Custom Ontologies for Expanded Network Analysis

Amy K. C. S. Vanderbilt, Ph.D. and George Strauss
Wave Technologies, Inc.
4465 Brookfield Corporate Drive, Suite 200A, Chantilly VA 20151, USA

avanderbilt@wvtec.com, gstrauss@wvtec.com

ABSTRACT

This paper discusses a new approach to answering Requests for Information (RFIs) from military commanders, intelligence analysts, individual soldiers and others received by reach-back information and intelligence collection repositories. This new approach avoids the previous ideals of either searching out a set of a thousand documents or building one large all-encompassing ontology and instead embraces the concept of custom ontologies based on each users query and returns to that user a concise and organized knowledge set along with visualizations that invite exploration and facilitate assimilation.

1.0 INTRODUCTION

When a software user requests information on one topic or another, they have only two choices to answer the need: document retrieval, and large scale living ontologies. These methods represent the best of unstructured and structured information, but come each with their own problems. This dichotomy led to the development of a third option that takes in the best of both: Custom User Ontologies. This method harvests the requested information from large amounts of unstructured text, and deposits it into an ontological structure that the user may explore for greater understanding. What follows is brief description of the two current methods for RFI response, details on the concept of custom user ontologies, and an ongoing application of this new method. Visualization of custom ontologies is a particular challenge. These ontologies are compound network structures that may contain diverse information types and will therefore require diverse concurrent visualizations. Potential methods for visualizing custom ontologies are discussed near the end of this paper.. In conclusion we will discuss the potential relevance and impact, and forge a vision for the future of this new method.

2.0 PREVIOUS APPROACHES TO RFI RESPONSE AND KNOWLEDGE DISSEMINATION

Previous approaches to RFI response and knowledge dissemination have fallen into two categories: text search / document retrieval and large scale living ontologies. We will consider the motivations, and eventual detriments of each.

2.1 Text Search and Document Retrieval

When searching a large corpus of free text documents for just the right information, the first thing many try is to do a web-type search using a search engine such as Google, Alta Vista, or other such document search and retrieval algorithm. These search engines are queried with a short word string that is terse by necessity and

Vanderbilt, A.K.C.S.; Strauss, G. (2006) Custom Ontologies for Expanded Network Analysis. In *Visualising Network Information* (pp. 6-1 – 6-10). Meeting Proceedings RTO-MP-IST-063, Paper 6. Neuilly-sur-Seine, France: RTO. Available from: http://www.rto.nato.int/abstracts.asp.

therefore fairly general in concept. The simplest of these searches are based on scoring functions and return an ordered stack of documents that have a certain percent match to the query.

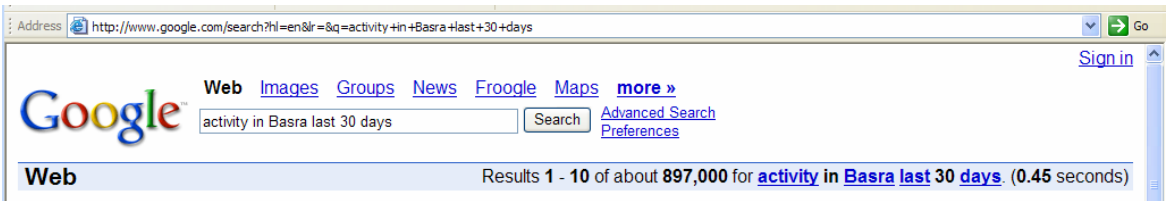

Figure 1: Google.com search results showing a return of 897,000 entries

The result is a set of one hundred to ten thousand or more documents that match the query to some degree and may or may not contain the specific type of information the user seeks. This satisfies the need to believe that we have taken in all available information on a topic, but forces the user to spend time paging through each document in the hopes of finding the dozen or so kernels of knowledge, and associated relationships that they seek. Essentially, the user is performing information extraction in their heads on every document that may contain the targeted information. The time-consuming nature of this process has led to another method at the other end of the spectrum.

