Pattern Discovery Using Semantic Network Analysis

Robin Burk[11], Alan Chappell[22], Michelle Gregory [23], Cliff Joslyn[24], Liam McGrath[25]

[1] *Battelle Memorial Institute*
505 King Ave.
Columbus, OH USA 43201
[1] burkr@battelle.org

[2] *Pacific Northwest National Laboratory*
P.O. Box 999
Richland, WA USA
[2] alan.chappell@pnnl.gov [3] michelle@pnnl.gov [4] cliff.joslyn@pnnl.gov
[5] liam.mcgrath@pnnl.gov

***Abstract*—Cognitive information processing at higher conceptual levels requires a computational approach to knowledge representation and analysis. Semantic network analysis bridges the gap between probabilistic pattern recognition techniques and symbolic representations by replacing cumbersome and computationally complex forms of logic-based semantic inference common in symbolic approaches with mathematical metrics on graph representations of labelled, directed semantic networked data. These metrics in turn support assessment of evidentiary support for the presence of patterns of interest in which entities play specified roles in complex event scenarios. The resulting system allows patterns to be specified at higher levels of conceptual abstraction while also remaining robust to conflicting and incomplete information.**

I. Introduction

Probabilistic computing approaches based on brain functions have resulted in significant advances in a wide variety of problems that correlate well to neurological sensory processing. Such approaches face some significant limitations when applied to recognition of patterns at higher levels of conceptual content and abstraction, however. As early as 1990 Minsky argued that intelligent systems would require both bottom-up, associative and top-down, symbolic approaches to knowledge representation and computation [1]. More recently the work of Barsalou on perceptual symbol systems [2] and grounded cognition [3] and of Cassimatis on heterogeneous reasoning approaches [4], among others, have offered general models for bridging the gap between probabilistic learning methods on the one hand and conceptual representation and algorithms on the other.

The explosive growth of digital information makes concept-based pattern detection and exploitation a pressing need today. Such a capability entails finding the meaningful patterns whose implications do not lie on the surface of the data alone, but which acquire their meaning from domain knowledge which in human thought and language is organized and expressed through relationships at the conceptual level. Semantically structured knowledge representations can support novel, statistical approaches to conceptually informed pattern discovery in data extracted from a variety of sources and source types ranging from unstructured human language texts through structured databases and the outputs from sensor processing and pattern recognition systems. In this paper we discuss graph-theoretical analysis on semantically structured and labelled networks of data as demonstrated in the Threat Anticipation Platform (TAP). TAP is a set of software components which can be configured for unattended or user-interactive discovery and assessment of higher-level cognitive patterns of interest in data which have been extracted from a wide array of sources types and semantically fused for analysis.

II. Ontologies and Semantic Data Networks

Initial approaches to computational knowledge representation and reasoning were dominated by propositional and predicate calculus assertions and associated logic-based inferencing systems. More recent work focuses on formal analysis of graphical structures [5] in the form of labeled directed graphs implemented in Resource Description Format [6] and stored in semantic databases [7]. These systems are structured according to ontological typing systems [8] which typically take the form of hierarchical directed acyclic graphs connected by *is-a* subsumptive and *has-part* compositional linkages. Fig. 1 gives the graph and set formalism for ontologies and associated data networks. Also defined in Fig. 1 are other formalisms used within TAP, specifically the knowledgebase and the event-based scenarios which form patterns to be discovered within the semantic network .

Ontologies can be extended through the use of description logics [9] which allow relationships beyond subsumption to form the basis for inferencing. Description logics permit both terminological (concept/class relationship) and assertion (specific fact) statements. Although ontologies and description logics offer improved efficiency of knowledge representation and processing compared to propositional and predicate calculus-based knowledgebases, they are at best unwieldy and are difficult to visualize. In addition, they are often not robust to conflicting or incomplete information. These limitations are among the motivations for a graph-theoretical approach to pattern discovery and analysis.

