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

Nicholas Sun
Rutgers – New Brunswick
nicholas.sun@rutgers.edu

Sheikh Motahar Naim
University of Texas at El Paso
snaim@miners.utep.edu

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

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

Introduction

Background

- Intelligence analysts draw connections between entities (suspects, organizations, etc.) using data.
- Connections help analysts construct conduct more effective investigations.
- Large amounts of data overwhelm the capabilities of human analysts.
- Computer software (Entity 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

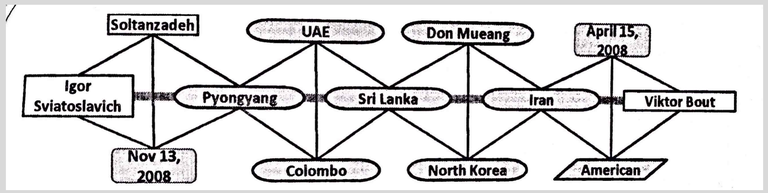
Existing algorithms have numerous limitations:

- 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

Example output from Jigsaw Visual Analytics
http://www.cc.gatech.edu/gvu/ii/jigsaw/image/views/Graph.jpg

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 so they 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



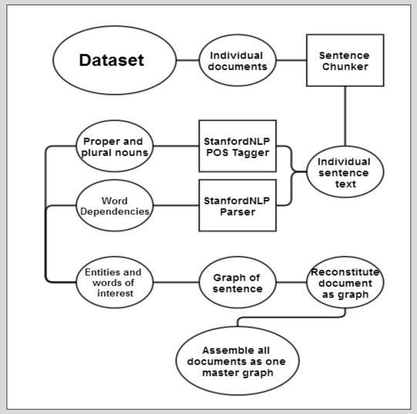
An example of a graph-based story in VAST_2010 with syntactic constraints

Hossain, M., Butler, P., Boedihardjo, A.P., Ramakrishnan, N. (2012). Storytelling in Entity Networks to Support Intelligence Analysis. KDD '12, August 12-16, 2012, Beijing, China. (Fig. 13)

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



"Last month, a unit of Saudi National Guards known as the Mujahidin issued a statement saying it had seized a large shipment of weapons – which included hand grenades, explosives, guns and ammunition – in Dafa valley when it intercepted arms smugglers led by a notorious arms merchant Saleh Ahmed, who fled back to Yemen."

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.

Saudi_NNP
National_NNP
Guards_NNP
Mujahidin_NNP
weapons_NNS
grenades_NNS
explosives_NNS
guns_NNS
Dafa_NNP
arms_NNS
smugglers_NNS
arms_NNS
Saleh_NNP
Ahmed_NNP
Yemen_NNP

[amod(month-2, Last-1), root(ROOT-0, month-2), det(unit-5, a-4), appos(month-2, unit-5), nn(Guards-9, Saudi-7), nn(Guards-9, National-8), prep_of(unit-5, Guards-9), vmod(Guards-9, known-10), mark(issued-14, as-11), det(Mujahidin-13, the-12), nsubj(issued-14, Mujahidin-13),…

We input the individual sentences into a parser (Klein, Manning, 2003) and part of speech tagger (Toutanova, Klein, Manning, & Singer, 2003).

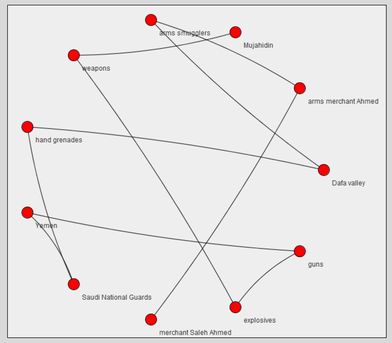
The tagger identifies entities and words of interest. The parser identifies the relationships between words (de Marneffe & Manning, 2008).

Saudi_NNP
National_NNP
Guards_NNP
→ Saudi National Guards

Dafa_NNP
valley_NN
→ Dafa valley

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.

Finally, the graphs of the sentences are combined using the entities as juncture points. The resulting "master graph" will display connections of the entities and can be used by analysts to find new leads and stories.

Future Work

- What is the most effective way to combine the graphs of the sentences (Hossain, Akbar, & Polys, 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 (Bradel, Self, Endert, & Hossain, 2013).
  - Can we emulate this using computers?

References

Bradel, L., Self J.Z., Endert, A., Hossain, M.S., North, C. Ramakrishnan, N., (2013). How Analysts Cognitively "Connect the Dots". Department of Computer Science, Virginia Tech, Blacksburg, VA.

de Marneffe, M., Manning, C., (2008). Stanford Dependencies Manual.

Hossain, M.S., Akbar, M., Polys, N.F., (2012). Narratives in the Network: Interactive Methods for Mining Cell Signaling Networks. Journal of Computational Biology, 19 (9), 1043-1059. DOI: 10,1089/cmb.2011.0244

Hossain, M.S., Butler, P., Boedihardjo, A.P., Ramakrishnan, N. (2012). Storytelling in Entity Networks to Support Intelligence Analysis. KDD '12, August 12-16, 2012, Beijing, China.

Jalal-Kamali, A., Hossain, M.S., Kreinovich, V., (2014). How to Understand Connections Based on Big Data: From Cliques to Flexible Granules. Department of Computer Science, University of Texas at El Paso, El Paso, TX.

Klein, D., Manning, C., (2003). Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430

Toutanova, K., Klein, D., Manning, C., Singer, Y., (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *In Proceedings of HLT-NAACL 2003*, pp. 252-259.

Toutanova, K., Klein, D., Manning, C., Singer, Y., (2014). Basic English Stanford Tagger (Version 3.4) [Software]. Available from http://nlp.stanford.edu/software/tagger.shtml

Please direct all correspondence to:
Nicholas Sun: nicholas.sun@rutgers.edu; Sheikh Motahar Naim: snaim@miners.utep.edu

Funded by U.S. Department of Homeland U.S. Department of Homeland Security's Science and Technology Directorate Office of University Programs under Grant Award Number 2008-ST-061-BS0001-05Science and Technology Directorate Office