

# TRAKS<sup>1</sup>: Terrorist Related Assessment using Knowledge Similarity

Boanerges Aleman-Meza, Chris Halaschek, Satya Sanket Sahoo

Department of Computer Science  
University of Georgia  
Athens, GA. 30602-7404  
{boanerg, ch, saho}@cs.uga.edu

## 1 Introduction

Since the terrorist attack on September 11, 2001, homeland security has been a major topic of both research and application. The identification and possible prevention of various contributing factors to terrorist activities, such as money laundering, identity theft, terrorist planning, etc, are in high demand. The terrorist attacks on September 11 have reinforced this area of interest dramatically. These attacks alone caused an estimated “\$120 billion of damage” [1] and claimed around 3000 lives [2]. Clearly, something must be done to prevent the types of events from happening in the future.

Currently, automated solutions for detection in this domain primarily include anti-money laundering software which is essentially based on data mining techniques, rule based mechanisms, etc. While this has had benefits, there are several disadvantages. First, for detection the relationships that compose money laundering operations depend on the structure of the database schema being used in a specific financial institution. For the system to be able to find a money laundering operation there must be a structural match between the data within the database and known rules and/or patterns. This lacks the implicit information when the data is semantically annotated. Hence it would be beneficial to find not only structure similarity but semantic similarities as well. Second, if the data is semantically annotated, there are possibilities of further finding interesting connections. For example, given a bank in Afghanistan that has personnel with semantic relationships with terrorists, a system based on semantic similarity will be able to consider this relationship relevant in the domain of potential terrorist activities. In contrast, this type of model would not be able to be automatically discovered in traditional fixed rule based systems. This is because for this to be possible, the database schema of the financial institution would have to model such relationships.

---

<sup>1</sup> <http://lsdis.cs.uga.edu/proj/traks/>

This document is structured as follows. Section 2 describes how we plan to use Semantic Web technologies in our approach to finding data based on templates. Section 3 provides the framework on which the design decisions were made. Our prototype system, TRAKS, and its architecture is described in detail in Section 4. The preliminary results are presented in Section 5. Finally, Section 6 gives our concluding remarks and lists future research directions.

## 2 Background and Motivation

“Tracking and intercepting the unlawful flow of drug money is an important tool in identifying and dismantling international drug trafficking organizations with ties to terrorism” [17].

With the advent of current technology, data is now being semantically annotated. Hence, there exist many sources that describe different characteristics of a given entity in diverse domains. We plan to use (and if needed extract) semantically marked up data along existing data from interested institutions to find potential terrorist activities. Our proposed approach will employ past known money laundering, id theft, and terrorist attack models/templates to discover potential threats in the knowledge base based on novel semantic similarity algorithms that we will develop. An example source for past real money laundering operations is available at [3]. Furthermore, financial institutions are required by the Bank Secrecy Act to identify suspicions of money laundering operations and notify authorities [18]. There are also money-laundering requirements result of the U.S.A. Patriot Act.

Approaches for detection of money-laundering operations include data-mining techniques, rule-based systems, among others. Rule-based systems “Tend to generate enormous numbers of trivial alerts” [20].

We aim to showcase these capabilities with a prototype that makes use of data represented in proposed knowledge representation languages for the Semantic Web. The vision of an extension of the current Web, termed the Semantic Web [5], is based on the prediction that information on the Web will be machine processable. In order to achieve this evolution, the current Web has foundations on technologies that allow interoperability whereas one of the first steps has been markup data via XML [6]. The basic language to provide meaning to data by marking up URIs is RDF (Resource Description Framework) [7]. Data annotated by following the RDF model will be *machine-processable* only if other machines *know* the ontology with respect to which the data is marked up. Whereas a machine will *know*

a given ontology by means of hard-coding the intended meaning of the annotations or by automatically inferring meaning programmatically is an open issue. We believe that RDF will follow the evolution that XML has had in industry. By adding *tags* to existing data in databases, companies will 'output' data marked up using RDF making reference to the ontology and will keep their data in relational databases. Whereas RDF is intended to mark up data, RDF Schema [8] is the counterpart of RDF that provides the means to define an ontology by specifying a hierarchy of classes and the relationships among them. In our project we make the assumption that financial institutions will use Semantic Web technologies to take advantage of the expressiveness of knowledge representation for diverse types of data (un-structured in nature) about their customers, such as citizenship, companies they own, business associations, credit history, etc.

