Information Network Research I3: Knowledge Discovery in Information Networks

Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign
Project I3: Knowledge Discovery in Information Networks

- **Goal**
  - Develop efficient and effective knowledge discovery mechanisms that automatically uncover patterns, knowledge, and other critical information in information networks for military applications

- **Organization of Project Tasks**
  - Task I1: Methods for scalable mining of dynamic, heterogeneous information networks (J. Han, X. Yan, and C. Faloutsos) [Collaborators: M. Faloutsos (UCR, IRC), J. Hendler (RPI, IRC) and B. Szymanski (RPI, SCNARC)]
  - Task I2: Real-time methods for mining spatiotemporal information-related cyber-physical networks (S. Papadimitriou, J. Han, and X. Yan) [Collaborators: S. Adali (RPI, SCNARC), T. LaPorta (PSU, CNARC)]
  - Task I3: Text and unstructured data mining for information network analysis (D. Roth, J. Han and H. Ji) [Collaborators: M. Faloutsos (UCR, IRC) and B. Szymanski (RPI, SCNARC)]

**Project Lead: Jiawei Han (UIUC)**
IPP I3.1: Mining Dynamic, Heterogeneous Info. Networks

- Classification of heterogeneous information networks
  - Background: Rank-based Clustering: RankClus, NetClus, iNextCube
  - GNetClass: Classification on heterogeneous info. net (ECMLPKDD’10)
- Pattern discovery in evolutionary heterogeneous info. networks
  - Discovery of hidden semantic relationships (e.g., identifying leaders and their followers by mining records of group activities) (KDD’10)
  - Evolution of heterogeneous information networks (MLG-10)
  - Graph association patterns in info. networks (SIGMOD’10 by UCSB & IBM)
- Anomaly and outlier discovery in dynamic heterogeneous info. networks
  - Anomaly discovery in large information networks (PAKDD’10: Christos Faloutsos)
  - Detection of community outliers in information networks (KDD’10)

Han (UIUC), Yan (UCSB), Faloutsos (CMU)
Classification: Knowledge Propagation

M. Ji et al, GNetMine, ECMLPKDD’10
Role Discovery: Discovery of Advisor-Advisee Relationships

- Input: Research publication network
- Output: potential advising relationship & their ranking: \((r, [st, ed])\)
Text and Network Modeling in Composite Information and Social Networks

A Sample Ticket Info Flow

<table>
<thead>
<tr>
<th>ID</th>
<th>Day</th>
<th>Entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>81</td>
<td>05-14</td>
<td>New Ticket: DB2 login fail</td>
</tr>
<tr>
<td>81</td>
<td>05-14</td>
<td>Transfer to Group SMRDX</td>
</tr>
<tr>
<td>81</td>
<td>05-14</td>
<td>Contacted Mary for recycling</td>
</tr>
<tr>
<td>81</td>
<td>05-14</td>
<td>Transfer to Group SSDSISAP</td>
</tr>
<tr>
<td>81</td>
<td>05-14</td>
<td>Status updated ...</td>
</tr>
<tr>
<td>81</td>
<td>05-15</td>
<td>Transfer to Group ASWWCUST</td>
</tr>
<tr>
<td>81</td>
<td>05-15</td>
<td>Web service checking</td>
</tr>
<tr>
<td>81</td>
<td>05-18</td>
<td>Could not solve the problem.</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>81</td>
<td>05-18</td>
<td>Transfer to Group SSSAPHWOA</td>
</tr>
<tr>
<td>81</td>
<td>05-22</td>
<td>Resolved</td>
</tr>
</tbody>
</table>

Composite Networks with Dynamic Information Flows

Yan's Group (UCSB)
IPP I3.2: Real-Time Methods for Spatiotemporal Information-Related Cyber-Physical Network Analysis

Objectives and Collaborations

- Knowledge discovery in cyber-physical networks ⇒ Interact with CNARC
- Spatiotemporal information-centered clustering ⇒ Interact with CNARC and SCNARC when modeling needs spatial-based reasoning
- Evolution analysis of cyber-physical networks ⇒ EDIN and IRC

Recent Progress and On-going work

- Mining knowledge from massive moving objects
  - MoveMine (SIGMOD’10 demo), SWARM pattern mining (VLDB’10), Periodica (Periodic movement pattern mining) (KDD’10)
- Online mining of dynamic and evolving cyber-physical networks
  - Tru-Alarm (ICDM’10)
- Exploring military applications in cyber-physical networks

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

← Bird flying paths shown on Google Earth
Mined periodic patterns by our new method →

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.
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.
Task I3.3: Mining Text and Unstructured Data for InfoNet Analysis