978-1-4673-1878-5/12/$31.00 ©2012 IEEE

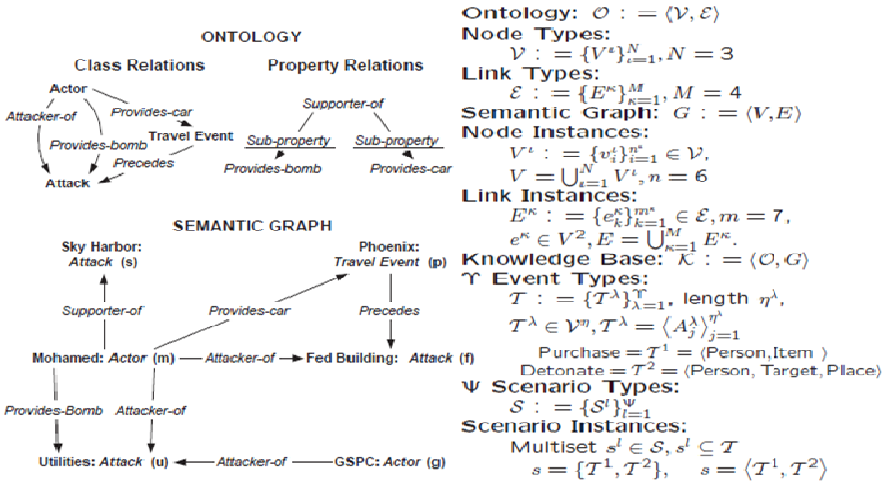

Fig. 1. Formalisms used within TAP.

The substantial literature on developing formal ontologies need not be repeated here. However, two issues often arise with regard to their use. Over time, an ontology may be modified to capture additional data or to represent existing data more accurately, sometimes resulting in duplicate or overlapping concept representations. Supporting research in the TAP project resulted in algorithmic methods for evaluating an ontology and detecting potential ambiguities and inconsistencies. TAP also includes methods for assessing the results of mapping one ontology to another [10]. This is of particular usefulness in data fusion programs built upon existing software and hardware systems in which varying representations occur for the same phenomena. Finally, TAP includes innovative graphical user interface methods to address the severe difficulties associated with user interaction with large ontologies, creating intuitively accessible representations of the ontology for analyst reference when the system is configured for interactive use [11].

III. Populating the Semantic Data Network

TAP includes an automated ability to extract information, including entities and relationships, from structured databases, semi-structured data, and libraries of text documents as well as from other data sources, and to record this data in an RDF semantic graph database. Novel ontology mapping techniques allow for the association of entities and relations to ontology categories, resulting in construction of an ontologically-compliant semantic graph from such input information. TAP can be configured to run automatically or as a user-driven tool; in the latter configuration the system allows an analyst to correct errors in automated information extraction and to resolve references to the same entity across multiple sources if they are not resolved during the extraction process. This is of particular utility when information is extracted from text documents since in such cases ambiguity of linguistic or pragmatic reference are more likely to occur than in more structured information sources. Metadata about the provenance and reliability of the information is maintained and can be modified by analysts as well. Provenance data is available both for manual inspection in interactive configurations and for computational exploitation, if desired, in automatic applications.

Fig. 2 illustrates a small portion of a notional semantic data network in the counterterrorism domain.

