</th>
<th>Precipitation (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>5</td>
</tr>
<tr>
<td>N/A</td>
<td>15</td>
</tr>
<tr>
<td>N/A</td>
<td>20</td>
</tr>
<tr>
<td>80</td>
<td>N/A</td>
</tr>
<tr>
<td>85</td>
<td>N/A</td>
</tr>
</tbody>
</table>
The Basic Assumption of Linked Objects

- **Linked objects tend to be in the same cluster**
  - Hard clustering: share the same cluster label
  - Soft clustering: have similar cluster membership vectors ($\theta$)

$$\theta_o = f(\theta_a, \theta_b, \theta_c, \theta_d)$$

- An object’s cluster membership is also dependent on its neighbors
Different Relations Carry Different Strengths

• An object is linking to other objects via different types of relationships

  (o, a): Family relationship
  (o, b): Friendship
  (o, c): Schoolmate
  (o, d): Colleague relationship

• For a certain clustering task, different relations carry different strengths

Voter segmentation in political campaign?

A set of attributes given by users serve as guidance to learn the strength of each relation
The Relation Strength-Aware Clustering Problem

• **Input:**
  - A heterogeneous information network, $G$
  - A subset of attributes associated with $G$
  - Number of clusters, $K$

• **Output:**
  - Soft clustering membership vector $\theta_i$ for each object $o_i$
  - Relation strength $\gamma(r)$ for each relation $r$
Case Studies of Relation Strengths

A paper's research area is more determined by its authors than its venue (13.30 vs. 3.13)
Integrating Meta-Path Selection with User-Guided Object Clustering [KDD’12]

- Goal: Clustering authors based on their connection in the network

Which meta-path to choose?
The Role of User Guidance

- It is users’ responsibility to specify their clustering purpose
  - Say, by giving seeds in each cluster

Seeds | Meta-path(s) | Clustering Result
--- | --- | ---
\{1\}, \{5\} | (a) AOA | \{1,2,3,4\}, \{5,6,7,8\}
\{1\}, \{2\}, \{5\}, \{6\} | (c) AOA + AVA | \{1,3\}, \{2,4\}, \{5,7\}, \{6,8\}
The Problem of User-Guided Clustering with Meta-Path Selection [KDD’12]

**Input:**
- The target type for clustering: \( T \)
- Number of clusters: \( K \)
  - Seeds in *some* of the clusters: \( L_1, L_2, ..., L_K \)
- Candidate \( M \) meta-paths starting from \( T \): \( \mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_M \)

**Output:**
- The quality weight for each candidate meta-path in the clustering process
  - \( \alpha_m \)
- The clustering results that are consistent with the user guidance
  - \( \theta_i \)
DBLP-Clustering Venues According to Research Areas

• **Task:**
  - Target objects: venues
  - Number of clusters: 4;
  - Candidate meta-paths: \( V-P-A-P-V, V-P-T-P-V \)

• **Output:**
  - **Weights:**
    - \( V-P-A-P-V: 1576 \) (0.0017 per relationship)
    - \( V-P-T-P-V: 17001 \) (0.0003 per relationship)

• **Clustering results:**

<table>
<thead>
<tr>
<th>#S</th>
<th>Measure</th>
<th>PathSelClus</th>
<th>LP</th>
<th>ITC</th>
<th>LP_voting</th>
<th>LP_soft</th>
<th>ITC_voting</th>
<th>ITC_soft</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Accuracy</td>
<td>0.9950</td>
<td>0.6500</td>
<td>0.6900</td>
<td>0.6500</td>
<td>0.6650</td>
<td>0.6450</td>
<td>0.5100</td>
</tr>
<tr>
<td></td>
<td>NMI</td>
<td>0.9906</td>
<td>0.6181</td>
<td>0.6986</td>
<td>0.6181</td>
<td>0.5801</td>
<td>0.5903</td>
<td>0.5316</td>
</tr>
<tr>
<td>2</td>
<td>Accuracy</td>
<td>1</td>
<td>0.7500</td>
<td>0.8450</td>
<td>0.7500</td>
<td>0.8200</td>
<td>0.8950</td>
<td>0.8700</td>
</tr>
<tr>
<td></td>
<td>NMI</td>
<td>1</td>
<td>0.6734</td>
<td>0.7752</td>
<td>0.6734</td>
<td>0.7492</td>
<td>0.8321</td>
<td>0.7942</td>
</tr>
</tbody>
</table>
Yelp-Clustering Yelp Restaurants into Categories

- **Task:**
  - Target objects: restaurants
  - Number of clusters: 6
  - Candidate meta-paths: \textit{B-R-U-R-B, B-R-T-R-B}.