2.2 Large Scale Living Ontologies

Instead of extracting information document by document in the users head, an alternate approach is to pre-extract all possible information from all possible documents related to a subject or set of subjects and form a large-scale ontology to house the information and relationships. This approach was exemplified in the Cyc research project that began in 1984. This project and others continue to evolve into ever larger editions of what we will term large-scale living ontologies [1]. Even to this day, current research groups are working to develop an approach that involves taking all available text, video, imagery and audio and extracting from it instances of known relations to populate a pre-structured ontology. Users then query the ontology with a terse and highly structured relation or topic and get back a series of relational statements and instance lists to consider.

These ontologies seek to encompass all possible topics, or all possible topics within a domain. This is, at least initially, an enticing method. It embodies the general concept of taking in vast amounts of unstructured data and converting it into structured data. Structured data, the user believes, will be so much more easily searched, and analyzed with a multitude of tools. This much is true in the general sense, but problems present themselves over the life of the ontology.

Figure 2: Map of high-level CYC Topics (www.cyc.com)

Although this approach returns more precise information, the problem with this approach is that the ontology must grow and be constantly modified to allow for new concepts and to update old ones. This leads to a number of problems.

- The ontology is soon so large that it takes almost as long to query the ontology as it would have taken to query the entire set of original documents and resources, and read the results. In figure 2 above we see a mind map of the CYC upper ontology. This diagram gives an idea of the scale to which the ontology is growing.
- The information within the ontology is so diverse that only a small portion of it is pertinent to the user and the large majority of it is useless to everyone.
- At the same time, if a concept does not exist in the ontology, it must be added in by an ontological engineer. That means that if a topic you are interested in is not in the ontology already, you will not get a return on your query, and will be forced to use a document search and retrieval engine just as before.
- To update the ontology, people must be continually engineering modifications and additions and the original resources must be constantly re-queried to populate those additions.
- Without the humans modifying the ever-growing ontology, it will become obsolete quickly.
- Each human ontological engineer may modify and grow the ontology differently using their own take on a topic and their own natural language, guarded only by basic ontological engineering semantic standards. This can lead to a mismatch between how the user would describe what he is looking for and how the ontology is housing that information. Great strides have been made in the last decade toward disambiguation of queries and concepts, but it has not yet been solved.
- The user may become disgruntled with the user-hostile environment of the ontological behemoth and

cease to use it. This serves no one.

An alternate Methodology is called for; one that settles appropriately between the two extremes of document retrieval and large-scale living ontologies; NOT a combination of these methods, but rather a new method entirely.

3.0 CUSTOM USER ONTOLOGIES

When considering the requirement to answer requests for information via analysis of vast amounts of unstructured information in multiple formats, the ideal application would be a system of ultimate efficiency; a system that would enable vast amounts of unstructured text to be harvested for only the concise kernels of information required for any one particular query. It was with this inspiration that the concept of custom user ontologies was formed.

