APP I3.1: QoI Mining of Noisy, Volatile, Uncertain, and Incomplete Heterogeneous Information Networks

Key Objectives:
- Develop robust and quality mining methods for noisy and inaccurate heterogeneous information networks
- Design substantially enhanced data mining methods to uncover hidden patterns and knowledge in two complementary directions

Deliverables:
- Q1: Methodology design for (i) two-stage mining and (ii) noise-aware mining, in heterogeneous information networks.
- Q2: Algorithm development for the two approaches
- Q3: Algorithm test and refinement for the two approaches
- Q4: System prototype demo of the two approaches

Impact:
Will enable tools to uncover hidden patterns and knowledge from info. networks despite of noise and uncertainty in the networks

Key Technical Innovations
- Two-stage mining framework by cleaning the data before mining the cleansed data
- Noise-aware mining model by directly mining the networked data with the consideration of certain portion of data may not be clean, complete, or reliable
- Novel approximate graph pattern mining methods by leveraging the concept of proximity pattern
- Algorithm of network clustering/classification by identifying and marking low credibility node/links
- Quality analysis using clustering and classification methods developed for heterogeneous info. networks

<table>
<thead>
<tr>
<th>Role</th>
<th>Researchers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lead</td>
<td>J. Han, UIUC, INARC</td>
</tr>
<tr>
<td>Primary</td>
<td>X. Yan, UCSB, INARC</td>
</tr>
<tr>
<td>Primary</td>
<td>C. Faloutsos, CMU, INARC</td>
</tr>
<tr>
<td>Collab</td>
<td>H. Tong, IBM, SCNARC</td>
</tr>
<tr>
<td>Collab</td>
<td>C.-Y. Lin, IBM, CNARC</td>
</tr>
<tr>
<td>Collab</td>
<td>M. Magdon-Ismail, RPI, SCNARC</td>
</tr>
</tbody>
</table>
APP I3.2: Modeling and Mining of Text-Rich Information Networks

Key Objectives:
- Structurally model a text-rich info. network and investigate methods for mining knowledge from such networks
- Enhance keyword search and knowledge discovery capability by the text-rich info. network model

Deliverables:
Q1: Methodologies for modeling and construction of multi-dimensional, relatively structured information networks by progressive information network analysis
Q2: Models for enhanced text data analysis using relatively structured, heterogeneous information networks
Q3: Methods for multi-facet search in text-rich info. networks
Q4: System prototype demo of the approaches

Impact:
- The modeling, principles, and methodologies developed for text-rich information network will be applicable to other genres of networks

Key Technical Innovations
Efficient algorithms to enrich text mining techniques with the information network topology
Structurally modeling text-rich information networks by progressive network analysis on text and interconnected data
Enhanced text data analysis using relatively structured, heterogeneous information networks
Multi-facet search and mining in text-rich information networks

<table>
<thead>
<tr>
<th>Role</th>
<th>Researchers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lead</td>
<td>D. Roth, UIUC, INARC</td>
</tr>
<tr>
<td></td>
<td>J. Han, UIUC, INARC</td>
</tr>
<tr>
<td>Primary</td>
<td>H. Ji, UCSB, INARC</td>
</tr>
<tr>
<td></td>
<td>X. Yan, UCSB, INARC</td>
</tr>
<tr>
<td>Collab</td>
<td>J.J. Garcia-Luna-Aceves, UCSC (CNARC)</td>
</tr>
<tr>
<td></td>
<td>M. Magdon-Ismail, RPI (SCNARC)</td>
</tr>
<tr>
<td></td>
<td>Z. Wen, IBM (SCNARC)</td>
</tr>
</tbody>
</table>
Research on Mining, Searching, Evolution and Trust Analysis of Information Networks
IPP and APP Summary

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

Urbana-Champaign, Oct, 2010
News

08/14/10
- Zhenhui Li, Jiawei Han, Ming Ji, Lu-An Tang, Yintao Yu, Bolin Ding, Jae-Gil Lee, and Roland Kays, "MoveMine: Mining Moving Object Data for Discovery of Animal Movement Patterns", 

