OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE

Knowledge Discovery and Dissemination (KDD)

LEADING INTELLIGENCE INTEGRATION

Dr. Arthur Becker
Program Manager
Arthur.becker@iarpa.gov

KDD Objectives

Objective

Enable analysts to **quickly** produce actionable intelligence from multiple, disparate data sources, including new unanticipated data sets available to analysts.

Program Thrusts / Contractor Tasks

Research in Data Alignment to quickly align the terminology and organization of new data sources to the analytic data model.

Research in Analytics to develop flexible algorithms that work across heterogeneous data sets.

Engineer a prototype that contains the research products so that they can be assessed in a realistic IC environment.

To meet flexibility objectives, Performer's research is used to create software components.

Components are loosely coupled in their prototype and combined into workflows to perform analytic tasks when needed.

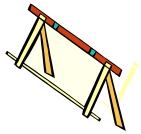
What We Aren't Doing

-- Scalability research
-- User interface research and user studies
-- Research to process media such as speech or video
-- Research to process foreign language data
-- Research to develop information processing architectures
-- Research focused on IC policy and cultural issues

Evaluating Research Components

Research Idea

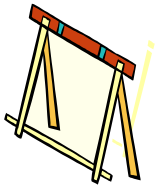
Innovative research idea for an alignment or analysis component. (Stage 1)



Standard Benchmark

For most research components there are accepted metrics and test data sets that are used to measure and compare performance. (Stage 2)

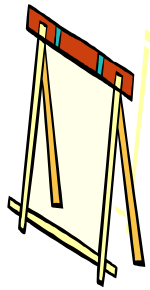
How does the research compare to state of the art performance?



Intelligence Benchmark

The component is retested on IC data using the same metric. Speed and scalability issues are addressed. (Stage 3)

Can it work similarly against real data?



Operational Utility Test

The component is integrated into a prototype system and tested by IC analysts using IC data to do specific tasks and answer challenge problem. Results are compared to a baseline system. (Stage 4)

Does it add value to analysis in realistic situations ?

KDD Research- Analytic Components

1) A graph based query method that uses the structure of the query rather than the keywords. Method had 99% accuracy querying DBPedia and 92% accuracy on field reports. Method is extremely fast in most situations. (Stage 4)

2) An algorithm that quickly extracts facts from large volumes of text and stores them in a simple searchable format. Analysts found this useful for getting a sense of what is in the data about a particular name, location, event or term; or to explore a new name or term. (Stage 4)

3) A new method to retrain statistical algorithms to extract entities and events from text involving a convenient review of system-nominated sentences. Reduced the number of retraining samples by over 95% with minimal loss of performance. (Stage 4)

4) A novel approach to adapt natural language processing to new problem domains reduced time by 95% (40 Hrs. to 2 Hrs.) with no decrease in performance. The approach combined a templating approach with semantic calculus. (Stage 4)

5) A new method for entity resolution on a large data set achieves 75% reduction in the error rate (8% to 2%) over current state-of-the-art methods. Method is also more amenable to streaming data and large data volumes. (Stage 2-3)

KDD Research: Analytic Components (continued)

6) A dynamic semantic graph capability where sufficiently detailed graphs can be computed on the fly (5 minutes) using NLP, structured alignment and entity extraction. This serves as an alternative to creating a semantic data warehouse. (Stage 3)

7) A new technique for network completion that has better performance than previous state-of-the-art techniques. The technique finds missing edges and nodes of semantic networks as well as social networks. Limited testing shows 80% accuracy on public data and similar results on IC data sets. (Stage 3)

8) A new technique for creating a summary of multiple documents. The method generates text from a discourse graph that captures logical relations between sentences that are important for readability. The result is a summary that is more readable and more amenable to multiple sources . (Stage2-3)

9) A new technique for event resolution that uses time, location, and event characteristics to assign similarity scores between events. Mechanism to translate locations into bounding boxes harmonizes vocabulary differences between report types. Thresholds for similarity criteria fully exposed and customizable by users. (Stage 4)

KDD Research
Alignment and Infrastructure Components

10) A semi-automated tool to align structured data sets reduces time by 90% on academic data and 80% on IC data sets. (Stage 4)

11) A Hierarchical realist ontology that can be quickly adapted to new domains and new data sets. Has been successfully applied in 4 KDD challenge problems and many transition partner data models. (Stage 4)

12) A quality of service workflow module that provides analysts with the ability to specify the amount of resources and time they are willing to allocate for a particular action. Tested and used in MapReduce environment and being moved to Storm. (Stage 4)

13) Provenance Capture Software: Automatically captures and stores analysts highlights as a citation URL describing source, record and offset. (Stage 4)

KDD Research
Evaluation and Methodology

14) Analytic Test Range: A stand-alone platform that isolates data services, analytic algorithms and user interfaces. The test range has been used for pre-transition evaluation of technology and is being considered as a training vehicle. (Stage 4)

15) Stand-alone Geospatial Visualizers: Stand-alone map server/visualizer using three open source products. (Stage 4)

16) Analytic Test Scenarios: Through the KDD evaluation a methodology for creating evaluation scenarios and data sets has been created and refined. (Stage 4)

KDD Performer Teams

Systems and Severable Components

- Use of a hierarchical model to rapidly build a problem ontology and adapting analyst tools to it.
- Use of Hadoop/Storm and HBase for parallelism and Oozie for workflow management.
- NLP accuracy.

CUBRC

Actionable Intelligence Retrieval System (AIRS)

SRI International Project Dovetail

- Rapid, trainable assimilation of new information sources, structured and unstructured.
- Comprehensive integration of two best-of-breed information extraction systems
- Automated construction of entity dossiers.

Partners and Transition Process

KDD addresses:

- ✓ How does the component work against state-of-the art tools?
- ✓ How does it scale and work with IC data?
- ✓ How does it work when analyst use it to solve real problems using real data?

To move to TRL 6 transition, partners want to know:

How does it work against my data?

How does it work against my problems?

KDD has at least Government Purpose Rights for nearly all software delivered.