Mining Trusted Information in Medical Science: An Information Network Approach

Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign

Collaborated with many, especially Yizhou Sun, Ming Ji, Chi Wang, Tim Weninger, Xiaoxin Yin, Bo Zhao

Acknowledgements: NSF, ARL, NASA, AFOSR (MURI), Microsoft, IBM, Yahoo!, Google, HP Lab & Boeing

November 28, 2012
Outline

- Why Information Network Approach for Medical and Health Informatics?
- Exploring Rich Semantics of Structured Heterogeneous Networks
  - From RankClus to RankClass
  - A PubMed Exploration
- Information Trust Analysis: An Info. Network Approach
  - From Truth Finder to Latent Truth Model
- Conclusions
The Real World: Heterogeneous Networks

- Multiple object types and/or multiple link types

Homogeneous networks are **information loss** projection of heterogeneous networks!

Directly mining information richer heterogeneous networks
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>
</tr>
</thead>
<tbody>
<tr>
<td>How are CS research areas <strong>structured</strong>?</td>
<td>Clustering</td>
</tr>
<tr>
<td>Who are the <strong>leading</strong> researchers on Web search?</td>
<td>Ranking</td>
</tr>
<tr>
<td>What are the most essential <strong>terms, venues, authors in AI</strong>?</td>
<td>Classification + Ranking</td>
</tr>
<tr>
<td>Who are the <strong>peer</strong> researchers of Jure Leskovec?</td>
<td>Similarity Search</td>
</tr>
<tr>
<td>Whom <strong>will</strong> Christos Faloutsos <strong>collaborate with</strong>?</td>
<td>Relationship Prediction</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>
</tr>
<tr>
<td>How was the field of Data Mining <strong>emerged</strong> or <strong>evolving</strong>?</td>
<td>Network Evolution</td>
</tr>
<tr>
<td>Which authors are <strong>rather different</strong> from his/her peers in IR?</td>
<td>Outlier/anomaly detection</td>
</tr>
</tbody>
</table>
Outline

- Why Information Network Approach for Medical and Health Informatics?
- Exploring Rich Semantics of Structured Heterogeneous Networks
  - From RankClus to RankClass
  - A PubMed Exploration
- Information Trust Analysis: An Info. Network Approach
  - From Truth Finder to Latent Truth Model
- Conclusions
**RankClus: Algorithm Framework**

- **Initialization**
  - Randomly partition

- **Repeat**
  - **Ranking**
    - Ranking objects in each sub-network induced from each cluster
    - Generating new measure space
      - Estimate *mixture model coefficients* for each target object
    - Adjusting cluster

- Until stable
NetClus on DBLP: Database System Cluster

database 0.0995511  
  system 0.0678563  
  data 0.0214893  
  query 0.0133316  
  management 0.00850744  
  object 0.00837766  
  relational 0.0081175

VLDB 0.318495  
SIGMOD Conf. 0.313903  
ICDE 0.188746  
PODS 0.107943  
EDBT 0.0436849

Surajit Chaudhuri 0.00678065  
Michael Stonebraker 0.00616469  
Michael J. Carey 0.00545769  
C. Mohan 0.00528346  
David J. DeWitt 0.00491615  
Hector Garcia-Molina 0.00453497  
H. V. Jagadish 0.00434289  
David B. Lomet 0.00397865

Rank-Based Clustering of Multimedia Data

RankCompete: Organize your photo album automatically!
M. Ji, M. Danilevski, et al., “Graph Regularized Transductive Classification on Heterogeneous Information Networks”, ECMLPKDD'10
Experiments with Very Small Training Set

- **DBLP**: 4-fields data set (DB, DM, AI, IR) forming a heterog. info. network
- Rank objects within each class (with extremely limited label information)
- Obtain High classification accuracy and excellent rankings within each class

<table>
<thead>
<tr>
<th>Top-5 ranked conferences</th>
<th>Database</th>
<th>Data Mining</th>
<th>AI</th>
<th>IR</th>
</tr>
</thead>
<tbody>
<tr>
<td>VLDB</td>
<td>KDD</td>
<td>IJCAI</td>
<td></td>
<td>SIGIR</td>
</tr>
<tr>
<td>SIGMOD</td>
<td>SDM</td>
<td>AAAI</td>
<td></td>
<td>ECIR</td>
</tr>
<tr>
<td>ICDE</td>
<td>ICDM</td>
<td>ICML</td>
<td></td>
<td>CIKM</td>
</tr>
<tr>
<td>PODS</td>
<td>PKDD</td>
<td>CVPR</td>
<td></td>
<td>WWW</td>
</tr>
<tr>
<td>EDBT</td>
<td>PAKDD</td>
<td>ECML</td>
<td></td>
<td>WSDM</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Top-5 ranked terms</th>
<th>data</th>
<th>mining</th>
<th>learning</th>
<th>retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>database</td>
<td>data</td>
<td>knowledge</td>
<td>information</td>
<td></td>
</tr>
<tr>
<td>query</td>
<td>clustering</td>
<td>reasoning</td>
<td>web</td>
<td></td>
</tr>
<tr>
<td>system</td>
<td>classification</td>
<td>logic</td>
<td>search</td>
<td></td>
</tr>
<tr>
<td>xml</td>
<td>frequent</td>
<td>cognition</td>
<td>text</td>
<td></td>
</tr>
</tbody>
</table>
MedRank: Discovering Influential Medical Treatments from Literature

