I3.2. A Network-Integrated Approach for Construction, Search, and Mining of Text-Rich Multi-Genre Networks

Technical Abstract

Text, in the form of twitters, blogs, news, text messages, reports, etc., has been extensively used by soldiers and commanders in military applications for information gathering and exchange. Textual information is not only ubiquitous but also essential for understanding dynamic situations and making critical decisions. To make effective use of text information in multi-genre network analysis, we propose to investigate mechanisms for effective construction, search and mining of text-rich multi-genre networks, consisting of information, social, and communication networks associated with rich text information in their nodes and links. Based on our recent rich results, we propose to study the following structured framework: (i) construction: from text and web to relatively structured, multi-genre heterogeneous networks by progressive network analysis, information extraction, and text mining, (ii) search: multi-dimensional search in text-rich multi-genre networks; and (iii) mining: topic propagation across multi-genre networks. The novelty of this study is that it develops a network-integrated approach to search and analysis of text-rich multi-genre networks (none of the other NSCTA task is doing network-integrated text search or analysis), which departs from traditional approaches on entity extraction, information retrieval and text mining, and leads to new analytical power in information understanding and network analysis.

Research Issue/Technical Approach

Roadmap: This task is a critical component in the Network Discovery thrust and will contribute to three other thrusts: Trust, EDIN, and QoI, each of which needs to deal with network-based text information. It addresses three key issues: (i) measure (on the effectiveness of modeling and mining), (ii) adapt (to the changes of situations), and (iii) predict (using mining results).

Key Research Questions and Initial Hypotheses

Q1: Will the traditional approaches work well in text-rich multi-genre networks on entity extraction, information retrieval and text mining?

Although text documents/messages are ubiquitous and are interconnected, most traditional text analyses do not explore networked information, but only study terms/topics in individual documents. However, since many mission-related text messages, similar to tweets, are brief and include limited semantic contexts, isolated analysis cannot be effective and a network-integrated approach becomes essential.

H1: A network-integrated approach is superior over traditional approaches in search and analysis of text-rich information networks

We will demonstrate the network-integrated approach brings strong analytical power, such as uncovering the underlying text-rich networks (e.g., adversary networks) as well as the implicit semantics in text and interconnected data.

Q2: What new analytical power the new approach may bring to the understanding and analysis text-rich multi-genre networks?

Our preliminary studies in APP Y2 show that combining text and heterogeneous information network analysis enhances topic modeling, object clustering and event discovery. We will extend our scope to multi-genre networks, especially on how the new approach may help in-depth analysis of social networks and communication networks.

H2: A network-integrated approach for modeling text-rich multi-genre network will enhance the modeling and analysis of communication, information, and social networks as well.

Research Problem Background

Text is extensively used for communications and information gathering in military applications. Textual documents are not only abundant, but also interconnected in various ways, forming text-rich multi-genre networks. Information retrieval and text mining [Manning2008, Baeza-Yates2011] have been studied extensively in text data analysis.

However, most previous studies either ignore the interplay between textual topics and network structures or merely integrate them with homogeneous networks. Since text data is usually associated with some explicit or implicit heterogeneous information networks, it is important to explore how text understanding and information network analysis can mutually enhance each other. Our recent studies have been moving towards this direction [Lin2008, Mei2008, Sun2009c, Zhang2009, Deng2011]. By incorporating information network structures into topic modeling, we achieved 16%-25% higher accuracy over state-of-the-art PLSA approach in both scientific paper domain and general news domain [Deng2011]. The enhanced topic modeling method then yielded 10% absolute F-measure gains in text mining tasks, such as event extraction, with significant reduction (75%) in annotation cost. Certainly, there is greater potential to extend this novel thread to thoroughly explore text-rich multi-genre networks.

Technical Approach

We propose to develop a new network-integrated approach and its associated theory for construction, search, and mining of text-rich, structured, multi-genre networks.

Figure1: A Network-Integrated Approach for Construction of Text-Rich Multi-Genre Networks

Subtask 1. Construction of Structured, Text-Rich Multi-Genre Networks: We proposed to develop a network-integrated approach for
Although mining can be done on many aspects, in Y3, in structured, text-rich multi-genre networks, that is, to study how information flows through communication networks. Therefore, this task is essential to maximize the use of text data and text-rich networks for building intelligent military systems, which will help derive semantic structures of the network as well.