- **Output:**
  - Weights:
    - \textit{B-R-U-R-B} : 6000 (0.1716 per relationship, compared with 0.5864 for clustering shopping categories)
    - \textit{B-R-T-R-B}: $2.9522 \times 10^7$ (0.0138 per relationship)

<table>
<thead>
<tr>
<th>%S</th>
<th>Measure</th>
<th>PathSelClus</th>
<th>LP</th>
<th>ITC</th>
<th>LP_voting</th>
<th>LP_soft</th>
<th>ITC_voting</th>
<th>ITC_soft</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>Accuracy</td>
<td>0.7435</td>
<td>0.1137</td>
<td>0.1758</td>
<td>0.2112</td>
<td>0.2112</td>
<td>0.2430</td>
<td>0.2022</td>
</tr>
<tr>
<td></td>
<td>NMI</td>
<td>0.6517</td>
<td>0.0323</td>
<td>0.0178</td>
<td>0.0578</td>
<td>0.0578</td>
<td>0.2308</td>
<td>0.2490</td>
</tr>
<tr>
<td>2%</td>
<td>Accuracy</td>
<td>0.8004</td>
<td>0.1264</td>
<td>0.1910</td>
<td>0.2202</td>
<td>0.2202</td>
<td>0.2762</td>
<td>0.2792</td>
</tr>
<tr>
<td></td>
<td>NMI</td>
<td>0.6803</td>
<td>0.0487</td>
<td>0.0150</td>
<td>0.0801</td>
<td>0.0801</td>
<td>0.2099</td>
<td>0.2907</td>
</tr>
<tr>
<td>5%</td>
<td>Accuracy</td>
<td>0.8125</td>
<td>0.2653</td>
<td>0.2200</td>
<td>0.2437</td>
<td>0.2437</td>
<td>0.3049</td>
<td>0.3240</td>
</tr>
<tr>
<td></td>
<td>NMI</td>
<td>0.6894</td>
<td>0.1111</td>
<td>0.0220</td>
<td>0.1212</td>
<td>0.1212</td>
<td>0.2252</td>
<td>0.2692</td>
</tr>
</tbody>
</table>
Funding Sources

- The Multimodal Information Access & Synthesis (MIAS) Center funded by a grant from the Department of Homeland Security (DHS)

- NSF (IIS-09-05215 “On-Line Analytical Mining of Heterogeneous Information Networks”)
  - 2009-2012

- The Information Network Academic Research Center (INARC) as part of Network Science-Collaborative Technology Alliance (NS-CTA) program, U.S. Army Research Laboratory (ARL)
  - $16.75 million, 2009-2014
My Other Work

- **Quality of Information Analysis on Information Networks**
  - Trustworthiness Analysis [WWW’11][IPSN’11]
  - Outlier Detection [KDD’10, KDD’12, ECMLPKDD’12]

- **Knowledge Ensemble of Heterogeneous Source Information**
  - [KDD’09], [NIPS’09], [TKDE’12]

- **Spatio-Temporal Mining on Cyber-Physical Data**
  - Atypical Event OLAP [ICDE’12]

- **Business Intelligence**
  - Online Promotion Analysis [EDBT’10]

- **Text Mining**
  - Query Log Analysis [WWW’07]
Outline

• Why Heterogeneous Information Networks?

• Dissertation Overview
  • Ranking-Based Clustering
  • Meta-Path-Based Similarity Search and Mining
  • User-Guided Relation Strength-Aware Mining

• Conclusion and Future Plan
Conclusion

- Mining heterogeneous information networks is a new game

- Three principles
  1. Use Holistic Network Information
  2. Explore Network Meta Structure
  3. User-Guided Exploration
## Visions and Long Term Goals

<table>
<thead>
<tr>
<th>Application Level</th>
<th>Mining Algorithm Level</th>
<th>System Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Media</td>
<td>E-Commerce</td>
<td>Healthcare</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
It is my great honor to receive this award. I’d like to thank the KDD dissertation award committee, my advisor Prof. Jiawei Han, my long list of collaborators, such as Philip S. Yu, Charu C. Aggarwal, Xifeng Yan, Nitesh Chawla, Jie Tang, and all the UIUC Data Mining Group members!

THANK YOU!