Ontology-based Information Content Security Analysis

Pan Yan
School of Management and Economics Beijing Institute of Technology, Beijing 100081, P.R.China
yanpan0327@bit.edu.cn;

Yanping Zhao
School of Management and Economics Beijing Institute of Technology, Beijing 100081, P.R.China
zhaoyp@bit.edu.cn

Cao Sanxing
Information Engineering School, Communication University of China, Beijing 100024, P.R.China
c3x@cuc.edu.cn

Abstract

This paper proposes an ontology-based information content security analysis framework, which adopts artificial intelligent system designing theory and ontology engineering method to improve information content security surveillance. This framework introduces ICSO (information content security ontology), by which we incorporate disparate and heterogeneous data sources into surveillance system , and an novel reasoning subsystem to utilize ICSO knowledge to support information content security surveillance in interpreting surveillance data and improving surveillance decision making. The ICSO is not only a knowledge base but also a SNA (social network analysis) model, thus we could easily implement relative SNA methods to focuses on individual's computer mediated communication (CMC) networks to effectively spot harmful characters. And we test our system on Enron Email Dataset and web pages, the result shows this framework is an efficient and effective solution in information content security and can be spread to other knowledge-based systems.

1. Introduction

Internet information content security is a critical factor to the security of political, economic, cultural, military and many other areas. At present, the major technique applied in information content security is the information filtration technology and social network analysis (SNA) technology. In the area of information filtration technology, including text, audio, picture and video filtration, the mainly used algorithm is known as multi-pattern (multiple keywords) matching algorithm, such as Aho-Corasick (AC) algorithm, AC-BM algorithm, Wu-Manber (WM) algorithm and BC algorithm [1]. In the area of SNA, many researches are carried out on the social networks focusing on the computer mediated communication (CMC), including analysis of social behavioral factors pertaining to online communities and analysis of information dissemination pattern of groups, such as affect Intensity analysis of U.S. and Middle Eastern extremist group forum postings [2] and a new model for evaluating similarities of Blogs, which can be used for isolating and tracking like-minded networks for surveillance [3]. These studies on social networks and their information dissemination may help us understand individual and association's communication pattern and gain underlying information behind rough data.

As the increase of information content security methodologies, how can we utilize each researcher's method and make these methods work cooperatively becomes a problem to researchers and security authority. On the other hand, the tremendous data from disparate and heterogeneous online content sources presents both opportunities and challenges for surveillance. The problem is how we can integrate disparate and heterogeneous online content sources into surveillance–system to discover unknown connections between different sources and make efficient use of extracted knowledge. To answer the questions we develop an ontology-based information content security analysis framework. This framework is a knowledge-based intelligent system framework that integrates disparate and heterogeneous data sources and deploys various statistical, knowledge-based and SNA method to support information content security surveillance in interpreting data and improving surveillance decision making.

The paper is organized as follows: section 2 introduces the overall skeleton of our framework; Section 3 introduces the main component: Part 3.1, the information content security ontology schema; Part 3.2, the ontology building; Part 3.3, the ontology integration; Part 3.4, the reasoning subsystem; Section 4 describes our test on Enron email dataset and relative web pages.

2. Ontology-based information content security analysis framework

978-0-7695-3305-6/08 $25.00 © 2008 IEEE
DOI 10.1109/FSKD.2008.364

Our ontology-based information content security analysis framework is an intelligent system that extracts, organizes, and normalizes valuable, meaningful information and contextual knowledge from many heterogeneous content sources and utilizes various statistical, knowledge-based and SNA analysis methods to support knowledge discovery and decision making for security authority and researchers. The essence of our approach is the use of ontologies to model security information and knowledge. Ontology is the term referring to the shared understanding of some domains of interest, which is often conceived as a set of classes (concepts), relations, functions, axioms and instances (Gruber, 1993) [4]. In AI (Artificial Intelligence) community Ontology is knowledge representation method, which encodes knowledge into well-defined formats that can be either understood by man or processed by computer and allows knowledge shareable and reusable. In recent years many intelligent system based on ontology has been built, such as PMVE (politically motivated violent events) [5] and BioSTORM (the Biological Spatio-Temporal Outbreak Reasoning Module) [6]. We used ontology to model information and knowledge of relative surveillance data and social networks of people, as well as to describe and implement rule-based knowledge reasoning.

The use of ontology provides a shareable, reusable, scalable knowledge repository to our system, and models various analysis methods proved to be efficient and effective to information content security surveillance. So the comprehensive analysis to disparate and heterogeneous surveillance sources becomes practical.