Research Areas
- Information network analysis
- Link-based data mining
- Graph and sequential mining
- Mining data streams

Graduating PhD 2011

Courses and Seminar
- CS412
- CS512
- CS591

Quick Link
Group Brochure
Research on Information Networks

- Search, OLAP, and query processing in info. Networks
- Knowledge discovery in information networks
- Moving object and cyber-physical networks
- Text mining in info. networks
- Evolution in information networks
- Looking forward for fruitful APP research
Search in Information Networks

- Neighborhood signatures of vertices are built to maintain indexing features: Effective search space pruning ability
- Processing (Query Decomposition): Decompose the query graph into a set of indexed shortest paths in Spath (VLDB’10)

**Network**

**Query**

**Histogram**

<table>
<thead>
<tr>
<th>distance</th>
<th>label</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>2</td>
</tr>
</tbody>
</table>

**ID-List**

<table>
<thead>
<tr>
<th>vid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>9</td>
</tr>
</tbody>
</table>

**A global lookup table**

**Neighborhood signature of v3**
Path Schema-Based Similarity Search in Information Networks

- **Motivation:** Similarity search in heterogeneous information networks
  - Applications: *Which terror group or event is similar to this one?*
- **Key ideas**
  - New similarity definition for heterogeneous information networks
    - **Feature space**
      - Traditional data: attributes denoted as numerical value/vector, set
      - Networked data: a relation/link sequence—“path schema”
    - **Measure defined on the feature space**
      - PathSim, not cosine, Euclidean distance, Jaccard coefficient, etc.
  - Y. Sun, J. Han, X. Yan, et al. “PathSim: Path Schema-Based Top-K Similarity Search in Heterogeneous Information Networks” (in submission)
- **Flickr pictures using image-tag-group paths:** (1) ITI vs. (2) ITIGITI

Which images are most similar to me?

Path: ITI

(a) top-1  
(b) top-2  
(c) top-3  
(d) top-4  
(e) top-5  
(f) top-6

Path: ITIGITI

(a) top-1  
(b) top-2  
(c) top-3  
(d) top-4  
(e) top-5  
(f) top-6

Find similar suspects in info-networks, based on user-specified similarity path!
iRIN: Image Retrieval in Image-Rich Information Networks (WWW’10 demo)

Flickr Image Search (query="navy")
Research on Information Networks

- Search, OLAP, and query processing in info. Networks
- Knowledge discovery in information networks
- Moving object and cyber-physical networks
- Text mining in info. networks
- Evolution in information networks
- Looking forward for fruitful APP research
RankCompete: A Competing Random Walk Model for Information Network Analysis [WWW’10]

- L. Cao, X. Jin, Z. Yin, et al. 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!
# Classification Accuracy: Labeling a Very Small Portion of Authors and Papers

## Comparison of classification accuracy on authors (%)

<table>
<thead>
<tr>
<th>(a%, p%)</th>
<th>nLB</th>
<th>wvRN</th>
<th>LLGC</th>
<th>GNetMine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A-A</td>
<td>A-C-P-T</td>
<td>A-A</td>
<td>A-C-P-T</td>
</tr>
<tr>
<td>(0.1%, 0.1%)</td>
<td>25.4</td>
<td>26.0</td>
<td>40.8</td>
<td>34.1</td>
</tr>
<tr>
<td>(0.2%, 0.2%)</td>
<td>28.3</td>
<td>26.0</td>
<td>46.0</td>
<td>41.2</td>
</tr>
<tr>
<td>(0.3%, 0.3%)</td>
<td>28.4</td>
<td>27.4</td>
<td>48.6</td>
<td>42.5</td>
</tr>
<tr>
<td>(0.4%, 0.4%)</td>
<td>30.7</td>
<td>26.7</td>
<td>46.3</td>
<td>45.6</td>
</tr>
<tr>
<td>(0.5%, 0.5%)</td>
<td>29.8</td>
<td>27.3</td>
<td>49.0</td>
<td>51.4</td>
</tr>
</tbody>
</table>