3.1 The Concept of Custom User Ontologies

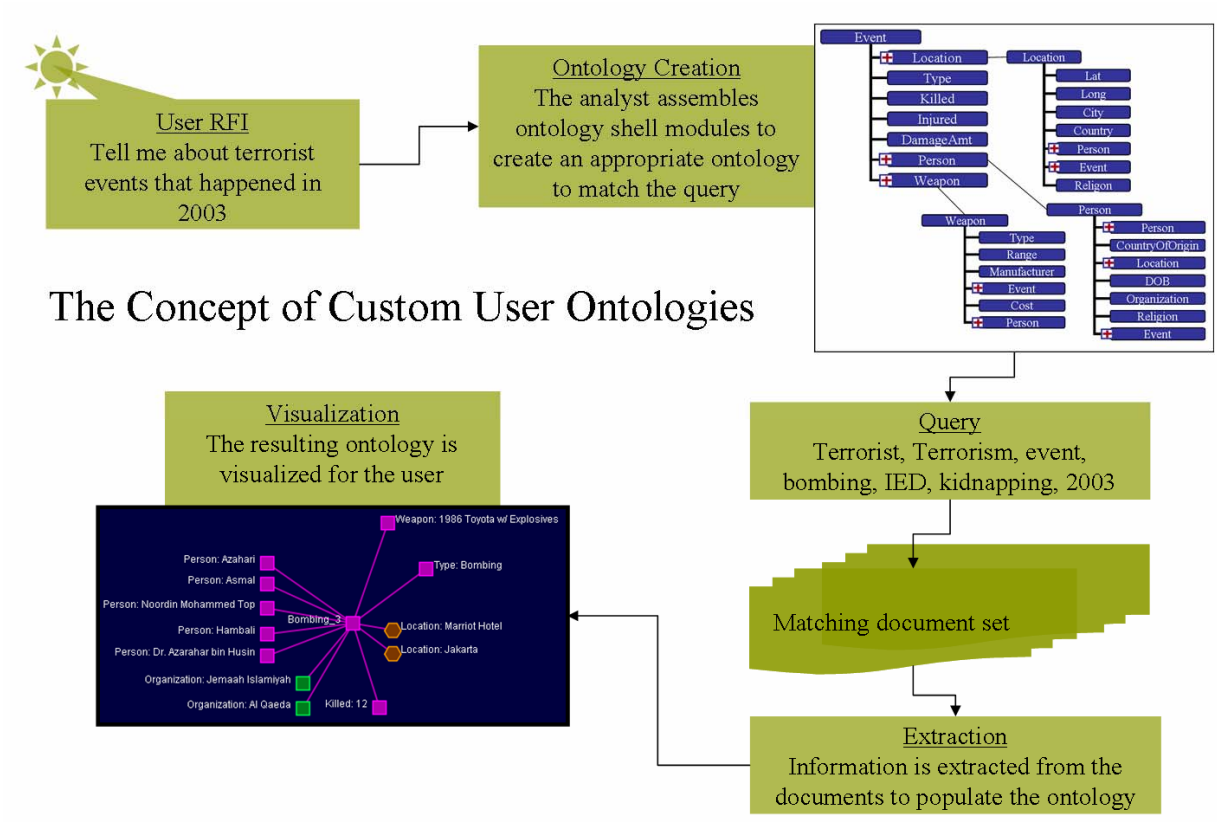

Figure 3: The Concept of Custom User Ontologies

Currently, one is forced to choose which method will be the basis of a given technical approach. Our own dissatisfaction with both extremes currently offered led to our development of a middle ground. Building on the idea of information fusion for common operational understanding [2], the concept of custom user ontologies was derived and detailed. At the cornerstone of the concept is the idea of creating a populated ontology complete with entities, relationships and appropriate visualizations for each user and each query. The

initial absurdity of the idea may seem to be rooted in the perceived onslaught of ontologies that would result, or the potentially long wait time required in searching a large corpus for the specific entities and relations. While it is true that many ontologies will result, each one is small. For any one user, a single query is looking for relatively little information: events in a certain location over the last month, the social network of a person in a photograph, new techniques for new tactical situations, and so on. The user craves not the drink from an informational fire hose yielded by popular methods to date, but a small bowl of concise knowledge that answers their curiosity of the moment. The issue of computability time can be solved in a number of ways from converting documents to text only, removing filler words, breaking the unstructured text into smaller blocks, and performing other pre-processing tasks. Figure 3 above details the concept of custom user ontologies as a process by which requests for information are filled by returning concise detail pulled from thousands of resources.

3.2 A Suggested Application of Custom User Ontologies

The concept of custom ontologies can form the basis for a unique search capability which could be housed at analysis centers, web portals, and other locations. The challenge is to provide distilled information, not by a snapshot in time, but as a continual understanding by the user of a given environment, the players in it and how it is changing.

At the heart of this effort are the data sources from which content-rich and relevant information is extracted and delivered to the user. Critical data sources might include archives and periodicals such as the Center for Army Lessons Learned (CALL), Janes, Lexis Nexis, Factiva, and many others. There is no shortage of available information to search and distil; quite the opposite, in fact. However, the user must be careful to choose sources that are trustworthy and reliable; the definition of which will vary greatly with the user.

Given portal access to such sources of structured and unstructured data and information, threaded spiders that crawl the portal in parallel can harvest updated information on a regular basis. These updates would then be tagged by source, date and other relevant metadata, and archived within an architecture that allows easy retrieval.

As seen in figure 4, critical content would be turned into concise actionable intelligence, through advanced algorithms, modelling and simulation, for everyday users and seasoned analysts alike that supports their need for information and understanding.

