Center for Visual & Decision Informatics

Novel Methods for Hidden Relation and Error Detection

PROJECT ID | 14-1 **TYPE |** [X] New [X] Continuing **START DATE |** July 2014

PROJECT LEAD/PARTICIPANTS | Yuan An, Tony Hu, Il-Yeol Song, Vijay V. Raghavan

DESCRIPTION | The objectives are to:

- Develop innovative methods and techniques for detecting hidden relations and inconsistent assertions in the structured information extracted from unstructured data sources.
- Develop application programming interfaces (API) with example graphical user interfaces (GUI).

EXPERIMENTAL PLAN | During the project period the team will work to

- Develop statistical inference methods such as probabilistic graphical models to reason about marginal and/or conditional probabilities for detecting possible hidden relations and errors. The models will capture syntactic and semantic dependencies among data items.
- Develop Conduct experiments to evaluate the effectiveness and efficiency of the proposed methods.

RELATED WORK | We have studied the problem of data error detection in electronic medical record systems. An article entitled "*Understanding the EMR Error Control Practices among Gynecologic Physicians*" was published in the iConference'2013. In addition, we have published a paper entitled "*An Error Detecting and Tagging Framework for Reducing Data Entry Errors in Electronic Medical Records (EMR) System*" in the IEEE Conference on Biomedicine and Bioinformatics 2013. In previous CVDI projects, we have developed prototypes for extracting semantic information from both structured and unstructured data. The outcomes of the previous projects provide solid foundation for developing novels methods on hidden relation and error detection.

HOW OURS IS DIFFERENT | We have investigated advanced probabilistic information extraction algorithms and developed the SemIntegrator tool for semantic information extraction. We also developed an innovative method for error detection in electronic medical record systems. The innovation of this proposed project lies in the new methods for capturing and reasoning about dependencies.

MILESTONES FOR YEAR |

3 months:

- Data sets collection and analysis. Building knowledge bases for data relation detection.

6 months:

- Develop probabilistic graphical models for capturing and reasoning about data dependencies.

12 months:

- Implement algorithms in prototypes.

DELIVERABLES |

1) Novel algorithms for inferring and reasoning about possible hidden relations and inconsistencies.

2) Implemented algorithms as prototypes.

BUDGET FOR YEAR |

Students	$57,272
Overhead	$5,728
Total	**$63,000**

ECONOMICS | The project helps industry members improve the process of using data for decision making.

POTENTIAL MEMBER COMPANY BENEFITS | The project provides techniques for extracting hidden knowledge from raw data for further analytics. It improves the productivity of data pre-processing.

PROGRESS TO DATE | We have investigated the problem of data error control and detection in electronic medical record systems. We also developed advanced information extraction and semantic discovery tools.

KNOWLEDGE TRANSFER TARGET DATE | 12 months

Center for Visual & Decision Informatics

Multi-Level and Multi-Source Visual Analytics of Evidence-Based Knowledge Diffusion Processes

PROJECT ID | 14-2 **TYPE |** [ ] New [X] Continuing **START DATE |** July 2014

PROJECT LEAD/PARTICIPANTS | Chaomei Chen, Prudence W. Dalrymple, Tony Hu, Erjia Yan (CCI, Drexel University); Leonard Samuels (College of Medicine, Drexel); Ryan Benton, Vijay V. Raghavan (UL Lafayette)

DESCRIPTION | The project aims to develop enabling and integrative techniques for multi-level and multi-source gap analytics across heterogeneous units of analysis. Building on the single-source gap analytics work in our CVDI Year 2 project, the project will support the entire workflow of gap analytics involving multiple sources. Prototypes will demonstrate the visual analytic workflow through key application cases such as translational and evidence-based medicine (esp. portfolios of clinical trials, medicine research, health informatics), user behavior in complex adaptive systems (e.g., interactive events initiated from users of a diversity level of experience, how do users' interactive behaviors differ between different levels?), and predictive studies of trigger events and trajectories of systemic changes. The objectives are to 1) develop computational solutions to visualize how information moves across multiple heterogeneous sources, 2) develop prototypes of showcases in relation to translational medicine, predictive analysis of the diffusion of information, and user adaptive behavior at different levels of expertise.

EXPERIMENTAL PLAN | The project will identify trajectories of organizations or other types of entities in various showcases with a focus on clinical trials, scientific literature, patents, and business performance. The project will extend the capability of visual analytics. The experiment will also apply the approach to the analysis of knowledge diffusion processes such as the intellectual fitness of organizations in terms of the competitiveness of their patent portfolio in a broader context and diffusion trajectories of their own innovations and their competitors' claims. Enhanced topic modeling and network-based predictive analysis techniques will be used for novelty detection.

