An Ontology Based Approach to Data Surveillance

Teresa Pereira[1] and Henrique Santos[2]

[1] Informatics Department
Superior School of Business Studies
Polytechnic Institute of Viana do Castelo
Valença, Portugal
tpereira@esce.ipvc.pt

[2] Information System Department
School of Engineering
University of Minho
Guimarães, Portugal
hsantos@dsi.uminho.pt

Abstract: Nowadays the terrorist threat took proportions that concern governments and the national security organizations, all over the world. A successful terrorist incident usually brings catastrophic results. However if a terrorist attack can be predicted and characterized, it may be possible to organize a proper intervention in order to avoid it or to reduce its impact. The management of information is becoming an important issue in the domain of security information systems. The information access and association, analysis and assessment, and finally exploitation have become the focus for all security information services and governments. Current surveillance approaches are not very efficient leading innocent citizen to the confrontation of law enforcement services. One reason for this, result from the difficulties of the current system to extract knowledge or concepts abstracted from massive databases of information. Knowledge based methods, such as ontologies can integrate data surveillance, and enable a proper data analyse improving the performance of the security information services. This paper intends to present a perspective about the use of ontologies in the context of data surveillance, and present its importance in the current security services domain.

Keywords: Ontology, surveillance, data surveillance, security information systems, knowledge representation.

1. Introduction

The tragic events occurred in 9/11, as well as the ones that followed, forced many countries to review the efficiency and the efficacy of their information systems security. The information management has become a main concern of the national security organizations. The amount of digital information available in the Web and its accessibility makes the Web an important information source and therefore a powerful weapon regarding terrorism. It must be noted that, currently, the Internet is the most used communication medium rather the traditional means. Consequently we observe the development of new security technologies which are readily available to everyone, in response to a continuously rise of threats from all over the world. Furthermore, it starts to be recognized the importance of the interoperability between diverse information systems, in order to promote the security information exchange.

Nowadays, security organizations collect a large amount of data from the Web and from different information sources in a daily basis, resulting in huge databases. On these databases an intensive surveillance is performed in order to track suspicious activities.

However, the increase amount of data to analyse and its inherent heterogeneity makes traditional analysis mechanisms inefficient, demanding the use of automatic and effective knowledge management techniques. Currently, governments and security organizations already have access to sophisticated data mining technologies with advanced statistical techniques for that purpose, but the results retrieved by those techniques are raising several questions regarding de false positives that can be generated, which bring potential negative side-effects, for instance, leading innocent citizen to the confrontation of law enforcement services (Maxwell 2005).

The main problems are the amount, the heterogeneity and dynamic nature of data available on the Web, and it becomes absolutely necessary to structure and organise it for knowledge retrieval. In fact it is very difficult to incorporate knowledge or concepts abstracted from the low level data, into statistical analyses.

To address these issues improved data surveillance approaches must be developed. Knowledge based mechanisms seems to be an appropriate strategy in order to enable a better data analyse and therefore a faster detection and characterization of threats and attacks.
In this context, the knowledge organised according to ontologies helps to abstract the essence of the concepts, besides to enable a better interoperability between different information systems, and consequently improving the performance of the security information services.

In this paper we describe how an ontology based approach to data surveillance can be helpful in the current security services domain, and is structured as follows: in the section 2 will be presented an overview about the data surveillance and the associated technologies currently used. In the section 3 a context of the ontologies and its evolution is provided. In section 4 it is introduced a perspective about the use of an ontology based approach in the context of data surveillance. Lastly, some conclusions are presented in section 5.

2. Data Surveillance

Terrorism and surveillance have always been directly related and will remain that way due to the continue urgent need of some sort of control over the proportions that terrorism threats reached. The United Nations presents a definition of terrorism in the report entitled *Larger Freedom* on 17 March 2005:

> "[any action] intended to cause death or serious bodily harm to civilians or non-combatants with the purpose of intimidating a population or compelling a government or an international organization to do or abstain from doing any act."[1]

Surveillance over terrorist groups' activities always existed and will continue. The abolishment of internal border control in Europe and the tragic events of the 9/11 as well as the ones that followed forced all governments to impose new requirements in the efficacy and the efficiency of their own security organizations. Today the paramount interests of all governments concerning the security are on: (1) the borders control and security; (2) illegal immigration and (3) protection against terrorist attacks (Evelien 2005).