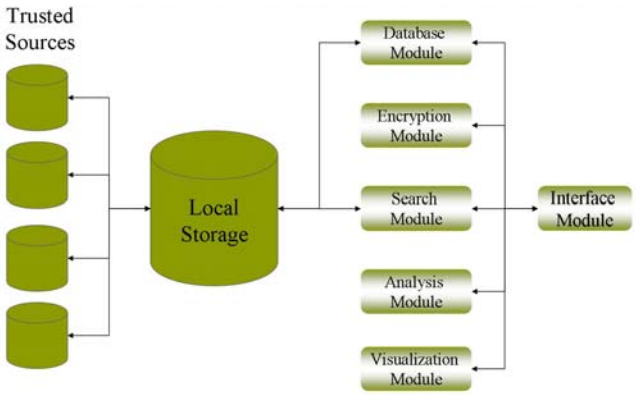

Figure 4: A Potential Concept of Operations

The process might look like the following. Users submit a Request for Information (RFI) to analysts by phone or email or via a website. Users submit requests in their own words, using their own natural language.

For example:

- I need to know everything that has happened in Town X in the last 90 days
- I need to find out what this weapon is in the attached picture
- Are there any new techniques for clearing buildings of type X?
- What do you know about person A?

The analyst then enters the query into the system, in a somewhat more detailed and concise form, aids the software as it builds the custom ontology and initiates information extraction based on that ontology. Such an application could be built in a modular way from largely existing components. For example, Carnegie Melon University's AutoMap and Ora software tools [3,4] could be integrated with organic algorithmic components to perform the search space reduction, ontology construction and information extraction necessary to make custom user ontologies a reality.

4.0 VISUALIZING CUSTOM USER ONTOLOGIES

All of these ontologies are of little use if they cannot be visualized so that the user may internalize and understand them. The challenge in visualizing custom ontologies is in the variety of data types that may exist within any one ontology, and the dynamic nature of subsequent updates to that ontology over time. The ontology itself can be generalized as a network whose nodes and links are made up of various types. For example, each custom ontology can be generalized as a fusion of multiple networks, such as social, geographic, financial, computer-based, logistical, and others.

4.1 Network Diversity and Possible Visualizations

Within each type of network, such as those listed above, there are various types of nodes or entities connected by various types of links or relationships. Each of these nodes and links will have its own set of metrics dependent upon the users needs. In the table below we put forth just a sample of possible network types, what nodes and links might represent within each, and metrics whose values may be of interest to the user.

Network Type	Example Nodes (Entities)	Example Links (Relationships)	Example Metrics (for nodes or links)
Social	People's names Organizations	Social relationships Political relationships Diseases passed	Centrality Influence Relationship strength
Geographical	Cities Terrain features	Transportation routes	Population Transport capacity
Financial	Account numbers Banking institutions	Transfers	Amounts Dates Times

Cyber	IP addresses	LAN / WLAN connection	Packets sent/received Bandwidth
Logistical	Geographical locations People's names	Weapons transferred Supplies transported Other goods transferred	Amount Frequency Dates Times Type of goods

Table 1: A Sampling of Network Types with Example Nodes, Links and Metrics

Given these network types, and their related metrics, appropriate visualizations of the ontology and its contents may consist of a combination of the following (some of which are seen in Figure 5):

- GIS or 3D landscapes,
- Timelines,
- Bar, line or other graph types,
- Timelines blended with GIS,
- Graphs overlaid on landscapes,
- Architectural views of the ontology itself which depict entities as nodes and relationships between entities as links in the network-like structure

With such a range of information to present to the user, the question becomes how to do so without overcrowding the scene. Couple the overcrowding problem with the fact that each user will work more efficiently with their own preferred type of visualizations, and you have a real challenge. One answer to this challenge may be to create visualizations that are as customized as the ontology – visualizations that are automatically tailored to the user based on previous knowledge of their preferences. This is an ideal not easily achieved in the near term. Until such visualizations are feasible, we suggest the use of "explorable visualizations" that show certain dimensions to the user at one time, and allow them to switch between dimension sets in various combinations. Variations on this theme include having each dimension on a layer that can be switched on and off, or having pre-determined pairs or triplets of dimensions that can be viewed. For example, time and location, bandwidth and packets transferred, or people with location and influence.