We indeed go one step further by designing and building our system architecture in terms of richer languages being pushed by the World Wide Web Consortium<sup>2</sup> such as OWL [8]. The main differences with RDF and OWL are in respect to the *reasoning* capabilities that can be used more directly with OWL DL, one of the three flavors of OWL. Due to the nature of our template-matching application, we do not make extensive usage of the *new* constraints available in OWL.

There have been proposals of query languages for RDF data, such as RDQL [9] by HP, TRIPLE [10], and RQL [11]. However, they do not completely provide means to query RDF data based on a template, as pointed out in [4]. Our project addresses template-based querying of semantically marked up data. Furthermore, we exploit the explicit (and implicit) implications of the nature of the hierarchy in order to find *semantically* similar matches. This is critical in scenarios of money laundering, terrorism activities and identity theft where known operations are being tracked by current money laundering systems. Their drawback is our strength by relying on *semantic similarity* to find those scenarios that are not an exact match of known operations.

We introduce the concept of a 'core' template that captures the essence a known scenario in either money laundering, terrorism activities and/or identity theft. After filtering out potential matches to known scenarios with the core template, the rest of the template definition provide means to our system to flag a set of entities that are related in a *similar* way to those known a priori.

### 3 Framework

---

<sup>2</sup> <http://www.w3.org/>

The foundations of TRAKS are based on the concepts of a Template, Core template, ontology, datasets, semantic similarity, and semantic ranking.

**Definition 1, Template.** A template consists of a set of classes and relationships that form a connected graph where the classes are nodes and the relationships are arcs connecting the classes. The set of interconnected classes and relationships aims at capturing a scenario of interconnected entities.

In a real-world scenario, a template could capture a known money laundering operation, an identity theft scheme and/or terrorism related activities. The purpose of the template is to be a small model of the entities that are to be found in a large dataset. In terms of information retrieval, it is simply a query. The potential of querying templates is, however, very important. An intelligence analyst could discuss a terrorism related scenario with colleagues and describe it by using a template. Later it could try to find potential matches of the template in a large dataset. Thereby, a process of hypothesis and testing is possible based on manufactured scenarios.

**Definition 2, Core template.** A core template is a subset of the classes and relations of a template (definition 1). The idea of a core template is to capture the essence a known scenario represented in a template. The purpose of a core template is to provide a means to filter out potential matches to a template thus making more efficient the template matching algorithm.

**Definition 3, Semantic Similarity.** We define two entities to be similar if (i) both belong to the same class, (ii) both belong to classes that have a common parent class, or (iii) one entity belongs to a class that is a parent class to which the other entity belongs. Furthermore, two relationships are similar if (i) both belong to the same class, (ii) both belong to classes that have a common parent class, or (iii) one relation belongs to a class that is a parent class to which the other relation belongs.

Two semantically similar entities are, for example 'John', who belongs to the class "CEO", and 'Anna', who belongs to the class "CTO" where both classes have a common parent "Management Board". Two semantically similar relations are, for example 'associated with' and 'member of' where 'member of' is a specialization of the relation 'associated with'.