## Comparison of classification accuracy on papers (%)

<table>
<thead>
<tr>
<th>(a%, p%)</th>
<th>nLB</th>
<th>wvRN</th>
<th>LLGC</th>
<th>GNetMine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P-P</td>
<td>A-C-P-T</td>
<td>P-P</td>
<td>A-C-P-T</td>
</tr>
<tr>
<td>(0.1%, 0.1%)</td>
<td>49.8</td>
<td>31.5</td>
<td>62.0</td>
<td>42.0</td>
</tr>
<tr>
<td>(0.2%, 0.2%)</td>
<td>73.1</td>
<td>40.3</td>
<td>71.7</td>
<td>49.7</td>
</tr>
<tr>
<td>(0.3%, 0.3%)</td>
<td>77.9</td>
<td>35.4</td>
<td>77.9</td>
<td>54.3</td>
</tr>
<tr>
<td>(0.4%, 0.4%)</td>
<td>79.1</td>
<td>38.6</td>
<td>78.1</td>
<td>54.4</td>
</tr>
<tr>
<td>(0.5%, 0.5%)</td>
<td>80.7</td>
<td>39.3</td>
<td>77.9</td>
<td>53.5</td>
</tr>
</tbody>
</table>

## Comparison of classification accuracy on conferences (%)

<table>
<thead>
<tr>
<th>(a%, p%)</th>
<th>nLB</th>
<th>wvRN</th>
<th>LLGC</th>
<th>GNetMine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A-C-P-T</td>
<td>A-C-P-T</td>
<td>A-C-P-T</td>
<td>A-C-P-T</td>
</tr>
<tr>
<td>(0.1%, 0.1%)</td>
<td>25.5</td>
<td>43.5</td>
<td>79.0</td>
<td>81.0</td>
</tr>
<tr>
<td>(0.2%, 0.2%)</td>
<td>22.5</td>
<td>56.0</td>
<td>83.5</td>
<td>85.0</td>
</tr>
<tr>
<td>(0.3%, 0.3%)</td>
<td>25.0</td>
<td>59.0</td>
<td>87.0</td>
<td>87.0</td>
</tr>
<tr>
<td>(0.4%, 0.4%)</td>
<td>25.0</td>
<td>57.0</td>
<td>86.5</td>
<td>89.5</td>
</tr>
<tr>
<td>(0.5%, 0.5%)</td>
<td>25.0</td>
<td>68.0</td>
<td>90.0</td>
<td>94.0</td>
</tr>
</tbody>
</table>
# Knowledge Propagation: List Objects with the Highest Confidence Measure Belonging to Each Class

<table>
<thead>
<tr>
<th>No.</th>
<th>Database</th>
<th>Data Mining</th>
<th>Artificial Intelligence</th>
<th>Information Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>data</td>
<td>mining</td>
<td>learning</td>
<td>retrieval</td>
</tr>
<tr>
<td>2</td>
<td>database</td>
<td>data</td>
<td>knowledge</td>
<td>information</td>
</tr>
<tr>
<td>3</td>
<td>query</td>
<td>clustering</td>
<td>Reinforcement</td>
<td>web</td>
</tr>
<tr>
<td>4</td>
<td>system</td>
<td>learning</td>
<td>reasoning</td>
<td>search</td>
</tr>
<tr>
<td>5</td>
<td>xml</td>
<td>classification</td>
<td>model</td>
<td>document</td>
</tr>
</tbody>
</table>

## Top-5 terms related to each area

<table>
<thead>
<tr>
<th>No.</th>
<th>Database</th>
<th>Data Mining</th>
<th>Artificial Intelligence</th>
<th>Information Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Surajit Chaudhuri</td>
<td>Jiawei Han</td>
<td>Sridhar Mahadevan</td>
<td>W. Bruce Croft</td>
</tr>
<tr>
<td>2</td>
<td>H. V. Jagadish</td>
<td>Philip S. Yu</td>
<td>Takeo Kanade</td>
<td>Iadh Ounis</td>
</tr>
<tr>
<td>3</td>
<td>Michael J. Carey</td>
<td>Christos Faloutsos</td>
<td>Andrew W. Moore</td>
<td>Mark Sanderson</td>
</tr>
<tr>
<td>4</td>
<td>Michael Stonebraker</td>
<td>Wei Wang</td>
<td>Satinder P. Singh</td>
<td>ChengXiang Zhai</td>
</tr>
<tr>
<td>5</td>
<td>C. Mohan</td>
<td>Shusaku Tsumoto</td>
<td>Thomas S. Huang</td>
<td>Gerard Salton</td>
</tr>
</tbody>
</table>