The Schengen Information System (SIS) is a huge database that stores information of millions of objects and individuals that is shared by different European countries for different purposes (Evelien 2005). However, SIS is not the only database used in Europe. Additional information systems such as those maintained by Eurodac and Europol collect and share information to control immigration and safeguarding security (Evelien 2005). In fact the primarily use of SIS was to control illegal immigration. However, the dramatic emergence of the terrorist threats, have been promoting the discussion to extend the use of the SIS to different purposes, namely the establishment of a new Visa Information System and the use and storage of biometrical data. These databases store, in a daily bases, an extensive amount of information, becoming difficult to perform manually surveillance on these data. The National Security Agency (NSA) of the United States of America (USA) with the cooperation of the nation's telephone companies, created a database with the telephone records of millions of Americans. This database included detailed information for every telephone call made within the nation's borders, enabling to track hidden terrorists cells (Garfinkel et al 2006).

The security services agencies noted that the Web is also an influent information source besides the telephone records and the management of transactional data gathered by the movements of citizens. The presence of new and sophisticated communication technologies and the increase use of digital media challenge the conventional technologies (such as telephone and fax) and become itself a potential threat to civilization. According to Davies, the perpetrators of September 11 attacks had been using the email to communicate with each other (Davies 2002). This increase use of the Internet is a consequence of its potential for anonymity when communicating in the worldwide (Ramasastry 2002). In fact the Web became a powerful weapon used by terrorists, not only to establish communication and to coordinate operations, but also to reach the public and get new recruits. According to Hsinchun Chen, director of the University of Arizona's Artificial Intelligence Lab, many Web pages are designed and maintained by recognized terrorist groups, including Al-Qaeda, the Iraqi rebels and many terrorist cells in Europe, to spread terrorism philosophies and to attract followers (Kotler 2007). "Around the year 2000, there were 70 to 80 core terrorist sites online; now there are at

[1] http://www.un.org/unifeed/script.asp?scriptId=73

least 7000 to 8000" (Kotler 2007). This fact alert governments and underlies the need to perform extensive surveillance on the Web.

Roger Clarke defines Data Surveillance “as systematic use of personal data system in the investigation or monitoring of actions or communications of one or more persons” (Clark 1997). One (traditional) surveillance approach relies on human capabilities to analyse and capture data. Besides the analyses of databases, teams of several participants with special language skills (such as Arab), participate in forums or online chat rooms with the purpose to track the presence of terrorists on the Web and identify possible threats. Recent observations show that terrorists do not exist only in Islamic world. Actually most of them belong to a second generation that lives in Occidental countries. In fact, the Web has been increasingly used to exchange messages in other languages besides Arabic, such as English, Spanish, French and even in Chinese, underlying the need of automated mechanisms to analyse messages posted in different languages (Rotstein 2007).

In this context, governments and industry began efforts to use technological means to collect and integrate data from distributed and heterogeneous information sources. The Department of Defence of the USA established the use of data mining technology, in addition to sophisticated and advanced statistical techniques, to find important patterns in massive databases, in order to be able to anticipate and prevent terrorist attacks (Garfinkel et al 2006). The Information Awareness Office (IAO) at Defence Advanced Research Projects Agency (DARPA) is one of the sponsored of the Total (Terrorism) Information Awareness (TIA) project. This project aimed to analyse information and detect potential terrorists. Massive databases were created to store an extensive volume of information collected from distributed information sources to be integrated and compared. On this data extensive analyses are performed in order to extract information through the use of data mining techniques. However, the efficiency of these databases and data mining technology used to foreseeing terrorism activities as not been proven in the academic literature (Maxwell 2005). Moreover the risks to privacy and other civil-liberties concerns several communities and raises important issues, as the likelihood of the false positives resulted from casual information associations, and meaningless positives (Anderson 2007). This problem result from the fact that it is very difficult to incorporate knowledge or concepts abstracted from the low level data, into statistical analyses. Furthermore, if the investigations of these positives are not done correctly, innocent people can be damaged.