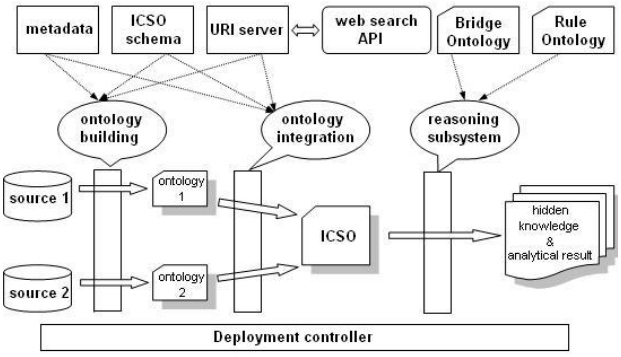

Figure 1. System architecture

The system architecture is shown in Figure 1: Data from disparate and heterogeneous sources pass through ontology building process to construct ontologies respectively. Ontologies built from different sources are integrated into ICSO in ontology integration process. The URI server, metadata, and ICSO schema are referred in both ontology building and ontology integration. The reasoning subsystem then reasons by using rule-ontology to configure analytical methods in logical sequence to make comprehensive analysis.

3. Main component

3.1. The Information content security ontology – ICSO schema

We develop ICSO by a middle-out approach, in which the most important concepts are identified first and then generalized and specialized into other concepts. Then we develop ICSO schema in Protégé [7], a mature methodology and software tool for building ontology, in OWL-DL [8] language by defining class and property.

The main Information content security classes are shown in Figure 2.

Figure 2. Main Information content security classes

We identify harmful character through two main sources: computer mediated communication (CMC), including emails, Blogs and so on, and online News. We collect relations and information from these sources and complete ICSO by ontology building and integration.

3.2. The ontology building

Surveillance data including news, blogs, BBS (Bulletin Board System), emails and online databases are different

from formats. We use information extraction method to convert semi-structured and unstructured sources to structured sources. To most blogs, BBS, emails which are presented on web pages, many sophisticated methods can be used to extract useful and valuable information from them. We use a customized information extraction tool to extract information from online sources and store extracted information in XML format.

In the process of ontology building, instances of ICSO classes are built according to extracted information and relation. We introduce URI (Uniform Resource Identifier) to guarantee the compatibility on the instances. Each entity is endowed with a unique URI. ICSO schema and metadata are generally referred in order to build datatype properties and object properties which are predefined. Two main steps in ontology building are class instantiation and property build-up.

3.2.1. Class instantiation. To build instance of ontology, we need to identify entity through data values provided by structured data sources and metadata. However, homograph and allomorph of character strings indicating entities within metadata brings ambiguity to entity identifying. To solve this problem, we introduce a web search API: The ZoomInfo API, which provides access to ZoomInfo's (http://www.zoominfo.com/) search capabilities and data for over 36 million people and 3.5 million companies. We remove ambiguity of entity resolution by comparing character strings being instantiated to the search results from web search API. In our test on Enron Email Datasets, in which ambiguity always occurs on employees and companies, this method basically satisfies our need in entity resolution.

We follow the instructions bellow in class instantiation process on URI server:

1. Group entities by name, class and attribute (such as human's gender and birthday and so on); entities with the same name, class and attribute are in the same name group.

2. If a name group only has one entity, assign a URI to the entity.

3. If a name group has 2 more entities, search the web search API and compare the returned entity with the entities in the name group:

3.1 If there is 0 return entity, we can't tell if the 2 more entities are the same, then we temporally consider entities in this name group are all different and assign them different URIs for checking later;

3.2 If there are some returned entities, we compare these returned entities with the 2 more entities, if one is the same as some of the 2 more ones in the name group, assign the same ones 1 URI, and assign the rest no match entities in name group different URIs.

3.2.2. Object and datatype property build-up. The extracted XML documents contain information and relation, which are resources in building instances' object and datatype properties. ICSO schema and metadata are referred in the building process. Because the entity resolution on URI server guarantees that each entity has a URI, the object property build-up becomes quite similar to datatype property build-up: the object property build-up assigns URI to the object of the statement (a statement is a RDF-Triple that includes a subject, a predicate and a object) that added to the ontology instance; the datatype property build-up assigns data value and data type to the object of the statement that added to the ontology instance.

The metadata repository store tables guiding the building of ontology instance. An example table showing corresponding relationships between XML tags and OWL terms is in Figure 3.