## Top-5 authors concentrated in each area

<table>
<thead>
<tr>
<th>No.</th>
<th>Database</th>
<th>Data Mining</th>
<th>Artificial Intelligence</th>
<th>Information Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>VLDB</td>
<td>KDD</td>
<td>IJCAI</td>
<td>SIGIR</td>
</tr>
<tr>
<td>2</td>
<td>SIGMOD</td>
<td>SDM</td>
<td>AAAI</td>
<td>ECIR</td>
</tr>
<tr>
<td>3</td>
<td>PODS</td>
<td>PAKDD</td>
<td>CVPR</td>
<td>WWW</td>
</tr>
<tr>
<td>4</td>
<td>ICDE</td>
<td>ICDM</td>
<td>ICML</td>
<td>WSDM</td>
</tr>
<tr>
<td>5</td>
<td>EDBT</td>
<td>PKDD</td>
<td>ECML</td>
<td>CIKM</td>
</tr>
</tbody>
</table>

## Top-5 conferences concentrated in each area
Role Discovery: Advisor-Advisee

- Input: Research publication network
- Output: potential advising relationship & their ranking: \((r, [st, ed])\)
Research on Information Networks

- Search, OLAP, and query processing in info. Networks
- Knowledge discovery in information networks
- Moving object and cyber-physical networks
- Text mining in info. networks
- Evolution in information networks
- Looking forward for fruitful APP research
Discovery of Periodic Patterns of Moving Object Clusters

- A system that mines moving object patterns: Z. Li, et al., "MoveMine: Mining Moving Object Databases", SIGMOD’10 (system demo)
- Z. Li, B. Ding, J. Han, and R. Kays, “Mining Hidden Periodic Behaviors for Moving Objects”, KDD’10

Bird flying paths shown on Google Earth

- Z. Li, B. Ding, J. Han, and R. Kays, “Swarm: Mining Relaxed Temporal Moving Object Clusters”. VLDB’10 (sub)

Swarm discovers more patterns

- Convoy discovers only restricted patterns
Zhenhui Li, Ming Ji, Jae-Gil Lee, LuAn Tang, Yintao Yu, Jiawei Han, and Roland Kays, “MoveMine: Mining Moving Object Databases” (system demo), Proc. 2010 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'10), Indianapolis, Indiana, June 2010.
From Moving Objects to Cyber-Physical Networks

Lu-An Tang, Xiao Yu, Sangkyum Kim, Jiawei Han, Chih-Chieh Hung, and Wen-Chih Peng, “Tru-Alarm: Trustworthiness Analysis of Sensor Networks in Cyber-Physical Systems”, Proc. of 2010 Int. Conf. on Data Mining (ICDM'10), Sydney, Australia, Dec. 2010.
Research on Information Networks

- Search, OLAP, and query processing in info. Networks
- Knowledge discovery in information networks
- Moving object and cyber-physical networks
- Text mining in info. networks
- Evolution in information networks
- Looking forward for fruitful APP research
Effective OLAP Exploration

- TEXplorer (submitted to VLDB’ 10): integrating keyword-based ranking and OLAP exploration.