One promising solution and which encourage this research, is the use of ontology based applications in the knowledge management. In the next section it will be presented an overview of the ontologies’ foundations, followed by a reflection concerning the improvements ontologies might bring to the data surveillance domain.

3. Ontology

In the recent years several studies concerning ontologies have enlarged their application scope from the theory of nature of human being, science and philosophy areas to other domains, mainly with the objective to support the sharing and reuse of the formally represented knowledge (Mika 2002). Furthermore, in logic programming area, ontologies are defined for two main functions: (1) “Provide a way of viewing the world, and hence for organising information”; (2) “The ontologies are required for interoperability, to define a shared vocabulary and meanings for terms with respect to other terms” (Miller 2000).

According to Gruber, an ontology is a formal explicit specification of a shared conceptualization (Gruber 1993). This means sharing common understanding by expertises from a specific domain, which is usually performed as a set of concepts, relations, functions, axioms and instances. Fensel express the Gruber’s definition accordingly to four basic components: “conceptualization”, concerning the abstract model of a phenomenon in the world; “formalization”, meaning it should be machine readable; “explicit” because the concepts used in a specific domain must be explicitly defined; and finally, it is “shared oriented” according to an agreement establish between the developers and users of ontologies (Dieter 2000). In brief, an ontology constitute a set of well defined concepts describing a specific domain. The concepts are defined using a subclass hierarchy, by assigning and defining properties and by defining relationships between the concepts (Rees 2003).

In this context the study of ontologies, today, completely fit the study of modern computer science and information technologies and is considered a promising solution to the Semantic Web and Knowledge Management areas (Gaitanou 2007). In fact it is noticed a continued growth of the information volume

and the emergence of sophisticated information and communication technologies. These facts have contributed to the recognition of ontologies' richness to express semantics in several areas and led to a new focus on information content management and organization. In fact the use of ontology based applications in the knowledge management as been pointed as one of the most promising ways to deal with the continued growth of the information volume available in the Web, which makes it difficult to find, organize, access and maintain information (Masuoka 2003). Standardized ontologies have been developed in many and different domains in order to enable researchers and worldwide user to share and annotate specific information.

Many organizations are now offering semantic enhancement tools to index content and map them to entities within the specified ontology. In fact the users of the *World Wide Web* think about content based on concepts that are embedded in the information resource (what the resource is about). In addition, if the information resources are organized on concepts, they can be related to each other, based on those concepts. This is the most interesting feature of an ontological approach, once it enables information resources to be related to a user profile and precisely delivered to specific user interest. This is the major challenge and the research issue regarding the use of an ontology based approach in the data surveillance domain.

4. Knowledge based Surveillance

The World Wide Web Consortium (W3C) is developing efforts in a language to encode knowledge on the Web pages, in order to make it understandable to electronic agents searching for information - this language is called Resource Description Framework (RDF). The DARPA and W3C are working together in the development of a DARPA Agent Markup Language (DAML) by extending RDF with more expressive structure in order to promote agent interaction on the Web (Noy et al 2001). In several areas researchers are now trying to develop standardized ontologies towards a common objective: to share and annotate information in their knowledge fields. Some relevant examples are presented in the area of the medicine. In this domain standardized and large structured vocabularies have been developed, such as SNOMED (http://www.snomed.org/) and the semantic network of the United Medical Language System (UMLS – http://www.nlm.nih.gov/research/umls/).

In public health domain several systems have been developed in order to detect disease-outbreak patterns and also with administrative and business purposes. One example is the billing and pharmaceutical sales records, collected for inventory and marketing purposes. Other example is the Realtime Outbreak and Disease Surveillance (RODS) project developed in the University of Pittsburgh to detect earlier outbreak of a disease.