Motivation and Goals

- Most military applications need to handle text data, including documents, e-mails, telecommunication messages, micro blogs and conversations
- Text and unstructured data form a critical part of information networks
- Multidimensional analysis of information networks associated with text data
- Combine text mining and network analysis and leverage the power of statistical topic modeling and discrete regularization

Recent Research Progress

- Multi-dimensional text information network analysis
  - **Text Cube, Topic Cube, iTopicModel**: Information network enhanced topic modeling
  - Dynamic language modeling:
- Text information extraction: **Wikification** (Linking entities to Wikipedia)
- **Google News analyzer**: Integrate text mining and information network analysis

Roth (UIUC), Han (UIUC), Heng Ji (CUNY)
Google News Analyzer and Text Mining

- **Motivation:** Military missions need to analyze a lot of textual and unstructured data

- **Google News Analyzer**
  - Integration of (1) information retrieval, (2) natural language processing, (3) multi-dimensional text cube, (4) information network analysis, and (5) text information integration and information quality

- **Recent progress**
  - Text/topic cube construction and usage
  - TopCell: Query-based ranking of text cube (ICDE’10)
  - MiTexCube: multidimensional text clustering

- **On-Going Research:**
  - Construct a Google News Analyzer for automated online multidimensional news analysis
Generative models for ticket resolution in expert networks

A set of tickets reported to the expert network

\[ T = \{ t_1, t_2, \ldots, t_m \} \]

Word description of tickets

\[ \mathcal{W} = \{ w_1, w_2, \ldots, w_n \} \]

An interconnected network of experts

\[ \mathcal{G} = \{ g_1, g_2, \ldots, g_L \} \]

Goal: Minimize the average length of routing sequences

\[ S = \sum_{i=1}^{m} \frac{|R(t_i)|}{m} \]

Routing sequence of tickets

\[ R(t) = g_{init}(t) \rightarrow \ldots \rightarrow g_{res}(t) \]

Generative models for ticket resolution in expert networks
Cross-Document Information Extraction, Tracking and Summarization

- **Motivation**: Most current Information Extraction analyzes single documents from single media in isolation. Net result is a set of unconnected, unranked, redundant, erroneous, and incomplete facts

- **Cross-document Information Extraction, Tracking and Summarization**
  - Cross-document information inference based on Markov Logic Networks (Ji et al., RANLP2009)
  - Improving Word Sense Disambiguation Using Web-Scale Phrase Clustering (Ji, IUCS2010)
  - Novel event discovery and Domain Adaptation (Ji, NAACL2009; Li et al., PACLIC2010, Ji et al., LREC2010)
  - Knowledge Discovery based on Web-scale Ngrams (Ji and Lin, 2009; 2010; Lin et al., 2010 Journal of LRE)
  - Graph-clustering based Cross-document Event Coreference Resolution (Chen and Ji, 2009; 2010)
  - Global implicit time discovery and reasoning (Gupta and Ji, ACL2009; Ji et al., HLT2010)
  - Information-aware Cross-document Summarization (Ji et al., 2010 MMIES Book Chapter; Ji et al., AIRS2010 Nominated as the Best Paper)

H. Ji’s Group (CUNY)
Knowledge Base Population

- **Goal**
  - Given a large collection (+1M) of news and web documents, and a small knowledge base, and any input queries
  - Automatically disambiguate entities, distill information about entities (e.g. birth-dates, spouse, employment, headquarter, residence), and expand a knowledge base (e.g. Wikipedia Infoboxes)
  - Rapid adaptation to new relation mining

- **Approach**
  - Combining pattern matching, information extraction and question answering (Parton et al., 2009)
  - Enhance by automatic query expansion, answer re-ranking and cross-slot reasoning (Chen et al., AIRS2010)

- **Results**
  - Top 5 among 33 teams in NIST TAC2010 evaluation
Research Publications, Awards and Out-Reach

- Christos Faloutsos: Test of Time Award for their SIGCOMM’1999 paper
- Christos Faloutsos: ACM 2010 SIGKDD Innovations Award
- Christos Faloutsos: Conf. Keynote Speech @ECML/PKDD'10, ICDM’10
- Jiawei Han: IEEE Computer Society, W. Wallace McDowell Award
- Jiawei Han: Conference Keynote Speech@ECML/PKDD'10
- Jiawei Han: Conference tutorial: DASFAA'10, SDM’10, SIGMOD’10, KDD’10
- Heng Ji: NSF Career Award
- Xifeng Yan: NSF Career Award
- Dan Roth: Fellow of AAAI
- Student Awards: IBM Ph.D. Fellowship: Jing Gao (2010-2012), DOE and NSF Ph.D. Scholarship/Fellowship: Tim Weninger (2009-2011, ...)
- Research paper publications (> 30)
  - Covers most major conferences in data mining, database systems, machine learning, natural language, WWW, social/info. network analysis, etc.
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>
Thanks!