13.3. Google News Project: Object Identification and Network Analysis

Top 20 nodes
Research on Information Networks

- Search, OLAP, and query processing in info. Networks
- Knowledge discovery in information networks
- Moving object and cyber-physical networks
- Text mining in info. networks
- Evolution in information networks
- Looking forward for fruitful APP research
Mining Evolution & Dynamics of InfoNet

- From network sequences to evolutionary communities
Case Study on Delicious.com

Delicious Schema

Week 1 - Week 2:
- Security
- Terrorism
- Politics
- Travel
- USA
- Airport
- Israel
- Obama
- CIA
- Afghanistan

Week 2 - Week 3:
- Google
- China
- Internet
- Privacy
- Politics
- Censorship
- Facebook
- Business
- Terrorism

Week 3 - Week 4:
- Security
- Google
- China
- Internet
- Microsoft
- Privacy
- Censorship
- Politics
- Browser
- USA

Week 4 - Week 5:
- Digg
- Politics
- Datenschutz
- Facebook
- USA

Week 5 - Week 6:
- Ipad
- Apple
- Iphone
- Technology
- Tablet
- Mobile
- Newspapers
- Kindle
- Media

Week 6 - Week 7:
- Haiti
- Photography
- BBC
- Earthquake
- Photos
- UK
- 2010
- Disaster
- Travel
- Wildlife
E3.3. Evolutionary Clustering and Analysis of Heterogeneous Bibliographic Networks

Clustering are more consistent.

**Table 1: Consistency versus prior weight**

<table>
<thead>
<tr>
<th>Prior wt</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>0.108</td>
<td>2.432</td>
<td>1.199</td>
<td>1.342</td>
<td>2.818</td>
<td>5.273</td>
</tr>
<tr>
<td>Term</td>
<td>0.470</td>
<td>3.105</td>
<td>2.168</td>
<td>2.222</td>
<td>3.646</td>
<td>6.024</td>
</tr>
<tr>
<td>Conf</td>
<td>0.567</td>
<td>2.730</td>
<td>1.800</td>
<td>1.326</td>
<td>3.293</td>
<td>6.709</td>
</tr>
</tbody>
</table>

**Table 2: Quality Variation with Prior Weight**

<table>
<thead>
<tr>
<th>Prior wt</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compactness</td>
<td>4594.17</td>
<td>2166.86</td>
<td>1978.39</td>
</tr>
<tr>
<td>Prior wt</td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
</tr>
<tr>
<td>Compactness</td>
<td>2932.19</td>
<td>4267.58</td>
<td>3972.62</td>
</tr>
</tbody>
</table>

**Dataset**
- DBLP (1993 to 2008, 654K papers, 484K authors, 107K title terms and 3900 conferences)
- Four area (DM, DB, IR, ML papers; 1993 to 2008, 24K papers, 26K authors, 12K title terms, 20 conferences)
Research on Information Networks

- Search, OLAP, and query processing in info. Networks
- Knowledge discovery in information networks
- Moving object and cyber-physical networks
- Text mining in info. networks
- Evolution in information networks
- Looking forward for fruitful APP research
Recent Research Publications

- (Li, et al., 2010b) Li, Z.; Lee, J.-G.; Li, X.; Han, J., “Incremental Clustering for Trajectories”, *Proc. 2010 Int. Conf. on Database Systems for Advanced Applications (DASFAA’10)*, Tsukuba, Japan, April 2010.
- (Gao, et al., 2010) Gao, J. (UIUC); Liang, F. (UIUC); Fan, W. (IBM); Wang, C. (UIUC); Sun, Y. (UIUC); Han, J. (UIUC), “On Community Outliers and their Efficient Detection in Information Networks “, accepted by **KDD’10**
- (Wang, et al., 2010) Wang, C. (UIUC); Han, J. (UIUC); Jia, Y. (UIUC); Zhang, D. (UIUC); Yu, Y. (UIUC); Tang, J. (Tsnghua U); Guo, J. (Tsnghua U.), “Mining Advisor-Advisee Relationships from Research Publication Networks”, accepted by **KDD’10**
- (Li, et al., 2010d) Li, Z. (UIUC); Ding, B. (UIUC); Han, J. (UIUC); Kays, R. (New York State Museum), “Mining Hidden Periodic Behaviors for Moving Objects”, accepted by **KDD’10**
- (Gupta, et al, 2010) Gupta, M.(UIUC), Aggarwal, C. (IBM), Han, J. (UIUC), Sun, Y. (UIUC), "Evolutionary Clustering and Analysis of Heterogeneous Information Networks", submitted.
- (Ji, M. et al., 2010) Ji, M. (UIUC); Sun, Y. (UIUC); Danilevsky, M. (UIUC); Gao, J. (UIUC); Han, J. (UIUC), “Graph-based Classification on Heterogeneous Information Networks”, submitted.
- (Li, et al., 2010c) Li, Z. (UIUC); Ding, B. (UIUC); Han, J. (UIUC); Kays, R. (New York State Museum), “Swarm: Mining Relaxed Temporal Moving Object Clusters”, submitted for publication, 2010.
- (Sun, et al., 2010) Sun, Y. (UIUC); Han, J. (UIUC); Wu, T. (UIUC); Yan, X. (UCSB); Yu, P. S. (UIUC); Liu, L. (UIUC), “PathSim: Path SchemaBased TopK Similarity Search in Heterogeneous Information Networks”, submitted.
Research Awards, Tutorials, and Invited Talks