RELATED WORK | Current approaches to gap analysis and predictive analysis commonly focus on predictions per se rather than giving explanatory information about underlying assumptions and critical evidence that leads to a particular prediction. It remains to be a challenge to convey how new information may alter the prediction holistically. The problem is in part due to the lack of information – not only on specific evidence per se, but how it fits in the entire chain of analytic reasoning and what alternative interpretation of the information might be. Translational science faces such challenges to an even greater degree because maintaining the provenance of evidence across the boundaries of distinct systems is much harder since distinct reference systems may be at play. Analysts need more effective support to identify the relevance of evidence and to synthesize uncertain and potentially conflicting information in a search, comparison, adoption, and other tasks

HOW OURS IS DIFFERENT | Our approach will extract the provenance of evidence from multiple heterogeneous sources of information. The provenance can be extracted algorithmically and synthesized from unstructured text without the need for an existing ontological structure. The approach is therefore more flexible and applicable to handle dynamic situations.

MILESTONES FOR YEAR |

3 months: Evidence extraction algorithms (medical, clinical, and adoption evidence)

6 months: Algorithms for modeling and visualizing trajectories.

12 months: Integrative prototypes and demonstrations of the workflow of visual analytics, with show cases in relation to translational medicine, competitive intelligence, and portfolio analysis.

DELIVERABLES |

3) Algorithms and implementations for modeling the structure and dynamics of knowledge transformation based on heterogeneous sources of data.

4) Prototypes for detecting and tracing trajectories of adoption in showcases related to translational medicine, adaptive user modeling, and other knowledge diffusion processes.

BUDGET FOR YEAR |

Students: DX: 1.25 FTEs	$59,390
Students: DX: hourly hire	$2,700
Supplies	$544
Equipment	$4,999
Travel	$3,000
Overhead (10%)	$7,090.84
Total	**$77,999.24**

ECONOMICS | Gap analyses and portfolio analyses are critical to business, industry, and the government to maintain a competitive edge. It is valuable for strategic management and tacit operations. The proposed project has the potential for both commercial exploitation and direct applications

POTENTIAL MEMBER COMPANY BENEFITS | The resultant prototypes would provide technical insights for the potential design of commercial products and fundamentally improve an organization's ability to assess a complex and evolving system.

PROGRESS TO DATE | We have developed algorithms and prototypes for portfolio analysis and gap analysis for grant proposals (NSF) and patent portfolios (NIH NCI). Industrial sponsors to our previous and current research include Pfizer and IMS Health.

KNOWLEDGE TRANSFER TARGET DATE | 12 months

VDI Center for Visual & Decision Informatics

A Predictive Analytics Framework for Spatiotemporal Hotspots

PROJECT ID | 14-3 **TYPE |** [X] New [ ] Continuing **START DATE |** July 2014

PROJECT LEAD/PARTICIPANTS | Jian Chen, Xiaohua Tony Hu, Ryan Benton, Raju Gottumukkala, Vijay Raghavan

DESCRIPTION | Data mining's future is predictive analytics. We will develop a predictive analytics framework for spatio-temporal data utilizing state-of-the-art big data technologies. In spatiotemporal data, people are interested in hotspots, areas of space and time with unusually high incidences of events. You may also want to predict how these hotspots may grow and where future hotspots may occur. To facilitate such analyzing and forecasting needs, this project has two specific objectives:

1) *developing hotspots detection tool applying spatiotemporal clustering techniques*
2) *empowering hotspots forecasting through evolutionary clustering, time series prediction and geospatial prediction*

Hotspots analysis and prediction is very useful in public health, homeland security, auditing selection, and business & marketing.

EXPERIMENTAL PLAN | During the project period, the team will work on:

1) *Data mining* Adapting spatiotemporal clustering techniques including those developed in our ongoing data mining project to find the hidden pattern and predictive information in the data, form hypothesis and detect interesting events.

2) *Predictive analytics* Investigating ways of integrating evolutionary clustering and classic prediction methods to predict hotspots evolution over time and space. Examples of such prediction methods are autoregressive integrated moving average (ARIMA), seasonal trend decomposition based on loess (STL), employing kernel density estimation, spatial scan statistics, and some machine learning methods.

3) *Case studies* Testing proposed framework in use cases of epidemic surveillance and tax auditing selection

RELATED WORK | We have a funded project to investigate spatiotemporal data mining approach for fraud detection using big data technologies. Experience gained and algorithms and tools developed can benefit the new project.

HOW OURS IS DIFFERENT | Most commercial predictive analytical tools developed by vendors such as SAS, SPSS, IBM, Insightful, etc. are short of spatiotemporal capabilities and more suitable for business and marketing. Little research has been done in area of spatio-temporal prediction using big data technologies. Our approach is based on evolutionary clustering algorithms and machine learning methods on top of Hadoop platform, provides powerful spatiotemporal analysis & prediction functions, which is the key for decision makers.