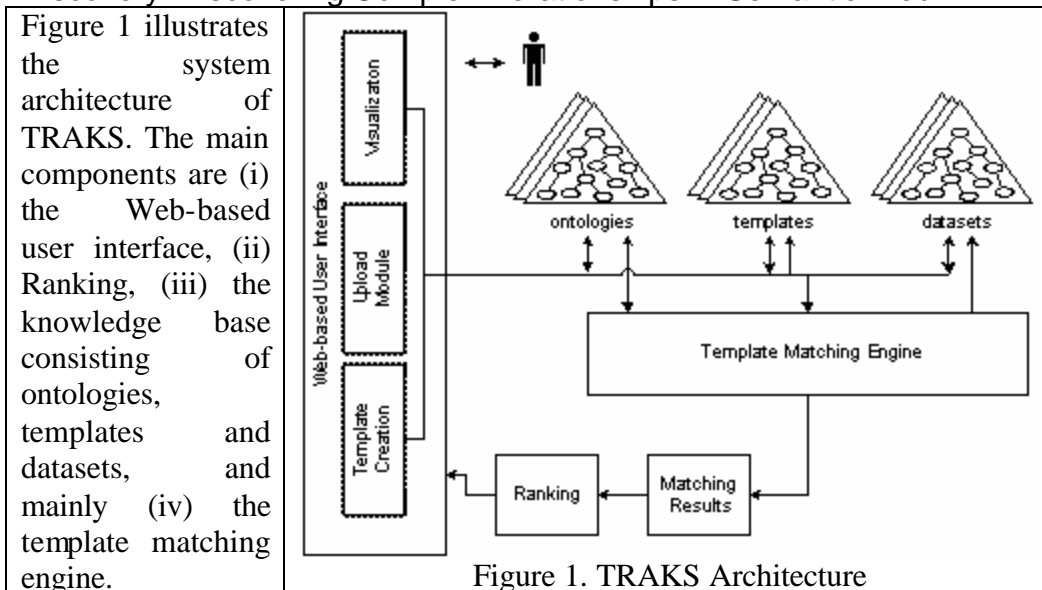
**Definition 4, Semantic Ranking.** The semantic ranking criteria considers the overall number of entities and relations that match a template by assigning an evaluation to each of them depending on how distant the class they belong is with the class of the slot in the

template. A perfect match is ranked higher and would consist of entities and relations that reflect exactly the classes of the template.

Semantic ranking is based on the consideration that entities or relations closer to their respective place in the template are more representative of the scenario that the template captures. Related research in similarity measures considers for example, that entities located at a lower place in an ontology are more specialized than those located at a higher place in the ontology [13].

#### 4 System Architecture

TRAKS system architecture was designed based on insight and experience gained by previous research in [12] and other ongoing projects at the LSDIS Lab such as the NSF-funded project ‘Semantic Discovery: Discovering Complex Relationships in Semantic Web<sup>3</sup>’.



##### Web-based User Interface

A design issue was to get feedback on the use of our system. By providing a web-based interface, there are no issues related to operating system or programming language compatibility. The interface contains an ‘upload’ module that allows the user to specify his/her own ontology, dataset and template. For simplification of the template definition, there is a module ‘template creation’ that provides a means to create a template without syntax complications.

Designing the visual component of the Web interface was facilitated by the usage of Touchgraph<sup>4</sup>. We have developed a module that transform an ontology defined in OWL into the format used by

<sup>3</sup> <http://lsdis.cs.uga.edu/proj/semDIS/>

<sup>4</sup> <http://www.touchgraph.com/>

Touchgraph. Furthermore, we have the extra functionality of labeled edges with the name of the relationship between entities. For the case of the visualization of templates, different colors differentiate the classes and relations that belong to the core template. Similarly, instance data and ranked results are as well presented to the user with the graphical Touchgraph environment. Figure 2 illustrates the visualization of a simple ontology.