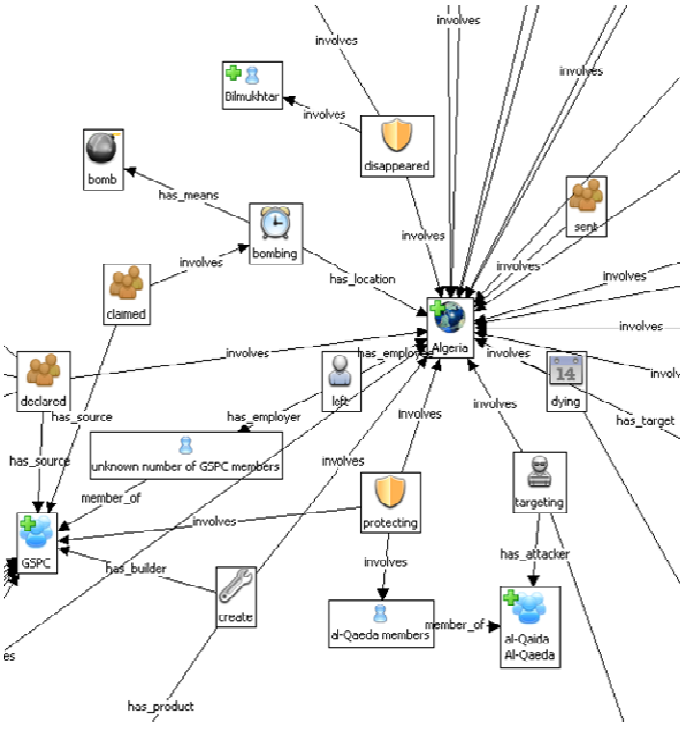

Fig. 2. A portion of a notional semantic data network for counterterrorism

IV. DEFINING HIGHER LEVEL PATTERN TEMPLATES

A. The Challenge

Among potential applications of higher level, cognitively informed pattern recognition, intelligence analysis presents some of the most significant challenges. Intelligence analysts must evaluate large amounts of information in order to anticipate potential threats to national security, military operations and other critical capabilities. This information is collected in a variety of formats ranging from unstructured text documents (Web postings, intelligence reports etc.) to structured databases and multimedia. Increasingly, these sources include information inferred by means of signal and sensor processing as well. These data sources generally are not collected using standardized vocabulary or data models and come from sources of varying provenance and reliability.

A key task in intelligence analysis is to fuse this information, evaluate its (often tacit) meaning and anticipate potential threats as a result of the relationships among entities (people, places and things), and the events in which they participated or might participate. Analysts draw upon broad domain knowledge of the likely patterns of behavior which signal potential threats, along with specific known or surmised facts about a given location, topic or other behavioral context.

Thus the patterns of interest to analysts consist of often-complex scenarios of evolving and interrelated events linked by people, places, organizations or other entities and their relationships to each other and to those events. Adopting the terminology of the analysts, we will sometimes refer to these scenarios or specific pattern instances as "hypotheses".

In most situations, analysts can specify (or hypothetically assume) specific entities for only a subset of the roles in the various events that make up a scenario of interest, or can only specify a broad class of related possible behaviors that might constitute an event. We will sometimes refer below to these partially-specified scenarios or pattern templates as "hypothesis templates".

Note that the need to discover patterns of interest which can be specified only in a partial way or at high levels of abstraction is not unique to the intelligence analysis task. Similar tasks exist in a wide variety of human professions and activities where we commonly use the term 'expertise' to denote the ability to apply knowledge both of domain principles and of background domain facts to assess the implications of a specific set of observed data.

TAP supports the definition of partially-specified pattern templates for any domain. To illustrate the pattern discovery approach embodied in TAP we will follow the intelligence analyst example hereafter since it illustrates the ability of TAP to discover patterns specified at a high level of conceptual abstraction. However, the event / scenario pattern model on which TAP is based reflects fundamental cognitive categories that have numerous applications across a wide range of domains. For instance, a financial transaction is an event and certain combinations of transactions signal the likelihood of fraud; medical test results are events and combined with symptom events signal likely disease occurrence; certain communications traffic patterns suggest cyber attacks or intrusion attempts, etc.

B. The AQIM Example

At the end of 2006 the Algerian Groupe Salafiste pour la Prédication et le Combat announced it had become an operational affiliate of al Qaeda and should henceforth be known as Al Qaeda in the Islamic Maghreb. If true, this announcement had a number of implications, including potential expansion of the group's terror activities to other parts of northern Africa. How might an analyst evaluate this claim and any new threats that GSPC/AQIM might pose?

It is generally not possible to find *direct* confirmation of such claims in physical data, especially since terror groups have been known to falsely claim affiliation for a variety of purposes. Therefore analysts must apply their expertise to assess the *indirect* evidence that might support the claim. In order to do this they hypothesize scenarios of events that would, if present in the data, suggest the claim is true.