Subtask 2. Multi-Dimensional Search in Structured Multi-Genre Networks: Our recent studies have shown that information search can be enhanced by exploration of heterogeneous information networks, such as, iTopicModel [Sun2009c] and biased topic propagation [Deng2011]. Previous studies performed topic modeling by examining individual documents. However, by viewing data as a text-rich heterogeneous (i.e., multi-typed) information network, as shown in Figure 1, one may explore the interactions among multi-typed objects, i.e., the rich semantics of a network, to enhance text modeling. For example, the discovery of latent topics and clustering of multi-typed objects can be mutually enhanced by biased topic propagation over heterogeneous information networks [Sun2009c, Deng2011], and trustworthiness of conflict information provided by multiple information providers can be evaluated by network analysis [Yin2008, Pasternack08, Pasternack2010]. In Year 3, we propose to investigate search mechanisms in text-rich networks as follows: (i) developing semantics-based topic search by exploitation of new semantics of both communication and social networks (e.g., the topics a user tends to associate may be affected by his friends or professional social groups, and topics could be influenced by spatiotemporal locality and communication locality), and (ii) exploring multidimensional search in structured, text-rich multi-genre networks, e.g., expert search, that explores ranking and semantic similarity in structured networks guided by meta paths.

Subtask 3. Mining Topic Propagation in Structured Text-Rich Multi-Genre Networks: Although mining can be done on many aspects, in Y3, we plan to focus on topic propagation in structured, text-rich multi-genre networks, that is, to study how information flows through heterogeneous networks, such as how a research topic (i.e., a cluster of terms) propagates through authors, papers, venues and other topics. This study will be tightly integrated with communication and social networks: Some rumors or secrets could be propagated (or leaked) via word-of-mouth, through certain people or communication channels; but general topics, such as breaking news could be propagated via internet, or social networks, such as Facebook and Twitter. In heterogeneous networks, top propagation will also disclose how the information flows through multi-typed entities, such as person, organization, location, communication channel, and sub-networks. To study topic propagation, semantic clustering of similar terms, top ranking, entity/link ranking (such as authors, venues, sub-network structures) will be explored. We will extend our newly developed RankClass framework [Ji2011]; geo-friends discovery method [Yu2011] and GPS/timestamp-associated pattern mining and topic modeling [Yin2011] (which studies geo-tagged photos and tweets sent from iPhones with timestamps) to compare information propagation across different network structures (e.g., topic propagation in information networks vs. in social networks).

Validation Approach

Our work will be evaluated using a variety of large, real data sets, including (i) 1 million news articles from New York Times, (ii) 10 million news and blog documents from Gigaword corpora, with rich entity and attribute annotations from NIST knowledge base population program 2009-2011 [Ji2011], (iii) 1.4 million tweets gathered from Twitter.com, and (iv) NIST 2002-2009 training corpora [Ji2009, Parton2009]. We will conduct two evaluations: (1) intrinsic evaluation, e.g., the impact of network structure on enhanced network structuring, search and text mining tasks; and (2) extrinsic evaluation, e.g., the impact of enhanced topic modeling and mining on the construction of improved network structures. In the meantime, we will work together with the IRC team to identify military needs in text analysis and simulate such kinds of data sets based on the collected corpora.

Key Deliverables and Planned Research Outputs

| Q1 | Methodology design & algorithm development for Subtask 1 (UIUC, UCSB, CUNY) |
| Q2 | Methodology design & algorithm development for Subtask 2 (UIUC, UCSB, CUNY) |
| Q3 | Methodology design & algorithm development for Subtask 3 (UIUC, CUNY, UCSB) |
| Q4 | Research paper publication & system prototype demo of the approaches (UIUC, CUNY, UCSB) |

Military Relevance/Impact on Network Science

This task investigates principles, methodologies and tools for construction, search and mining of text-rich multi-genre networks. It will develop powerful text analysis techniques for military personnel to obtain network-based semantic information, well beyond keyword-based information retrieval. It will play an important role in situation understanding, and detecting adversaries and other potential threats by exploring text-rich integrated networks. Therefore, this task is essential to maximize the use of text data and text-rich networks for building intelligent military system for threat assessment and efficient command & control.
Transition Opportunities

The work will generate new principles, methodologies, algorithms and implementations for construction, search and mining of text-rich multi-genre networks. We will facilitate transition of our research results to application-oriented (6.2) research at IRC as well as potential transition to ARL. As far as we know, ARL researchers have been exploring our tools developed at IPP and APP (Y2). We will be actively exploring such technology transition opportunities by working with ARL researchers, IRC and INARC experimentation and technology transfer contacts.