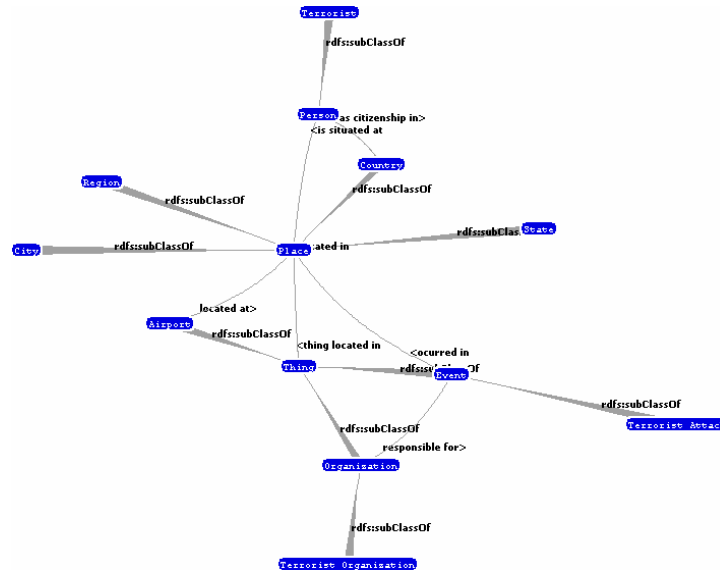


Figure 2. Visualization of an ontology

Visualization of the results and/or templates gains importance due to the complexity of some well known money laundering operations. As stressed in [19], a visualization tool provides a means to model the data and also is valuable to investigate a potential money laundering case.

### (Semantic) Ranking

The ranking component assigns a estimate value to a matching result based on how the set of entities and relationships reflect those captured by the template. Ranking a set of interconnected entities and relations is more complex than ranking a set of documents or paths of semantic associations [14].

### Knowledge Base

This module organizes the ontologies, templates, and datasets provided via Web by users. Besides the Web-accessible uploaded ontologies, we have designed an internal structure that relates ontologies to templates and to datasets. Instead of processing XML serializations of these sets, we store them in a prototype knowledge base. The operations of the knowledge base include typical

maintenance utilities as well as an API that is used by the template matching engine. Currently, the API is based in spirit to that of the JENA project [15].

## 5 Initial Results

Available data for our experiments is based on the research of the NSF-funded project “Semantic Association Identification and Knowledge Discovery for National Security Applications” [12]. Therefore, the preliminary ontology is within the domain of terrorism. This has been extended to include entities in the financial domain, such as banks (domestic and international), CEO’s of companies, and so on.

With a mini set of the ontology in [12], we have identified two cases that match templates. These results are then ranked based on distance similarity measures that take into consideration the ontology itself. We have used Touchgraph to visualize the results in a graphical manner.

## 6 Conclusions and Future Work

Detection of money laundering activities, identity theft operations, and terrorism related activities is a driving motivation to enhanced knowledge discovery research. We address the problem of identifying known scenarios described in form of a template by searching in large datasets. Furthermore, we make use of semantics to go one step closer to realistic situations where variations of known scenarios are intended to be hidden by traditional exact matching algorithms. Thereby we describe the TRAKS system, where it is possible to identify *similar* known scenarios of money laundering, identity theft and terrorism related activities. The architecture of TRAKS was designed by considering open standards for knowledge representation that are becoming more popular with the vision of the Semantic Web. TRAKS is a system that is Web-based and therefore platform independent and makes use of XML and RDF technologies for an open access to it. Therefore, it is possible for anyone who has data annotated with an ontology to use our system. The only step required, *template creation*, is carried out also in a Web-based form. Furthermore, TRAKS visualization provides an easier understanding of the results found, unstructured in nature. The results, ranked based on degree of semantic similarity, provide means to financial institutions and/or intelligence agencies to detect potential scenarios of importance for increased economic and national security.

In respect to the API to the knowledge base we will evaluate the proposed API by [16] that was developed considering different approaches taken by tools that read RDF or OWL data, such as ontology editors, knowledge sources management, and inference engines among others.

Several research directions will address scalability issues. Among the first is the limitation of the search space by considering only the newly added information to the knowledge base.

