INARC_Task I3 Network Construction and Reliable Knowledge Synthesis with Noisy and Untrustworthy Text Data

Project I3: Year 4: Task I3: Network Construction and Reliable Knowledge Synthesis with Noisy and Untrustworthy Text Data

Project Co-Lead: Dan Roth and Jiawei Han, UIUC  
Email: danr@cs.uiuc.edu, hanj@cs.uiuc.edu  
Phone: 217-333-6903

Primary Research Staff

<table>
<thead>
<tr>
<th>Collaborators</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. Roth, UIUC (INARC)</td>
</tr>
<tr>
<td>J. Han, UIUC (INARC)</td>
</tr>
<tr>
<td>H. Ji, CUNY (INARC)</td>
</tr>
<tr>
<td>X. Yan, UCSB (INARC)</td>
</tr>
</tbody>
</table>

| D. Basu, BBN (IRC) |
| J. Hendler, RPI (SCNARC, IRC) |
| N. Chawla, ND (SCNARC) |
| B. Uzzi, Northwestern (SCNARC) |
| L. Bowman ARL (S, Kase, ARL; Kirk, ARL) |
| L. Kaplan, ARL |
| C. Voss, ARL |

Most real-world data is unstructured, interconnected and noisy, and often in the form of text. This raises technical challenges including that of constructing from such unstructured data an organized, semi-structured network of information, enriching multi-genre networks, and mining advanced knowledge from these constructed and enriched networks. All these challenging tasks have broad military application.

Built on the multi-year success of Task I3.2, our Year-4 missions are: (1) constructing high quality, semi-structured, multi-genre networks by role and structure discovery, and network enhanced information distillation/consolidation from noisy, unstructured, military data; (2) synthesizing reliable knowledge from unreliable information distillation to detect malicious or adversary statements and activities using truth modeling; and (3) opinion mining from signed multi-genre information networks, which is based on social network's balance theory and information-network enhanced topic modeling/inference. Subtask 1 investigates general principles of information network construction. Subtask 2 exploits meta-models of leveraging information networks for information retrieval and knowledge discovery. Subtask 3 examines a key application of (signed) information networks—opinion mining, an increasingly important problem in social/information networks. We will use news data, Twitter corpus, Wikipedia, discussion forums and web blogs to validate our methods and models. This project will be linked closely with dynamic graph query processing (I2), the new IPAN (Information Processing across Networks) Thrust, and multiple SCNARC (on text-rich social network analysis), Trust (on trust analysis), and EDIN (on network evolution) tasks, and with IRC on experimentation and technology transfer.

In most military scenarios, an information network is usually integrated with rich text and unstructured data, which becomes a text-rich information network. We will extend existing studies to mine knowledge from these types of networks. Unlike most information retrieval and text mining studies, our proposed integrated text and information network analysis will (i) structurally model a text-rich information network by progressive network analysis on text and interconnected data, (ii) enrich text data analysis with such networks, and (iii) facilitate multi-facet search in such networks.

This project aims to generate a set of effective methods for QoI mining and text-rich information network modeling of heterogeneous information network. The research will be closely linked with SCNARC to explore its potential in social network analysis, and will be integrated with IRC’s experimentation for technology transfer.

Subtask 1. Construct high-quality, semi-structured, multi-genre networks from noisy data

Subtask 2. Synthesizing reliable knowledge from unreliable information network construction to detect malicious or adversary statements and activities using truth modeling

Subtask 3. Opinion mining from signed multi-genre information networks

Oct. 15 (Monday), 2012.  (Teleconf., meeting)  Discussion on the general on-going research and research plan of Year 4.

Jiawei Han [Presentation slides]

Heng Ji [Presentation slides]

Dan Roth [Presentation slides]

Next week schedule: 2-3pm ET (1-2pm CT/11-12pm PT) Mov. 12 (Monday), 2012.  (Teleconf., meeting)
(1) Xifeng’s research plan briefing, (2) discussion on the research experiment platform by Dan Roth

2011-12012

Meeting Jan. 17 2011:  1-2pm CT Teleconf.  Briefing the current work and work plan (each 12 minutes on briefing + discussions)

Jiawei Han [Presentation slides]

Heng Ji [Presentation slides]

Dan Roth [Presentation slides]

Xifeng Yan [Presentation slides]

CMU [Presentation slides]

Meeting March 8, 2011:  3-4pm CT Teleconf.  Discussion on Task-level Midyear Review

Jiawei Han [Presentation slides (I3.1)  (I3.2) ]

Heng Ji [Presentation slides (I3.2) ]

Dan Roth [Presentation slides (I3.2) ]

Xifeng Yan [Presentation slides (I3)]

CMU [ ]

Task-Level Presentations

Task I3.1: QoI Mining of Noisy, Volatile, Uncertain, and Incomplete Heterogeneous Information Networks (Faloutsos (CMU), Han (UIUC), Yan (UCSB))

Task I3.2: Modeling and Mining of Text-Rich Information Networks (Han (UIUC), Ji (CUNY), Roth (UIUC), Yan (UCSB))

Tasks:

Task I3.1: QoI Mining of Noisy, Volatile, Uncertain, and Incomplete Heterogeneous Information Networks (Faloutsos (CMU), Han (UIUC), Yan (UCSB))

Task I3.2: Modeling and Mining of Text-Rich Information Networks (Han (UIUC), Ji (CUNY), Roth (UIUC), Yan (UCSB))