Yet another variation on this theme would be to begin the visualization with the architectural view, revealing visualizations of various dimensions as the user navigated closer to the relevant nodes and links. Such a visualization would be a good bit more complex to create, but could have benefits to the user.

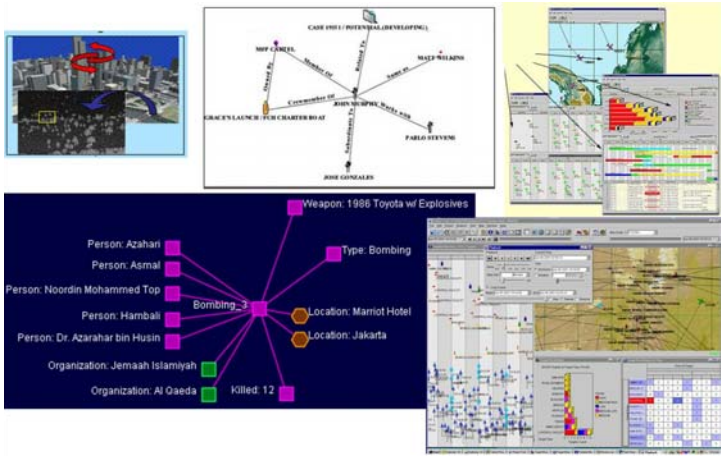

Figure 5: Potential Visualization Styles

A framework for visualizing network information would be a great step forward towards making the creation of such visualizations both possible and practical.

4.2 Impact of the Concept

The relevance of this method to social and other network analysis is of particular importance. When analyzing a social network for person X, more than a tree of associate persons is required. Users will likely need information about the events, groups, people, money, and supplies at what locations and at what times associated with that person and those with which each are, in turn, associated. Such an expanded social network may reveal far more about a person or social network, than would a people-only network. Similarly, information about a computer network that contained not only the network structure, but geographic locations, associated administrators, etc. may serve the user more than a network map alone. In the grandest of all impacts, and certainly farther into the future, is the impact that this concept can have on everyday information searches, and the way we store and use information of various formats.

5.0 A VISION FOR THE FUTURE

Expansion of this concept could include a number of ideas. Search agents for each user based on previous searches and/or user defined parameters can allow continual update of previously created ontologies for a number of days or perpetually. Agents could alert the user by email when an updated ontology is ready. This would involve the agent automatically defining searches for similar entities and relationships and learning from the user's previous searches to decide what might be useful to the user. Continuing, once a user has two or more completed searches, we need the ability to combine related ontology results. This will require the application and integration of ontology merging and negotiation techniques which are not yet mature [5].

The positive aspect of the custom ontology approach is that is does not matter what the resources are, how many times we update them, or how we update them. In theory, one could take the same "search engine" and apply it to any database of trusted resources on any topic. Each custom ontology is as disposable as an internet search results page or can be stored for future use. This flexibility and ease of maintenance make custom ontologies an ideal method for RFI response.

6.0 REFERENCES

[1] Lenat, D. and Guha, R. 1990. Building Large Knowledge Based Systems. Reading, Mass.: Addison-Wesley.

[2] Amy K. C. S. Vanderbilt, Robert I. Desourdis Jr.. Information Fusion for Common Operational Understanding. Extended Abstract. NATO RTO Workshop on Visualization and The Common Operating Picture (VizCOP) (IST-043/RWS-006). Toronto Canada, September 14-17, 2004.

[3] Jana Diesner and Kathleen M. Carley. AutoMap 1.2 – Extract, Analyze, Represent and Compare Mental Models from Texts. CASOS Technical Report CMU-ISRI-04-100, January 2004, Carnegie Melon University, Pittsburgh, PA.

[4] Kathleen M. Carley and Jeff Reminga. ORA: Organization Risk Analyzer. CASOS Technical Report CMU-ISRI-04-106, January 2004, Carnegie Melon University, Pittsburgh, PA.

[5] Bailin, S. and Truszkowski, W. Ontology Negotiation between Agents Supporting Intelligent Information Management. Workshop on Ontology in Agent Systems (OAS-2001), to be held in conjunction with the Fifth Annual Conference on Autonomous Agents, Montreal, Canada, May 28 – June 1, 2001.