For example, we can imagine that analysts might expect to see events such as the following if GSPC has become an operational affiliate of al Qaeda:

- Communication between GSPC/AQIM members and al Qaeda members
- New funds transfers to GSPC/AQIM, especially from sources or through channels known to fund other al Qaeda groups and especially if it follows after communications between the groups
- Travel of GSPC/AQIM members to al Qaeda training camps outside Algeria or establishment of new training camps within Algeria, especially if this follows new funding
- Acquisition or creation of methods of attack characteristic of al Qaeda, e.g. specific types of bombs, IED construction etc.
- Attacks on new types of targets characteristic of al Qaeda

Note that the analyst must evaluate the strength of evidence for the overall claim in terms of scenarios consisting of temporal sequences of events or event alternatives, some of which are specified in general terms (funding), others more specifically (certain types of attacks on certain types of targets). The requirement to apply varying levels of conceptual abstraction along with alternate sub-patterns (event type A or B) within a single pattern discovery activity illustrates why statistically-based pattern discovery techniques alone are inadequate to support analytic efforts of this sort.

C. Formally Modeling the Hypothesis Space

TAP's novel method for representing pattern instances as hypotheses about event scenarios uses a structure based upon the grammar of natural language. An *hypothesis* is defined as an expression over a set of *events*. An *event* is a verb with a set of syntactic *roles* dependent upon the verb. Each *role* in a specific event or hypothesis is filled by exactly one entity. An

hypothesis expression then combines events using Boolean connectors (e.g., *And*, *Or*), temporal constraints derived from Allen's interval algebra [12], and constraints on which entities can participate in which roles among events .

An *event class* consists of a verb and a set of *role classes*. A *role class* is a structured description of the set of entities that may participate in that *role*. Constraints may be applied to participants across events and roles; for instance, the participants in a subset of roles may be constrained to either be the same entity or be different entities (e.g., the object in Event 1 must be the same as the Subject in Event 2). A *completely specified event* is one where some specific entity is associated with every role. A completely specified hypothesis is one in which all events are completely specified.

Fig. 3 illustrates a portion of a pattern template. In this template Planner1, Planner2, Buyer1 and Buyer2 are specific entities. Builder and Location are entity classes. Purchase1 and Purchase2 are different, specific events of the Purchase class.

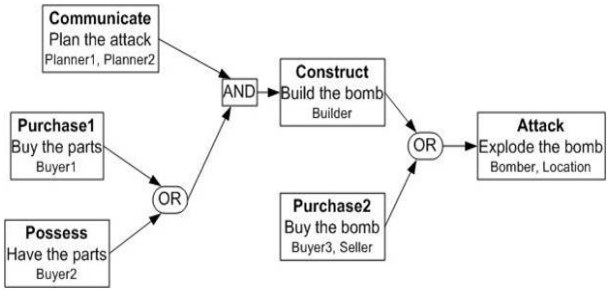

Fig 3. Extract from a partially-specified pattern template showing event and role types.

Because events and roles are linked to the subsumption and compositional hierarchies in a domain ontology, the TAP pattern representation scheme is extremely flexible. Both verbs and role assignments in events can be specified at a very detailed level or at a higher (class-wide) level. In this way, using a symbolic domain knowledge representation in the form of an ontology provides the capability to specify and discover meta-pattern instances at a wide variety of conceptual levels of granularity against a single semantically-organized data collection.

V. The TAP System

Fig. 4 illustrates the overall flow in an interactive TAP application. TAP processing flow consists of several stages. First, information about entities and events is extracted from sources using natural language and other techniques, as guided by the ontology or ontology facet applicable to this domain. This information is stored in a knowledgebase, along with meta-information regarding the source of each assertion. The entity and event information forms a semantic network which captures the relationships among entities and between entities and events.