- Heuristics: A good treatment is likely to be found in good medical articles published in good journals and written by good authors and successful in clinical trials
- Data (PubMed) and Ontology
  - 20M articles, forming a gigantic heterogeneous infonet
  - Use only those treatments that passed Clinical Trial Phase III
- MeSH: Medical ontology used
- Exploring rich semantics of structured heterogeneous networks
  - Star schema
  - MedRank (extension to NetClus)
  - Ranked treatments on popular and non-popular diseases

Star Schema for PubMed InfoNet
Experiments: Ranking Medical Treatments

**Table 1: Size of sub-networks in categories**

<table>
<thead>
<tr>
<th>TypeNs</th>
<th>ALS</th>
<th>HB</th>
<th>AIDS</th>
<th>D2</th>
<th>RA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article</td>
<td>7975</td>
<td>33679</td>
<td>48962</td>
<td>50732</td>
<td>70736</td>
</tr>
<tr>
<td>Author</td>
<td>16637</td>
<td>67320</td>
<td>86481</td>
<td>99060</td>
<td>108234</td>
</tr>
<tr>
<td>Journal</td>
<td>1256</td>
<td>2936</td>
<td>4272</td>
<td>3308</td>
<td>3963</td>
</tr>
<tr>
<td>Treatment</td>
<td>383</td>
<td>669</td>
<td>937</td>
<td>1121</td>
<td>1401</td>
</tr>
<tr>
<td>Clinical Trial</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>25256</td>
<td>104609</td>
<td>140657</td>
<td>154316</td>
<td>184339</td>
</tr>
</tbody>
</table>

**Ranking influential treatments for diseases from MEDLINE data**

1. Zidovudine/therapeutic use (0.1679)
2. Anti-HIV Agents/therapeutic use (0.1340)
3. Antiretroviral Therapy, Highly Active (0.0977)
4. Antiviral Agents/therapeutic use (0.0718)
5. Anti-Retroviral Agents/therapeutic use (0.0236)
6. Interferon Type I/therapeutic use (0.0147)
7. Didanosine/therapeutic use (0.0132)
8. Ganciclovir/therapeutic use (0.0114)
9. HIV Protease Inhibitors/therapeutic use (0.0105)
10. Antineoplastic Combined Chemotherapy (0.0103)

**Treatments of 5 diseases**

- ALS: Amyotrophic Lateral Sclerosis
- HB: Hepatitis B
- AIDS:
- D2: Diabetes Mellitus Type II
- RA: Rheumatoid Arthritis

**MedRank vs. baselines using AO (average over sum of weighted overlaps of 1st d els)**
Guidance: Meta Path in Bibliographic Network

- Relationship prediction: meta path-guided prediction
- Meta path relationships among similar typed links share similar semantics and are comparable and inferable

- Co-author prediction (A—P—A) using topological features also encoded by meta paths, e.g., citation relations between authors (A—P→P—A)
Meta-Path Based Co-authorship Prediction in DBLP

- 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

### Meta-Path

<table>
<thead>
<tr>
<th>Meta-Path</th>
<th>Semantic Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A \rightarrow P \rightarrow P - A$</td>
<td>$a_i$ cites $a_j$</td>
</tr>
<tr>
<td>$A \leftarrow P \leftarrow P - A$</td>
<td>$a_i$ is cited by $a_j$</td>
</tr>
<tr>
<td>$A \rightarrow P - V - P - A$</td>
<td>$a_i$ and $a_j$ publish in the same venues</td>
</tr>
<tr>
<td>$A \rightarrow P - A - P - A$</td>
<td>$a_i$ and $a_j$ are co-authors of the same authors</td>
</tr>
<tr>
<td>$A \rightarrow P - T - P - A$</td>
<td>$a_i$ and $a_j$ write the same topics</td>
</tr>
<tr>
<td>$A \rightarrow P \rightarrow P \rightarrow P - A$</td>
<td>$a_i$ cites papers that cite $a_j$</td>
</tr>
<tr>
<td>$A \leftarrow P \leftarrow P \leftarrow P - A$</td>
<td>$a_i$ is cited by papers that are cited by $a_j$</td>
</tr>
<tr>
<td>$A \rightarrow P \leftarrow P \leftarrow P - A$</td>
<td>$a_i$ and $a_j$ cite the same papers</td>
</tr>
<tr>
<td>$A \leftarrow P \leftarrow P \rightarrow P - A$</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>$A \leftarrow P \leftarrow P \leftarrow A$</td>
<td>0.0077</td>
<td>***</td>
</tr>
<tr>
<td>$A \rightarrow P \rightarrow V \rightarrow P \leftarrow A$</td>
<td>1.2974e-14</td>
<td>****</td>
</tr>
<tr>
<td>$A \rightarrow P \rightarrow A \rightarrow P \leftarrow A$</td>
<td>1.1484e-12</td>
<td>****</td>
</tr>
<tr>
<td>$A \rightarrow P \rightarrow T \rightarrow P \leftarrow A$</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>$A \rightarrow P \leftarrow P \leftarrow P \leftarrow A$</td>
<td>0.0647</td>
<td>*</td>
</tr>
<tr>
<td>$A \rightarrow P \rightarrow P \leftarrow P \leftarrow A$</td>
<td>9.7641e-11</td>
<td>****</td>
</tr>
<tr>
<td>$A \leftarrow P \leftarrow P \rightarrow P \leftarrow A$</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])