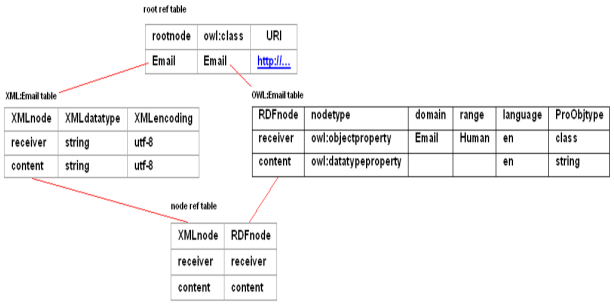

Figure 3. An example of relationships between XML tags and OWL terms

3.3. The ontology integration

Different researchers has different understandings to ontology integration, here we consider ontology integration as integrating separate ontologies into a global or reference ontology. In Fig.1, ontologies built from different sources are integrated into ICSO to provide an overview of all sources and enrich semantic networks of ICSO. ICSO schema is referenced in all ontology building process, and ontology integration is carried out in instance level. The main issue of integration lies on the lack of information to recognize a unique instance from different ontologies, and integration rules changed by the correlation between sources and reliability of sources. For example, in two highly related sources, such as Enron email dataset and Enron email research web pages, "Sally Beck" extracted from both sources denotes the chief operating officer of Enron, however, "Sally Beck"

extracted from online news can't be considered as the chief operating officer of Enron just by name. So we just import instances in ICSO and consider all instances different, then in the reasoning subsystem we set rules to identify equivalent individual based on the correlation between sources to complete the ontology integration.

3.4. The reasoning subsystem

In modern intelligent systems, the domain knowledge is separate with the reasoning module for the reuse of domain knowledge and the extensibility of inference rules. We follow this principle in designing our system. The reasoning subsystem reasons on the domain knowledge represented by ICSO and make conclusions by inference rules. To enhance the ability of inference rules, we wrap analytic methods, including traditional statistical techniques operating on low-level data, SNA method in order to discover social knowledge and knowledge-based approaches capable of reasoning about semantic knowledge, into inference rules. We categorize our rules by resources they operate on, and the action they perform and implement rules and methods as software routines to conform to our surveillance requirements. To make such rule structure explicit and practical, we model rules in rule-ontology which referenced in system procedures. With rule-ontology, rules become processable and the reasoning subsystem becomes independent of the domain knowledge. Rule-ontology facilitates system modification to enhance rule designing, method reuse and helps security analyzer understand the system.

The inference rules has the typical IF (condition) THEN (action) structure. When the domain knowledge represents by ICSO satisfy the IF (condition) part of a rule, the THEN (action) part of the rule is executed. Both the IF (condition) part and the THEN (action) part can invoke various methods. Those methods within their inputs and outputs format are all predefined in our system libraries.

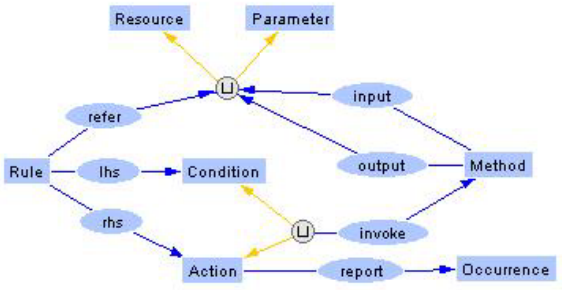

Figure 4. The rule-ontology schema

Rule-ontology is to describe rules to end users and to be referred to organize system procedures as the logical sequence it shows.

Figure 4 describes the classes in the rule-ontology: Rule identifies rules in rule-ontology; Condition represents the IF part of rules; Action represents the THEN part of rules; Method represents every analytical method in system library; Resources identifies instances and properties that analytical methods operate on; Parameter represents the data value and data type that analytical methods operate on, and it has subclasses such as String, Integer and so on; Occurrence identifies modification of input resources. They are used to provide continuous and rational reasoning.

Figure 5 is an example of rule instance:

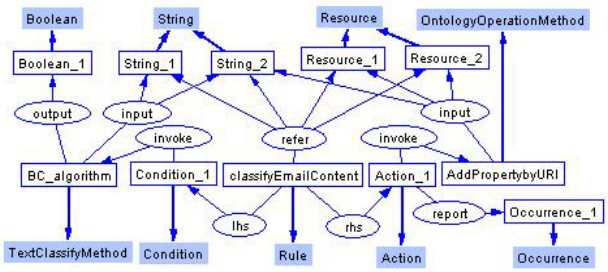

Figure 5. An example of rule-ontology

classifyEmailContent, the IF part of the rule invokes the BC (Byte-Coding) algorithm [1] method, which is used to make fast semantic filtration to text, to test if String_1, which indicates the test text, is classified into String_2, which indicates a category of theme; String and Boolean are subclass of Parameter indicating the data type of parameters; the THEN part of the rule invokes the AddPropertyByURI method, which is a ontology operation method, to add Resources_2 (a datatype property) to Resource_1(an instance), with the data value String_2.