It is recognised the increase use of ontology based applications in the knowledge management of data surveillance, in particularly in bioterrorism surveillance, in order to early detect and characterize an epidemic threat resulting from bioterrorism act. According to Buckeridge an effectiveness intervention depends on how quickly an epidemic can be detected, how well can be characterized and how rapidly a response is initiated (Buckeridge et al 2002). An experimental system named BioSTORM (Biological Spatio-Temporal Outbreak Reasoning Module) is a knowledge based framework for real time epidemic surveillance (Buckeridge et al 2002). In fact, the use of ontologies to model and annotate information and knowledge involved in syndrome and epidemic surveillance is the main feature of the BioSTORM system approach.

As stated before, the data surveillance approach currently used to detect suspicious terrorist activities is performed through the use of sophisticated data mining technologies with advanced statistical techniques to analyse large amounts of data (Popp 2006). However, most of the algorithms used were developed for completely different contexts. Predicting future terrorist attacks can not be compared to the use of massive databases to indicate the future consumer behaviour or organize supply chain management. In fact the general consumers do not usually hide or mask their transactions and those who are concerned pay their products with cash (Maxwell 2005). Similarly, suppliers are definitively interested to cooperate with the retailers, providing accurate information in order to have greater sales. On the other hand, terrorists will actively try to hide their identities and actions, making more difficult the detection. In fact there is no evidence that data repositories and data mining systems have the capacity to identify and detect terrorisms activities (Maxwell 2005). Consequently several questions have been rising about the efficiency of this systems because of the risks of misidentify innocent individuals, and the consequence of these results.

The information collected by the data surveillance systems are characterized as heterogeneous due to the fact that is usually provided by diverse and distributed data sources such as databases and files without semantic description or syntactic structure. To enable analyses throughout these data sources, knowledge about how to characterize and combine different sources and types of data must be specified precisely. This general knowledge is properly modelled by ontologies.

Ontologies consist on specifications of concepts, properties and relationships for describing a domain of expertise. They provide structured and queryable frameworks for modelling the semantics of knowledge data and encoding them, in spite of the way they are internally represented (Crubezy et al 2003). Ontologies enable the description of features, types, and relationships of data provided from different sources (Pincus et al 2003). Finally, ontologies have been widely used in the computer science community to achieve semantic integration of distributed data sources.

In the context of data surveillance the use of ontologies to model and annotate information and knowledge, represent an outstanding challenge to be followed in the surveillance of terrorism activities. The combination of distributed data sources is crucial to perform efficient data analysis and consequently to detect and characterize suspicious activities. This way, the interoperability between different Information systems appears as a pre-requisite, which can be satisfied by an ontology based approach. Knowledge based mechanisms seems to be an appropriate strategy in order to enable a structured organization of information. This will enable to perform a better data analyse and thus a quickly detection and characterization of suspicious activities.
The definition of an ontology in the domain of data surveillance consist in providing a concrete specification of term names and term meanings, which includes description of concepts and their relationships. A complete description of the specific concepts in the domain of data surveillance will promote to catalogue and to distinguish various types of data objects and their relationships, and consequently contribute to a rapid analyse of data rather than the traditional means currently used.

In practical terms the ontology development will include: (1) the definition of the data source features; (2) analyse of a data set of terrorism surveillance activity in order to select the key elements in the data surveillance domain; (3) definition of taxonomy of the data attributes that describes the data context and the relationship with related data; (4) metadata description of the data elements, accordingly to standard and widely used vocabularies. The metadata description will enable the contextualization of the many different types of data involved in the data surveillance process. For these tasks it is necessary to analyse and determine the standard vocabulary that best describe features of the data source used. The definition of the metadata structure must include a variety of metadata elements that conforms the necessary descriptions requirements of the information resources used.

In fact the General Accounting Office (GAO) recommends the establishment of common metadata standards for electronic information as a strategy to integrate and manage homeland security functions, including new procedure for data sharing across governmental (Maxwell 2006).

