Finding More Effective Algorithms for Connecting the Dots Between Entities in Intelligence Analysis

Introduction

Background
- Intelligence analysts draw connections between entities (suspects, organizations, etc.) using data.
- Connections help analysts construct more effective investigations.
- Large amounts of data overwhelm the capabilities of human analysts.
- Computer software (Entify Workspace, Jigsaw, etc.) is needed to help find meaningful connections from data.
- Current algorithms have flaws: runtime, lack of domain knowledge, little support for explanations of connections, etc. (Hossain, Butler, Boedihardjo, & Ramakrishnan, 2012).
- By improving algorithms, we can assist national security by providing agencies with more coherent, reasonable leads.

Purpose
- Design an algorithm which connects the dots between entities, alleviating the issues of current intelligence algorithms.

Analytical Considerations
Existing algorithms have numerous limitations:
- Lack of support for “evidence marshalling”
- Lack of support for explanations of stories
- Little support for directed searches and manual exploration
- No support for syntactic constraints (people, places, etc.)
- Lack of support for entity extraction/disambiguation

Algorithm Design Goals
- We will examine documents from the VAST_2010 dataset, composed of phone calls, emails, and field reports.
- This allows us to use our algorithm to examine a variety of different documents.
- We will only use specific data entities as juncture nodes (syntactic constraints).
- We will use Natural Language Processing (NLP) software to extract entities from documents.
- We also aim to extract relevant information from the text besides entities, for example, mentions of weapons, money, etc.
- We will also attempt to identify entities composed of multiple words that are not repeated in the story.
- Our algorithm will create graphs of text documents (vertices and edges)
- Provides an intelligible story which will help human analysts identify new connections that might have been previously hidden

Advantages of Design
- Removes need for resource-intensive “evidence marshalling”
- By using natural language processing software, our design is more likely to provide meaningful stories.
- By using the most well-researched software libraries, our process has the best chance of extracting entities and finding their role in the story.
- By representing the entities as vertices in a graph connected by junctures, directed searches for relationships are easier.
- Manual exploration is possible with this design.
- Many of the essential elements of the documents are parsed by the software.
- By being selective with our parser, we can determine what parts of the sentences are of interest (people, places, organizations, etc.).
- Far easier to implement syntactic constraints.
- Implementing a graph-based approach also allows us to see connections between numerous entities (identifying cliques).

Design
From a dataset, we select which documents we would like to examine.
Once we have the individual documents, we break them apart into sentences using a rule-based engine.
We input the individual sentences into a parser (Manning, 2003) and part of speech tagger (Thomas, Klein, Manning, & Singer, 2003).

The tagger identifies entities and words of interest. The parser identifies the relationships between words (de Melo & Manning, 2003).

We combine words which constitute single entities with the information from the parser and tagger.

Design (Continued)
We output the sentence as a graph with the vertices being the extracted entities and words of interest.
Connections between words are drawn based upon their relationships in the original sentence.

Future Work
- What is the most effective way to combine the graphs of the sentences (Hossain, Aklar, & Pires, 2012)?
- What is the best way to display the output graphs?
- Find the most effective representation for human readability
- Improved sentence extraction, requires more natural sentence model
- Possible affects of clustering on data organization
- Development of search algorithm within master graph
- Calculations using “weight” or “distance” heuristics
- Incorporation of k-cliques or fuzzy granules in search for best stories (Jalal-Kamali, Hossain, & Kreinovich, 2014).
- Use more advanced Natural Language Processing techniques to extract better graphs from data
  - Use NLP to label the edges of the graph, and learn how to best use these labels when processing data.
- Human analysts still have a cognitive advantage.
- Learn how to elicit and use expert knowledge when processing data.
- Humans sometimes think in a disorganized manner which can improve storytelling (Bracken, Self, Endert, & Hossain, 2013).
- Can we emulate this using computers?

References

We would like to thank the following individuals for their contributions:
Nicholas Sun: nickolas.sun@rutgers.edu; Sheikh Molaher Naim: sheikh.naim@miners.utep.edu; Dr. Mahmud Shahriar Hossain: mhossein@utep.edu; Dr. Vladik Kreinovich: vladik@utep.edu

Dr. Mahmud Shahriar Hossain
University of Texas at El Paso
mhossein@utep.edu

Dr. Vladik Kreinovich
University of Texas at El Paso
vladik@utep.edu