Dependencies

This task is a continuation of Task I3.2 of the APP in INARC, which invented a methodology of network-integrated text analysis, and generated many research papers. One important novelty of our approach on text modeling is that other NSCTA tasks simply treat documents as bags of words without exploring network structures whereas we explore inter-document network structures in topic modeling, which is shown semantically powerful and less sensitive to noises. Y3 will push this frontier forward substantially.

This task is linked closely with the following tasks: (1) "S2.1: Adversary Social Networks, Detection, Evolution, Stability and Hierarchy". Our study can help identify context-specific information flows and internal hierarchies in discovered groups; identify relationships between communities, by exploring network-integrated text modeling. (2) "T2.2: Cognitive Models of Trust", and "T2.4: Network Behavior-based Indicators of Trust in Composite Social and Information Networks". Our model can provide deeper methods for understanding text data over social/information networks, and can benefit from the output of the TRUST tasks to obtain better semantics of network structures. Also, we will collaborate with C1.3 team on integrating communication networks in this study. Specifically, we will develop collaborations with Adali, Magdon-Ismail, Pirolli, BayNoy and their teams.

Bibliography


[Deng2011] Hongbo Deng, Jiawei Han, Bo Zhao, Yintao Yu, Cindy Xide Lin, "Probabilistic Topic Models with Biased Propagation on Heterogeneous Information Networks", in Proc. of 2011 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'11), 2011.

[Ding2010] Bolin Ding, Bo Zhao, Cindy Xide Lin, Jiawei Han, and Chengxiang Zhai, "TopCells: Keyword-based Search of Top-k Aggregated Documents in Text Cube", in Proc. of 2010 Int. Conf. on Data Engineering (ICDE'10), 2010.


[Lin2008] Cindy Xide Lin, Bolin Ding, Jiawei Han, Feida Zhu, and Bo Zhao, "Text Cube: Computing IR Measures for Multidimensional Text Database Analysis", Proc. 2008 Int. Conf. on Data Mining (ICDM'08), Pisa, Italy, Dec. 2008.

[Lin2010] C. X. Lin, B. Zhao, Q. Mei, and J. Han, "A Statistical Model for Popular Event Tracking in Social Communities", in Proc. of 2010 ACM Int. Conf. on Knowledge Discovery and Data Mining (KDD'10), 2010.


[Sun2009c] Yizhou Sun, Jiawei Han, Jing Gao, and Yintao Yu, "iTopicModel: Information Network-Integrated Topic Modeling", Proc. of 2009 Int.
[Wang2010] Chi Wang, Jiawei Han, Yuntao Jia, Jie Tang, Duo Zhang, Yintao Yu, and Jingyi Guo, "Mining Advisor-Advisee Relationships from Research Publication Networks", Proc. 2010 ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD'10), Washington D.C., July 2010.

[Weninger2011] Tim Weninger, Marina Danilevsky, Fabio Fumarola, Joshua Hailpern, Jiawei Han, Ming Ji, Thomas J. Johnston, Surya Kallumadi, Hyungsul Kim, Zhijin Li, David McCloskey, Yizhou Sun, Nathan E. TeGrenenhuys, Chi Wang, and Xiao Yu, "WinaCS: Construction and Analysis of Web-Based Computer Science Information Networks", Proc. of 2011 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'11), (system demo), Athens, Greece, June 2011.

[Yin2008] Xiaoxin Yin, Jiawei Han, and Philip S. Yu, "Truth discovery with multiple conflicting information providers on the web", IEEE Trans. on Knowl. Data Eng., 20(6):796--808, 2008.

[Yin2011] Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang, "Geographical Topic Discovery and Comparison", Proc. of 2011 Int. World Wide Web Conf. (WWW11), Hyderabad, India, Mar. 2011

[Yu2011] Xiao Yu, Ang Pan, Lu-An Tang, Zhenhui Li and Jiawei Han, "Geo-Friends Recommendation in GPS-based Cyber-Physical Social Network", Proc. of 2011 Int. Conf. on Advances in Social Network Analysis and Mining (ASONAM'11), Kaohsiung, Taiwan, July 2011.