An ontology based approach in the domain of data surveillance is just a proposal solution and needs further studies. However, the novelty of this solution regards the use of an ontology to enhance the abstract metadata rich view on data semantics resources. And further, this metadata will support the integration of heterogeneous data collected in the data surveillance process, enabling a uniform analytic and interpretation process to the data resource.

This is a new and an ambitious proposition in the data surveillance domain in order to improve the current mechanisms used. The view presented in this paper is consistent with the vision of the Tim Berners-Lee – the creator of the Web - that now considers ontologies to be a critical part of his latest work on the Semantic Web, envisioning the Semantic Web as being machine processable, leading to better understanding of the content of the Web pages by the machines (Berners-Lee 2000).

5. Conclusions

In this paper we proposed an ontology based approach to data surveillance as a way to improve the current data surveillance process. It was also presented the shortcomings of the data surveillance mechanisms currently used. Data mining technologies and statistical techniques have been widely

used to perform data analysis collected from distributed information sources. However the efficiency of these techniques to combat terrorism has been highly questionable, due to the fact that can lead to false positives that can be generated and bringing potential negative side-effects, for instance, leading innocent citizens to the confrontation of law enforcement services. In fact data mining systems have already proved its efficiency in the information extraction of data warehouses in business and commercial domains. However, the use of data mining systems for national security needs to be evaluated not only against the citizen privacy being subject of abuse, but also the likelihood of goal success.

On the other hand, ontologies have moved beyond the domains of philosophy to knowledge representation with unquestionable proves. Analyst research companies use ontologies to support browsing and search for e-commerce and to support interoperability in order to facilitate the knowledge management and configuration (McGuinness 2002). Concerning the bioterrorism, the recent anthrax attacks forced new surveillance developments. In fact the BioSTORM project already follows ontology based approaches in order detect and characterize early epidemic threats and syndromes (Buckeridge et al 2002).

The development of an ontology approach to integrate and process data surveillance will provide the principles foundation for implementation and evaluation of mechanisms to be used in data surveillance process. In fact the analysis and interpretation of information collected in data surveillance process is the focus of all process. A data surveillance system needs an ontology infrastructure to be used by different analytic strategies defined accordingly to the different incoming surveillance data. The use of ontologies in the domain of data surveillance will help in the specification and configuration of information structure collected from the Web, introducing an ambitious challenge to the conventional surveillance mechanisms currently used.

References

Anderson, Shannon R., (2007). Total Information Awareness and Beyond. The Dangers of Using Data Mining Technology to Prevent Terrorism, [on-line], http://www.bordc.org/threats/data-mining.pdf.

Berners-Lee, Tim, (2000) “Semantic Web on XML”, Keynote presentation for XML 2000. Slides available at: http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide1-0.html. Reporting available at: http://www.xml.com/pub/a/2000/12/xml2000/timbl.html

Buckeridge, David L., Graham, Justin, O’Connor, Martin J., Choy, Michael K., Tu, Samson W., Musen, Mark A., 2002. Knowledge-Based Bioterrorism Surveillance. *In AMIA Annual Symposium*, San Antonio, TX. [on-line], http://bmir.stanford.edu/file_asset/index.php/1147/SMI-2002-0946.pdf.

Clarke, Roger, 1997. *Data Surveillance: Theory, Practice & Policy*. Thesis (PhD). National Australian University. [on-line], http://www.anu.edu.au/people/Roger.Clarke/DV/PhD.html

Crubézy, M. and Musen, M.A. (2003). *Ontologies in support of problem solving*. Handbook on Ontologies. S. Staab and R. Studer, Springer-Verlag: 321-341.

Davies, Simon, (2002) *A Year After 9/11: Where Are We Now?*
The dissention over technology has migrated from the East to the West over the last 12 months. *Communications of ACM*, vol. 45 nº 9.