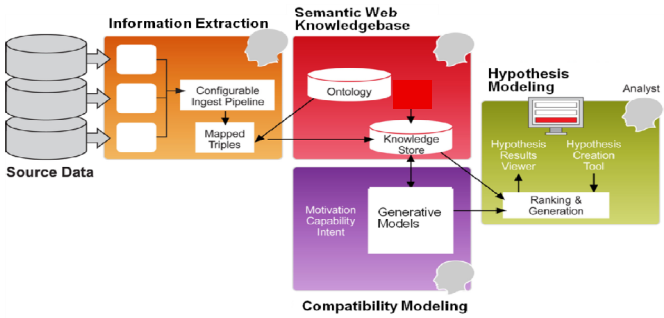

Fig. 4. TAP Flow

In an interactive application of TAP, analysts use the software's graphical user interface to define a pattern template. Alternately, pattern templates can be pre-defined within an automatic application and applied to data streams without user interaction. In either type of application, TAP then utilizes semantic network analysis, along with generative domain models if desired, to identify specific pattern instances and rank them according to the weight of supporting evidence from the source data.

VI. Semantic Network Analysis

The facts within the semantic knowledgebase constitute a network or graph in which entities are vertices or nodes, and the specific relationships between entities (e.g., object properties) are directed edges or links. Such a graph is, in reality, multi-modal or labeled: each vertex has a type (e.g., person, organization, etc.) and each edge has a type (e.g., employee-of, has-experience-with). While analysis of simple, unlabeled, and undirected networks has been addressed in research for some time, analysis of labeled, directed graphs is not nearly as mature and presents a number of computational challenges.

A core TAP capability is to provide a score for the amount of evidential support of events and hypotheses within the semantic graph database. In turn, a major factor of this score is the computation of a semantic distance between entities.

Distance measures are a key element in clustering and other machine learning techniques. However, defining "distance" with regard to higher level concepts requires both a formalism and also tractable computational algorithms which can be applied against instances of that formalism. This is made more difficult by the multi-dimensional character of conceptually-based knowledge as captured in data structures.

Most simple graph-theoretical measures of distance only consider the shortest untyped, undirected, unweighted path between pairs of nodes. In order to support complex analyses and to allow tradeoffs between granularity of discovery and computational complexity, TAP has multiple distance metrics available. The primary distance metric most often used assesses all paths between those nodes. Another metric allows users to specify the path types (including link type and direction) they are interested in. In this way, TAP can not only factor in the length of the paths as measured in links, but also

link and entity type, number, and topology. TAP uses this semantic distance as a proxy for the likelihood of there being an existing but unknown association or a possible future association between the entities. Where less granularity of analysis is required to discover and assess patterns of interest, simpler distance metrics may also be applied.

Not all relationships are of equal pragmatic implication. A key challenge in representing and applying higher level conceptual knowledge lies in the need to capture and apply domain insights which are key (and often tacitly invoked) in analysis as executed by a human domain expert. To address this challenge, TAP accepts weights applied to either or both the entities (vertices) and the relationships (edges) in the network. This can be used to make the entities or relationships more or less important based upon other criteria such as their semantic type. In addition TAP includes a semantic path type weighting mechanism which effectively replaces a path consisting of a sequence of directed relationships of a certain type with a single, weighted, direct relationship.

Unlike most simple methods of network analysis which are only able to compute the distance between pairs of individual nodes, TAP computes the distance between two sets of nodes and the variance of a set – a measure of how far the entities in a set (e.g., event) are from each other. These abilities enable the scoring of events and hypotheses.

TAP uses these concepts of distance, coupled with spectral analysis of matrices, to cluster the nodes in a network. These methods have proven highly effective on both canonical network analysis datasets and on the networks derived from semantic knowledge bases of terrorist activity. TAP combines clustering with other transformations such as filtering of data by type to create techniques for identifying clusters of activity within graphs with some level of noise (i.e., irrelevant data).