MILESTONES FOR YEAR |

Q1: Develop hotspots detection algorithm
Q2: Develop hotspots prediction algorithm
Q3: Develop prototypical predictive analytics framework for hotspots detection & prediction
Q4: Conduct case studies

DELIVERABLES |

5) Algorithms for spatiotemporal hotspots detection and prediction
6) A prototypical predictive analytics framework for analyzing and forecasting spatiotemporal hotspots

BUDGET FOR YEAR |

Students	$ 50,000
Supplies	$ 3,000
Equipment	$ 2,870
Travel & Training	$ 7,500
Overhead (10% of everything)	$ 6,337
Total	**$ 69,707**

ECONOMICS | A predictive analytics framework integrates the analytic power of data mining, and prediction. It is the ideal decision making tool in big data era: turning big data to big impact via high level knowledge, proactively not reactively.

POTENTIAL MEMBER COMPANY BENEFITS | Public health (disease outbreaks), marketing (emerging markets, cross-selling, advertisement optimization), securities (terrorism, crimes, cyber security), tax (auditing selection)

PROGRESS TO DATE | We have discussed use cases of epidemic surveillance and tax auditing selection with interested IAB members. Several de-identified contagious diseases surveillance datasets have been collected and preliminarily analyzed.

KNOWLEDGE TRANSFER TARGET DATE | 12 months

VDI Center for Visual & Decision Informatics

Analyzing, Modelling and Summarizing Social Media and Linked Datasets

PROJECT ID | 14-4 **TYPE |** [X] New [ ] Continuing **START DATE |** July 2014

PROJECT LEAD/PARTICIPANTS | Tony Hu, Ryan G. Benton, Yuan An, Vijay Raghavan, Chaojiang Wu, Erjia Yan

DESCRIPTION | The objectives are to:

- Utilize data from multiple source systems for analysis purposes, by integrating both social media and linked data. This integration empowers internal business users with the capability analyzing diverse information and data from multi-dimension and multi-view perspectives at different aggregation and granular levels.
- Develop approaches that can specifically address the needs of perceiving and analyzing the heterogeneity of linked data.
- Develop methods to summarize the information extracted from the linked data, in a coherent, consistent fashion.

EXPERIMENTAL PLAN | During the project period, the team will work to:

- Develop data integration tools to clean, extract, and pre-process multiple heterogeneous data sources at different granular levels such that data will include internal data, social media data, news data and so forth.
- Develop algorithms and methods to analyze, mine, model, and summarize large and diverse heterogeneous soclai media and linked data sets by integrating business domain knowledge.
- Work with IAB members to evaluate the effectiveness and efficiency of the prototype systems in the real-world setting and transfer the prototype systems to the IAB members.

RELATED WORK | We have done extensive work in mining and modeling large and diverse data sets in real-world applications in financial institute, e-business, healthcare, customer service and management, etc. We have demonstrated the ability to merge multiple social media sources for applications such as emerging event detection well as extracting meta-data from such sources. Our team also has experience in automated hypothesis discovery for bio-medical literature, link prediction, and automated ontology generation from Wikipedia and graph analysis.

HOW OURS IS DIFFERENT | The current proposal is motivated to resolve two hindering limits: (1) utilize external data sources (third party, social media, Web), not captured by the enterprise, (2) develop a unified framework to develop algorithms/methods/model for predictive analysis and summarization of results. Moreover, by fusing older static data (past user surveys) with more dynamic sources (social media, news), there is the ability to have grounded, known data tempered with up-to-date information.

MILESTONES FOR YEAR |

*6 months:*Develop dynamic schema and wrapper to integrate multiple heterogeneous data sources including internal data as well as external data in different granular level

9 months: Develop summarization techniques for classes of data/discoveries.

12 months: Design algorithms for clustering high-profit margin customer, modeling top candidate for customer marketing, cross-selling, up-selling, identify potential churners/attriters

DELIVERABLES |

1. Data schema and wrapper for multiple large and diverse data sources (both social media and linked external data sources).
2. A set of algorithms for analyzing, mining, modeling and prediction of large and diverse data sets.
3. Methods to summarize classes of data and discoveries.

BUDGET FOR YEAR |

Total **$48,000**

ECONOMICS | Mining to Analyze, Model and Make Prediction using social media and linked data sets is essential for almost all industry sectors to improve the competitiveness in the global economy in the Big Data age. Moreover, the ability to provide coherent summaries for certain types of data and analysis will increase the ability of decision makers to make quick, decisive decisions.

POTENTIAL MEMBER COMPANY BENEFITS | This project will benefit IAB members by providing a set of novel approaches and methods to gain advantages in retaining high-profit customers, acquiring new customers, increasing business revenue, increase sales with cross-selling and up-selling, enhance customer loyalty and targeted marketing through personalized service.