<table>
<thead>
<tr>
<th>Rank</th>
<th>Hybrid heterogeneous features</th>
<th># Shared authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Philip S. Yu</td>
<td>Philip S. Yu</td>
</tr>
<tr>
<td>2</td>
<td>Raymond T. Ng</td>
<td>Ming-Syan Chen</td>
</tr>
<tr>
<td>3</td>
<td>Osmar R. Zafi^\text{e}</td>
<td>Divesh Srivastava</td>
</tr>
<tr>
<td>4</td>
<td>Ling Feng</td>
<td>Kotagiri Ramamohanarao</td>
</tr>
<tr>
<td>5</td>
<td>David Wai-Lok Cheung</td>
<td>Jeffrey Xu Yu</td>
</tr>
</tbody>
</table>
Outline

- Why Information Network Approach for Medical and Health Informatics?
- Exploring Rich Semantics of Structured Heterogeneous Networks
  - From RankClus to RankClass
  - A PubMed Exploration
- Information Trust Analysis: An Info. Network Approach
  - From Truth Finder to Latent Truth Model
- Conclusions
Enhancing the Quality of Heterogeneous Info. Networks

- Info. networks could be untrustworthy, error-prone, missing, ...
- TruthFinder [KDD’07]: Inference on trustworthiness by mutual enhancement of info provider and statement trustworthiness
- Latent Truth Model (LTM) [VLDB12]: Modeling two-sided quality to support multiple true values per entity for truth-finding

**Generating Implicit Negative Claims:**

- High Precision, High Recall
- Low Precision, Low Recall

Web sites

- $w_1$
- $w_2$
- $w_3$
- $w_4$

Facts

- $f_1$
- $f_2$
- $f_3$
- $f_4$

Objects

- $o_1$
- $o_2$

Harry Potter
Truth Discovery:
Effectiveness of Latent Truth Model

Experimental datasets: Large and real

- **Book Authors from abebooks.com** (1263 books, 879 sources, 48153 claims, 2420 book-author, 100 labeled)
- **Movie Directors from Bing** (15073 movies, 12 sources, 108873 claims, 33526 movie-director, 100 labeled)

Effectiveness of Latent Truth Model:

<table>
<thead>
<tr>
<th>Results on book data</th>
<th>Results on movie data</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-sided error</td>
<td>Two-sided error</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td><strong>Recall</strong></td>
</tr>
<tr>
<td>LTMinc</td>
<td>1.000</td>
</tr>
<tr>
<td>LTM</td>
<td>1.000</td>
</tr>
<tr>
<td>3-Estimates</td>
<td>1.000</td>
</tr>
<tr>
<td>Voting</td>
<td>1.000</td>
</tr>
<tr>
<td>TruthFinder</td>
<td>0.880</td>
</tr>
<tr>
<td>Investment</td>
<td>0.880</td>
</tr>
<tr>
<td>HubAuthority</td>
<td>1.000</td>
</tr>
<tr>
<td>AvgLog</td>
<td>1.000</td>
</tr>
<tr>
<td>LTMpos</td>
<td>0.880</td>
</tr>
<tr>
<td>PooledInvestment</td>
<td>1.000</td>
</tr>
</tbody>
</table>

- Model source quality in other data integration tasks, e.g. entity resolution.
- Trustworthiness in multi-genre networks (text-rich networks, social networks, etc.)
Outline

- Why Information Network Approach for Medical and Health Informatics?
- Exploring Rich Semantics of Structured Heterogeneous Networks
  - From RankClus to RankClass
  - A PubMed Exploration
- Information Trust Analysis: An Info. Network Approach
  - From Truth Finder to Latent Truth Model
- Conclusions
Conclusions

- Heterogeneous information networks are ubiquitous
  - Most datasets can be “organized” or “transformed” into “structured” multi-typed heterogeneous info. networks
    - Examples: DBLP, IMDB, Flickr, Google News, Wikipedia, ...
  - Surprisingly rich knowledge can be mined from such structured heterogeneous info. networks
    - Clustering, ranking, classification, data cleaning, trust analysis, role discovery, similarity search, relationship prediction, ......
    - Meta path holds a key to effective mining and exploration!
  - Knowledge is power, but knowledge is hidden in massive, but “relatively structured” nodes and links!
  - Much more to be explored in information network mining!
From Data Mining to Mining Info. Networks

Han, Kamber and Pei, Data Mining, 3rd ed. 2011
Yu, Han and Faloutsos (eds.), Link Mining, 2010
Sun and Han, Mining Heterogeneous Information Networks, 2012