Dieter, F. (2003). *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. Springer-Verlag, New York, Inc., Secaucus, NJ, 2003.
Evelien, Brouwer, (2005) “Data surveillance and border control in the EU: Balacing efficiency and legal protection of third country nationals”, [on-line]
http://www.libertysecurity.org/article289.html?var_recherche=Data%20Surveillance

Gaitanou, Panorea, 2007. Ontology Semantics and Applications. *In Proceedings of the 2nd International Conference on Metadata and Semantics Research* (CD-ROM), Ionian Academy, Corfu, Greece

Garfinkel, Simon, Smith, Michael D., (2006) “Data Surveillance”, *IEEE Security & Privacy*, vol 4, nº 6, November/December 2006, pp. 15-17, [on-line] http://www.computer.org/portal/site/security/menuitem.6f7b2414551cb84651286b108bcd45f3/index.jsp?&pName=security_level1_article&TheCat=1015&path=security/2006/v4n6&file=gei.xml&;jsessionid=FspyzWtyVhrBzqhC5LqmZrmdQJnKJll96L5hhZq99fhlbQwY8FNB!1287236228

Gruber, Tom (1993). A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199-220, 1993. [on-line],http http://tomgruber.org/writing/ontolingua-kaj-1993.pdf.

Kotler, Steven. (2007). “'Dark Web' Project Takes On Cyber-Terrorism”, [on-line], FoxNews.com, http://www.foxnews.com/story/0,2933,300956,00.html

Masuoka, R., Labrou, Y., Parsia, B., Sirin, E. (2003). *Ontology-enabled pervasive computing applications*. IEEE Intelligent Systems. 18(5) pp.68-72.

Maxwell, Terrence A., (2005) “Information Policy, Data Mining, and National Security: False Positives and Unidentified Negatives”, In: Proceedings of the 38th Hawaii International Conference on System Sciences. Hawaii.

McGuiness, Deborah L. (2002) *Ontologies Come of Age*. In Dieter Fensel, Jim Hendler, Henry Lieberman, and Wolfgang Wahlster, editors. Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. MIT Press [on-line], http://www.ksl.stanford.edu/people/dlm/papers/ontologies-come-of-age-abstract.html

Miller, Libby, (2000) “Ontologies and Metadata”. A Draft Discussion of issues raised by the Semantic Web Technologies Workshop, 22-23 November 2000 [on-line], http://ilrt.org/discovery/2000/11/lux/

Mika, Péter, 2002. Applied Ontology-based Knowledge Management: A Report on the State-of-the-Art. Thesis (Master). Vrije Universiteit: Amsterdam. [on-line], http://www.cs.vu.nl/~pmika/thesis/pmika-thesis-full.doc

Natalya F. Noy and Deborah L. McGuinness, (2001). ``Ontology Development 101: A Guide to Creating Your First Ontology''. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001. [on-line], http://www-ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html

Pincus Z. and Musen, M.A. (2003). Contextualizing heterogeneous data for integration and inference. In: Proceedings of the American Medical Informatics Association Annual Symposium, Washington, D.C.: 514-8.

Popp, Robert and Poindexter, John, (2006). “Countering Terrorism through Information and Privacy Protection Technologies”, *IEEE Security & Privacy*, vol. 4, no. 6, November/December 2006, pp. 18-27, [on-line], http://www.computer.org/portal/site/security/menuitem.6f7b2414551cb84651286b108bcd45f3/index.jsp?&pName=security_level1_article&path=security/2006/v4n6&file=popp.xml&xsl=article.xsl&

Ramasastry, A.(2002). “Why we should be concerned about “Total Information Awareness” and other anti-terrorism strategies for the Internet”, [on-line], FindLaw, Legal News and Commentary, http://writ.news.findlaw.com/ramasastry/20021231.html

Rotstein, Arthur H. (2007). “Tracing Terrorist’s Social Web”, [on-line], Technology News, http://www.technewsworld.com/story/60294.html

Rees R. (2003). Clarity in the usage of the terms ontology, taxonomy and classification, Amor R. (editor) Proceedings of the CIB W78's 20th International Conference on Construction IT, Construction IT Bridging the Distance, CIB Report 284, ISBN 0-908689-71-3, Waiheke Island, New Zealand, 23-25 April 2003, pg. 432-440. http://itc.scix.net/cgi-bin/works/Show?w78-2003-432