PROGRESS TO DATE | We have developed several algorithms/methods in data integration, data pre-processing, and mining. Research prototypes from our previous research work, such as Dragon Toolkit and MapOnto could be adapted for use in this project. Graph visualization techniques and phrase/concept extraction techniques already exist. We have demonstrated our ability to fuse data from multiple social media sites as well as extracting meta-data from social media, which can aid in the linking process.

KNOWLEDGE TRANSFER TARGET DATE | 12 months

Center for Visual & Decision Informatics

Visual Analytic Methods for Dynamic Graphs

PROJECT ID | 14-5 **TYPE |** [X] New [ ] Continuing **START DATE |** July 2014

PROJECT LEAD/PARTICIPANTS | Raju Gottumukkala, Christoph Borst, Chaomein Chen

DESCRIPTION | This project focuses on interactive analysis of complex graphs. Organizations increasingly face a need to understand phenomena from real-world, real-time data sources such as social media, health or financial records, sensor or click streams. Example applications include environmental sensing, patient health and disease monitoring, financial and marketing decisions, etc. Relational data may be represented as time-varying graphs, where nodes represent entities and the edges represent inter-node interactivity or relationships. Both graph structure and attributes of nodes and edges may change as new information arrives or in response to interactive exploration (e.g., filtering or changing edge weight scheme). Visual analysis of the graphs is a challenging problem, especially as they become large. Managers need to integrate multiple graphs, e.g., community networks from social media, consequence analysis networks, and transportation networks. Interactive analytics and visual exploration techniques are key to understanding complex graph structure and behavior. Relevant features need to be extracted and aggregated dynamically based on user input. The primary project goal is to develop integrated high-performance visual analytic techniques that combine graph analysis with visualization and interface techniques for interactive mining of multi-modal, multi-relational graph mining methods.

EXPERIMENTAL PLAN | During the project period, the team will work on the following tasks:

1. Develop high-performance data mining techniques to support interactive analysis of large dynamic graph-based datasets
2. Extend and develop visual data browser : a) visually represent more complex graph data, and b) develop innovative interactive elements (multi-touch display techniques) for controlling data views, filtering, prediction, etc.

RELATED WORK | There has been a lot of work in integrating visualization, graph mining on relatively static graphs or data, or exploration using basic mouse-type techniques. We specifically focus on exploration of more dynamic structures. There are some existing multi-touch interaction approaches for more basic data plots (e.g., Microsoft's TouchVis for charts on tablets, and TouchWave for stacked plots). New datasets and applications will guide innovative techniques.

HOW OURS IS DIFFERENT | An equivalent integrated framework for visual analytics of large dynamic graphs does not exist. We propose integrated real-time analysis and a multi-touch exploration interface for large dynamic graph models. We seek novel techniques for interactively querying graphs. We will take advantage of multi-touch displays for improved interaction (e.g., two-handed selection or menus, multi-user interaction) instead of mimicking mouse interfaces.

MILESTONES FOR YEAR |

3 months: Obtain & prepare data sets for the project

6 months: Data modeling to derive new relationships from existing data

12 months: Performance evaluation, visualization & demonstration.

DELIVERABLES |

1. Framework for real-time analysis and visual exploration of graphs
2. Visual analysis and mining algorithms.
3. Demo visual data browser with multi-touch browsing.
4. Practical knowledge about implementation, feasibility, and tradeoffs of approaches.

BUDGET FOR YEAR |

Students (3 UL)	$ 58,025
Supplies	$ 4,000
Equipment (1 workstation, Display)	$ 5,001
Travel (IAB meeting, conferences)	$ 14,792
Overhead (10% of everything)	$ 8,182
Total	**$ 90,000**

ECONOMICS | Government agencies need to understand public perception of various governments' efforts, particularly in emergency management and public health, where government can benefit by taking proactive steps to improve outreach strategies to targeted areas. Real-time BI solutions with high-end visual analytics capabilities bring insights to organizations based on most recent data rather than days old data. Multi-touch displays are becoming a standard available interface type, supporting newer interfaces on affordable displays.

POTENTIAL MEMBER COMPANY BENEFITS | Faster insight into emerging events provides companies more time to react to events, which would the influence the outcome via mitigation and/or proactive management. New interaction techniques potentially improve usability and success of visualization systems. Multi-touch displays support additional interaction types.

PROGRESS TO DATE | The project infrastructure includes hardware, software and distributed processing infrastructure from the previous project, and the data and event detection techniques for other CVDI projects will be leveraged. We have tested preliminary multi-touch display compatibility in our sensor data browser from Years 1-2, allowing basic sensor/plot selection, and demonstrating how our software could be adapted to work in both 2D and 3D environments.

KNOWLEDGE TRANSFER TARGET DATE | 12 months