VII. Finding and Ranking Pattern Instances

As stated earlier, the main purpose of TAP is to identify and evaluate specific pattern instances (hypotheses) within a defined pattern class, based on the facts or evidence assembled in the semantic network. To do this, TAP relies predominantly upon statistical inference rather than logical inference to select candidate entities to fill specified roles. While logical inference is the traditional method of processing semantic knowledge, such inference relies heavily upon the accuracy, self-consistency, and, to some degree, the completeness of the knowledge. In particular, in the absence of self-consistency, inference via two-valued propositional logic fails entirely, enabling either no or any conclusions to be drawn. In addition, two-valued propositional logic provides no means to evaluate one hypothesis against another – a hypothesis may only be shown to be possible or impossible. These limitations are particularly problematic in intelligence analysis, where the available data sources may not be reliable and may yield information that is incomplete and contradictory. Similar conditions obtain for many other areas of expert judgement, such as in detection of financial market manipulation, insurance fraud and early diagnosis of non-typical medical conditions.

By contrast to formal logic-based inference, TAP does not require or apply a rigid two-value truth function. Rather, facts are weighted in terms of various dimensions including confidence and relevance.

TAP scores hypotheses based on a statistical preponderance of evidence. The resulting ability to handle inaccurate, incomplete, and inconsistent data is critical in many domains and has historically been a key advantage of probabilistic machine learning approaches over systems based on symbolic knowledge representation and logical inference. TAP bridges the gap between these methods, bringing the resilience of probabilistic machine learning techniques to conceptually-structured information processing.

TAP computes the marginal relative likelihoods for partially-specified events and hypotheses (i.e., where only some roles are associated with specific entities). This includes the ability to determine the relative likelihood of each entity in the knowledgebase to participate in each role in each event in the hypothesis, subject to the constraints imposed in the template.

The ability to compute marginal relative likelihoods for each entity enables the final, powerful ability to generate lists of the most likely events and hypotheses by assembling combinations of the most likely entities.

VIII. Assessing Role Compatibility

As described above, TAP discovers and ranks specific instances by evaluating possible entity/role assignments for those roles whose participants are not uniquely specified in the pattern template.

When humans evaluate hypotheses, we draw on our knowledge of the domain and of the entities and events in questions. This knowledge, often integrated at a tacit level in professional experts, provides an important ability to assess whether specific entities are really likely to play a given role in an event or event type within a given, broader context or scenario or whether their presence in the data is in fact coincidental to the issue under evaluation. Especially in the case of human relationships and social interaction, statistical correlation can be a poor indicator of intent, motivation or other characteristics germane to the purpose of a pattern discovery application. (This is illustrated by the relatively low level of information conveyed by many ‘friend’ relationships in social media such as Facebook and Twitter.)

Although ontologies are a powerful way to model many conceptual relationships within a domain that would be applied by such an expert, capturing all of the key relevant knowledge and expertise of a human is sometimes awkward or even impossibly cumbersome within this formalism alone. This is especially true when the patterns, events and entities involved are at a higher, broader conceptual level. Therefore, TAP offers the capability to augment semantic graph analysis with domain-specific methods and knowledge beyond those embodied in the ontology for a given application. This

capability is typically utilized to augment evaluation of compatibility.

Compatibility is a measure of how likely an entity (e.g., a person) is to play a certain role (e.g., attacker) in a certain type of event (e.g., suicide bombing) relative to all other entities in the knowledgebase. For instance, compatibility assessment in TAP as applied to the domain of terrorism utilizes models that assess motivation, capability, and target value of certain entities. Motivation quantifies the relative likelihood of an actor (e.g., a person or organization) to play any role in an event. Capability quantifies the relative likelihood of an actor to perform a certain action (e.g., how capable is a certain person of constructing a certain type of weapon). Target value quantifies how desirable a target is to a certain class of attackers.

An automated, quantitative model for motivation assesses facts in the semantic knowledgebase in light of input from socio-cultural experts specialized in terrorism. This model assigns compatibility values to all of the actors present in the knowledgebase for a given role/event combination. The compatibility assessment is then factored with the semantic network analysis metrics when assigning entities to roles and ranking the resulting specific hypotheses.