- **Conference tutorial**: Jiawei Han, Zhenhui Li, and Lu An Tang, “Mining Moving Object, Trajectory and Traffic Data”, Conference Tutorial of 2010 Int. Conf. on Database Systems for Advanced Applications (DASFAA’10), Japan, April 2010.

- **Conference tutorial**: Jing Gao, Wei Fan, and Jiawei Han, “On the Power of Ensemble: Supervised and Unsupervised Methods Reconciled”, Conference Tutorial of 2010 SIAM Int. Conf. on Data Mining (SDM’10), Columbus, OH, April 2010.

- **Conference tutorial**: Jiawei Han, Yizhou Sun, Xifeng Yan, and Philip S. Yu, “Mining Knowledge from Databases: An Information Network Analysis Approach” (tutorial), Proc. 2010 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD’10), Indianapolis, Indiana, June 2010.

- **Distinguished seminar**: Jiawei Han, “Mining Heterogeneous Information Networks By Exploring the Power of Links”, Department of Computer Science, Univ. of California at San Diego, Feb. 2010.

- **Distinguished seminar**: Jiawei Han, “RankClus: Integrated Clustering and Ranking at Mining Heterogeneous Information Networks”, Department of Computer Science, Univ. of North Carolina at Charlotte, Feb. 2010.

- **NSF IGERT invited talk**: Jiawei Han, “Mining Moving Object and Traffic Data”, Department of Computer Science, Univ. of Illinois at Chicago, March 2010.

- **Conference keynote speech**: Jiawei Han, ECML/PKDD 2010 (Barcelona, Spain)

- **IBM Ph.D. Fellowship**: Jing Gao (2010-2012)  

- **DOE and NSF Ph.D. Scholarship/Fellowship**: Tim Weninger (2009-2011, …)
Thank you

May, 2010, Urbana-Champaign
Discovering Structures on the Web

WWW’10 – CETR: Content Extraction via Tag Ratios

Table 4: F1-measures for each algorithm in each source. Winners are in bold.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>CleanEval-Eng</th>
<th>CleanEval-Zh</th>
<th>CleanEval</th>
<th>Myriad 40</th>
<th>Big 5</th>
<th>NyTimes</th>
<th>BBC</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>CETR-TM</td>
<td>94.10%</td>
<td>86.55%</td>
<td>90.33%</td>
<td>91.44%</td>
<td>81.27%</td>
<td>97.11%</td>
<td>98.26%</td>
<td>91.45%</td>
</tr>
<tr>
<td>CETR-KM</td>
<td>94.68%</td>
<td>86.50%</td>
<td>90.50%</td>
<td>94.17%</td>
<td>91.23%</td>
<td>98.40%</td>
<td>97.01%</td>
<td>93.66%</td>
</tr>
<tr>
<td>CETR</td>
<td>94.72%</td>
<td>86.62%</td>
<td>90.67%</td>
<td>94.72%</td>
<td>91.82%</td>
<td>98.21%</td>
<td>97.46%</td>
<td>93.93%</td>
</tr>
</tbody>
</table>