Rules are not processable without assigning values to Resource and Parameter, because the rules are not only designed for reasoning on ICSO, but also designed for reasoning on other ontology repositories; this structure allows our system adopt other information security ontologies as our knowledge repository and makes our reasoning rules flexible and reusable. When we assign URI to Resource and data value to Parameter, a rule is complete.

For example, to complete the rule classifyEmail-Content. We assign every Email instance's URI to Resource_1 and value of datatype property "content" to String_1, and assigns datatype property "theme" to Resource_2, while String_2 is assigned a text value "financial" . Then the rule is complete as: For every

instance of Email E, IF the content of E is classified into theme "financial", THEN add the datatype property "theme" with the value "financial" to E.

4. The test

Based on the above work, the information content security ontology repository which include all essential factors we need in analysis: when, where, who, what and how security event happens. We choose Enron scandal for experiment; establish the ontology repository from online sources including Enron Email Dataset and web pages such as http://enron.trampolinesystems.com.

We test ICSO repository by the example rule and rule visualizeFinancialNetwork. The rule visualizeFinancial-Network includes SNA and visualization technologies, which are in order to presents the potential community and the core member of the specific scope [9]. The visualization result of financial networks is shown in Figure 6.

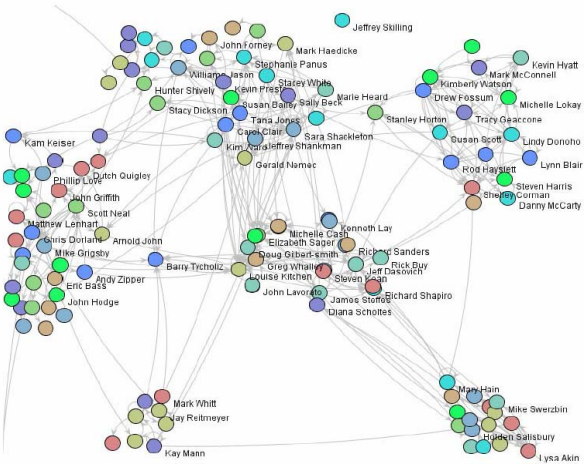

Figure 6. The visualization of financial networks

5. Conclusion

We carry out knowledge extraction and organization from rough data on the web pages and databases, and design the structure of ICSO, which makes information content security knowledge normalization and reusable, the updating of knowledge automatic, the designing of rules in purpose of security inference flexible, the analysis on detecting relations of individuals and non-obvious knowledge visible, and provides information push to the information content security analyzer. We test our system on Enron Email dataset and correlative web pages, and the result basically matches our expectation. The method we present in constructing knowledge repository and reasoning subsystem can be easily applied to other ontology-based research. In the future research, we are aiming at the following goals: first, enhance the ability of natural language understanding of our system to gain more information through computer mediated communication materials; second, enhance the ability of knowledge inference and visualization .

6. Acknowledgment

This research is supported by the National Science Foundation of China, the project code: 70471064, and Defence Research Foundation of National Innovation Base in Philosophy and Social Science of 985 II, the project No. 107008200400024.

7. References

[1] Zhao, Yanping; Lu, Wei; An Efficient Algorithm for Content Security Filtering Based on Double-Byte, Intelligence and Security Informatics, 2007 IEEE 23-24 May 2007 , pp. 300 – 307

[2] Abbasi, Ahmed; Chen, Hsinchun; Affect Intensity Analysis of Dark Web Forums Intelligence and Security Informatics, 2007 IEEE 23-24 May 2007 , pp. 282 – 288

[3] Choi, Hwan-Joon; Krishnamoorthy, Mukkai S.; Categorization of Blogs through Similarity Analysis Intelligence and Security Informatics, 2007 IEEE 23-24 May 2007 , pp. 160 – 165

[4] T.R. Gruber, Atranslation approach to portable ontology specification, Knowledge Acquisition 5 (1993) , pp. 199–220.

[5] Wennerberg, P. Oezden; Tanev, H.; Piskorski, J.; Best, C.; Ontology Based Analysis of Violent Events Intelligence and Security Informatics, 2007 IEEE 23-24 May 2007 , pp. 373 – 373

[6] Crubezy, M.; O'Connor, M.; Pincus, Z.; Musen, M.A.; Buckeridge, D.L.; Ontology-centered syndromic surveillance for bioterrorism Intelligent Systems, IEEE Volume 20, Issue 5, Sept.-Oct. 2005 pp:26 – 35

[7] Protégé, Available: http://protege.stanford.edu/

[8] OWL Web Ontology Language Overview, Available: http://www.w3.org/TR/owl-features/

[9] Zhao, Yanping; Feng, Lei; Chen, Lei;Detection of Multi-Relations Based on Semantic Communities Behaviors ,Service Systems and Service Management, 2007 International Conference on 9-11 June 2007 , pp. 1 – 7.