Because compatibility assessment draws on domain-specific insights, TAP does not constrain the computational method used for this purpose, requiring only that the method provide a numeric compatibility score. TAP applications can therefore be tailored as closely as desired to a given domain and problem set. Methods for assessing compatibility include manual assignment, elicited expert knowledge in the form of factor models, domain rules, models generated through machine learning and hybrid approaches. These methods are particularly useful for factoring socio-cultural differences into role compatibility evaluation.

IX. Other Methods for Incorporating Expert Opinion

The degree to which expert understanding can be captured in a tractable ontology differs widely among domains and problems of interest. Therefore, in order to apply key semantic network analytic techniques effectively across a wide range of applications, TAP allows expert opinion to be included in a number of other ways if desired. In addition to the generative compatibility models described above, expert opinions can be entered as facts into the knowledge base. Analyst expert opinion can also be included by manually setting compatibility and confidence values.

Two features of TAP are of particular and critical value in the context of expert opinion. First, multiple – even conflicting - opinions can be included in the models, facts, and the compatibility and confidence values. This allows the insights or perspectives of multiple experts to be included without requiring that they be reconciled or one selected over the other. And second, the TAP scoring algorithms are inherently robust to contradictory information, evaluating the alternatives based on preponderance of evidence, including any additional expert opinion that may have been added.

The resulting capability is resilient, open to enhancement over time and able to be tuned to as fine a degree of analysis as is cost-beneficial in terms of processing, model construction etc. to meet a given application requirement.

X. Conclusion

Cognitive techniques for processing information using higher level conceptual structures have the potential to bridge the gap between powerful clustering and other pattern recognition methods, on the one hand, and human-level knowledge analysis on the other hand. The Threat Anticipation Platform demonstrates the utility of semantic network data representations and the use of mathematical rather than logic-based analyses to identify and rank specific pattern instances of interest. In recognition of the complexity and often tacit nature of human domain expertise, TAP also allows integration of other computational models in order to incorporate domain knowledge into the analytic process.

Acknowledgment

This work was supported by research and development funding from Battelle Memorial Institute.

References

[1] Minsky, M., “Logical vs.Analogical or Symbolic vs. Connectionist or Neat vs. Scruffy”, *Artificial Intelligence at MIT, Expanding Frontiers*, Patrick H. Winston (Ed.), MIT Press, 1990, vol.1.

[2] L. Barsalou, “Perceptual Symbol Systems”, *Behavioral Brain Science*, vol. 28, pp. 577-660, 1990.

[3] L. Barsalou, “Grounded Cognition”, *Annual Review of Psychology*, vol. 59, pp. 617-635, 2008.

[4] N. Cassimatis, “A Cognitive Substrate for Achieving Human Intelligence”, *AI Magazine*, vol. 27, no. 2, 2006.

[5] B. Gantner, S. Gerd and R. Wille, Eds., *Formal Concept Analysis: Foundations and Applications*, Springer-Verlag, 2005.

[6] http://www.w3.org/RDF .

[7] B. McBride, “Jena: a Semantic Web Toolkit”, *IEEE Internet Computing*, Vol. 6, No. 6, pp. 55-59, 2007.

[8] http://www.w3.org/TR/owl-features .

[9] P. Hitzler, M. Krőtzsch, B. Parsia, P. Patel-Schneider and S. Rudolph, “OWL 2 Web Ontology Language Primer”, http://www.w3.org/TR/2009/REC-owl2primer-20091027/ .

[10] C. Joslyn, P. Paulson, A. White, “Measuring the Structural Preservation of Semantic Hierarchy Alignments”, *Proc. 4th Int. Wshop. on Ontology Matching (OM-2009)*, CEUR, v. 551, 2009.

[11] A. Chappell, A. Bladek, C. Joslyn, E. Marshall, L. McGrath, P. Paulson, S. Stolberg, A. White, “Supporting the Analytic Knowledge Manager: Formal Methods for Ontology Display and Management”, *2009 Conf. on Ontologies for the Intelligence Community (OIC 2009),* CEUR Vol. 555, 2009.

[12] J. Allen, “Maintaining Knowledge about temporal intervals”, *Communications of the ACM*, Vol. 26, No. 11, pp. 832-843, 1983.