WWW’10 – WebTables: Outer join with tables on the Web via links

KDD Explorations Dec. Issue (sub) – Explores problems in finding lists on the Web

Planned:

TransferPaths@WWW’11 – Learning structures for entity page discovery

VisualLists@WWW’11 – Automatically extracting general lists on the Web
Summary and Looking Forward

- Data Mining Group has been making good progress and fruitful results on Information Network Research
- We are working on many exciting research frontiers
- Forming APP brainstorming and proposal discussion groups
- Looking forward to collaborations with other groups at UIUC, other sites at INARC, other NS-CTA networks, and ARL
Discovering and Mapping Entity Pages

WSDM’11 (sub) – T. Weninger, F. Fumarola, et al.: Find Entity Pages by growing parallel paths:

<table>
<thead>
<tr>
<th>Domain</th>
<th>Num. Entities</th>
<th>Baseline</th>
<th>Precision</th>
<th>Recall</th>
<th>k-Shortest Paths</th>
<th>Precision</th>
<th>Recall</th>
<th>Path Removal</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS Faculty</td>
<td>1,410</td>
<td>1.0</td>
<td>1.1</td>
<td></td>
<td>81.6</td>
<td>34.4</td>
<td>96.0</td>
<td>82.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UIUC CS Courses</td>
<td>84</td>
<td>7.0</td>
<td>8.3</td>
<td></td>
<td>96.7</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UIUC CS Groups</td>
<td>36</td>
<td>9.0</td>
<td>23.0</td>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Representatives</td>
<td>441</td>
<td>12.0</td>
<td>11.0</td>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Senators</td>
<td>100</td>
<td>17.0</td>
<td>16.0</td>
<td></td>
<td>55.3</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Senate Committees</td>
<td>40</td>
<td>2.0</td>
<td>5.0</td>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>House Committees</td>
<td>45</td>
<td>6.0</td>
<td>13.3</td>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Computer science department person-entity retrieval results

Table 1: Baseline mapping results.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Google MRR</th>
<th>Prec. @1</th>
<th>Adjacent Only Prec.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBLP</td>
<td>.7516</td>
<td>71.58</td>
<td>74.19</td>
</tr>
<tr>
<td>IMDB</td>
<td>.4240</td>
<td>23.96</td>
<td>38.97</td>
</tr>
</tbody>
</table>

Table 2: Link paths mapping results

<table>
<thead>
<tr>
<th>Domain</th>
<th>Strict Matching</th>
<th>K-short</th>
<th>Loopless</th>
<th>≈ Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec.</td>
<td>Recall</td>
<td>Prec.</td>
<td>Recall</td>
</tr>
<tr>
<td>DBLP</td>
<td>96.34</td>
<td>52.86</td>
<td>98.63</td>
<td>54.49</td>
</tr>
<tr>
<td>IMDB</td>
<td>99.69</td>
<td>37.70</td>
<td>99.69</td>
<td>41.85</td>
</tr>
</tbody>
</table>

Figure 2: Precision and Recall tradeoff for DBLP data with K-shortest loopless paths as λ varies from 0 to 1.
Mining Multi-Dimensional Text Databases

- Multi-dimensional Text Data: text associated with multi-dimensional structured attributes
  - News (Text + Time, Location, other entities)
  - Product Reviews (Text + Price, Brand, Model, etc..)
  - Social Network Updates (Text + Author, Time, etc..)
  - Pilot Reports (Text + Sensor, Airport, etc..)

Scalable Data Cube Systems:
- TextCube (ICDM08),
- TopicCube (SDM09),
- iNextCube (VLDB09 demo)

Knowledge Discovery:
- Topic Tracking (KDD10),
- Correlated Topics (CIDU10),
- Aspect Detection (ongoing),
- Entity Resolution (ongoing).

Effective Data Access:
- Object Ranking: TopCells (ICDE10, CIDU10)
- Object Exploration: TEXplorer (to ICDE11)
Effective OLAP Exploration and Event Tracking

- **TEXplorer**: Integrating keyword-based ranking and OLAP exploration
- **PET (KDD’ 10)**: tracking popularity and textual representation of events in social communities (twitter)