Presentation issues can help improve usability. Our initial design of the module of template creation could be further improved with one that only uses Touchgraph events manipulation.

## 7 References

- [1] The Economic Cost of Terrorism.  
<http://usinfo.state.gov/topical/econ/mlc/02091004.htm>
- [2] September 11, 2001: A day of terror.  
<http://www.cnn.com/2003/US/03/10/spri.80.2001.terror/index.html>
- [3] Financial Action Task Force on Money Laundering Homepage.  
<http://www1.oecd.org/fatf/index.htm>
- [4] <http://lists.w3.org/Archives/Public/www-rdf-interest/2003Nov/0057.html>
- [5] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. Scientific American, May 2001
- [6] T. Bray, J. Paoli, C. M. Sperberg-McQueen, and E. Maler. Extensible Markup Language (XML) 1.0 (Second Edition). W3C Recommendation, 6 October 2000.
- [7] O. Lassila and R. Swick. Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation, 1999.
- [8] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, L. A. Stein. OWL Web Ontology Language Reference, W3C Candidate Recommendation, 18 August 2003.
- [9] A. Seaborne. RDQL: A Data Oriented Query Language for RDF Models. 2001.
- [10] M. Sintek and S. Decker. TRIPLE ---A Query, Inference, and Transformation Language for the Semantic Web. International Semantic Web Conference (ISWC), Sardinia, June 2002.
- [11] G. Karvounarakis, S. Alexaki, V. Christophides, D. Plexousakis, M. Scholl, RQL: A Declarative Query Language for RDF, The Eleventh International World Wide Web Conference, May 7-11, 2002, Honolulu, Hawaii, USA.
- [12] Amit Sheth, Boanerges Aleman-Meza, I. Budak Arpinar, Clemens Bertram, Yashodhan Warke, Cartic Ramakrishnan, Chris Halaschek, Kemafor Anyanwu, David Avant, F. Sena Arpinar, Krys Kochut. Semantic Association Identification and Knowledge Discovery for National Security Applications. Technical Memorandum 03-009, August 2003 (to appear in Special Issue of Journal of Database Management on Database Technology for Enhancing National Security, Eds: L. Zhou and W. Kim).
- [13] M. Rodriguez, and M. Egenhofer, "Determining Semantic Similarity among Entity Classes from Different Ontologies", IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 2, March/April 2003.
- [14] Boanerges Aleman-Meza, Chris Halaschek, I. Budak Arpinar and Amit Sheth, Context-Aware Semantic Association Ranking, Proceedings of the First International Workshop on Semantic Web and Databases, pp. 33-50. Berlin, Germany, September 7-8, 2003

- [15] B. McBride "Jena: Implementing the RDF Model and Syntax Specification", in: Steffen Staab et al (eds.): "Proceedings of the Second International Workshop on the Semantic Web - SemWeb'2001", May 2001.
- [16] Sean Bechhofer, Raphael Volz and Phillip Lord. Cooking the Semantic Web with the OWL API. Proceedings of the 2nd International Semantic Web Conference, Sanibel Island, Florida, USA. 2003.
- [17] DEA Congressional Testimony, April 24, 2002. Available online at: <http://www.usdoj.gov/dea/pubs/cngrtest/ct042402.html>
- [18] How Well Do You Need To "Know Your Customer?", Alvin D. Lodish, Bilzin Sumberg Baena Price & Axelrod LLP. December 2, 2003. Bank Systems and Technology, online at: <http://www.banktech.com/story/amLaundering/showArticle.jhtml?articleID=16401323>
- [19] Jonathan Thomson, Matthew Brace, and Richard Hurst. "A tour of anti-fraud technology". Fraud Intelligence, June 2003, issue 58, pp. 10-12.
- [20] Jason Kingdon, "Applying technology to fight money laundering". Available online at: [http://www.bankingmm.com/money\\_laundering.htm](http://www.bankingmm.com/money_laundering.htm)