Knowledge Sharing and Reuse in Digital Forensics

Jingshan Huang[1] and Alec Yasinsac
School of Computer and Information Sciences
University of South Alabama
Mobile, AL, U.S.A.
{*huang, yasinsac*}*@usouthal.edu*

Patrick J. Hayes
Florida Institute for Human and Machine Cognition
Pensacola, FL, U.S.A.
phayes@ihmc.us

Abstract

Digital investigation involves examining large volumes of data from heterogeneous sources. We offer a framework for facilitating examination and synthesis of this mountain of data using ontology matching and machine learning technology.

Keywords

digital forensics; knowledge sharing; ontology matching; heterogeneous semantics; and semantic integration

I. Introduction and Research Motivation

The rapid emergence of new techniques in Computer Science and Information Technology provides potential innovation for digital investigations, though substantial challenges remain. Briefly speaking, the major concerns have concentrated on the challenge to maintain the integrity of evidence found by different parties (usually from distributed geographic areas, or even with cultural barriers), the accurate interpretation of evidence, and the trustworthy conclusion drawn thereafter. The state-of-the-art efforts have resulted in various ad-hoc and/or proprietary formats for storing the contents of evidence, analyzing these contents, and maintaining metadata for evidence. Different parties are likely to adopt different formats according to their specific needs. Therefore, the seamless communication among different parties, along with the knowledge sharing and reuse that follow, become a non-trivial problem.

Traditional manual approaches involve a lot of human intervention, which is tedious, time-consuming, and error-prone. Researchers have proposed to adopt ontologies for digital investigations. An example is, the Digital Forensics Research Workshop (DFRWS) formed a working group [4] in 2005 with a goal of defining a standardized Common Digital Evidence Storage Format (CDESF). This group aimed to define an open data format that can store both digital evidence and related metadata. For example, the CDESF could contain a bit-wise image of a hard disk as well as the location from where the image was made, a digital photograph of the hard disk, the name of the person who made the image, and the case number. A different instance of the CDESF could contain a contraband file along with the unique identifier of the hard disk image from which it was extracted, the name of the investigator, and its original file name path. Unfortunately, despite of the importance of this topic to Digital Forensics, the CDESF working group was disbanded in August 2007.

While the CDESF attempted to design a common data format for digital investigations, the forms that they introduced are more suited to be represented as independently designed ontology structures [9] with bridges between heterogeneous constructions.

Since digital investigation deals with massive conceptual complexity in multiple layers of abstraction. Therefore, *there is no such central ontology that is large enough to include all concepts of interest to every individual criminal investigator*. Each need for a conceptual model from any individual party will have to provide its own particular extensions, which will be different from and likely incompatible with the extensions added by other parties. The intuition behind this phenomenon is: ontologies are a formal conceptualization of part of the world; considering the fact that anyone can design ontologies according to his/her own conceptual view of some domain, ontological heterogeneity among different parties becomes an inherent characteristic. Therefore, we believe that an agreed-upon, global, and general-purpose ontology is not a feasible solution. Instead, different groups should maintain their own conceptual models, but will utilize ontology-matching technology to synthesize their data with others' models, effectively decoupling the evidence semantics from its logical description and organization.

We propose a systematic mechanism, Digital Investigation Evidence Acquisition Model Based on Ontology Matching (DIEAOM), to facilitate knowledge collection from disparate, heterogeneous evidence sources, knowledge sharing and reuse, and decision support for criminal investigators.

[1] Corresponding author Tel./Fax: 1-251-460-7612/1-251-460-7274.

II. Background

A. Background Knowledge

Ontology is a computational model of some portion or domain of the world. The model describes the semantics of the terms used in the domain. Ontology is often captured in some form of a semantic network, i.e., a graph whose nodes are concepts or individual objects and whose arcs represent relationships or associations among the concepts. The semantic network is augmented by properties and attributes, constraints, functions, and rules, which govern the behavior of the concepts. In brief, an ontology consists of a finite set of concepts, along with these concepts' properties and relationships. In addition, most real-world ontologies have very few or no instances, i.e., they only have the aforementioned graphical structure (a.k.a. "schema").

Ontology Heterogeneity is an inherent characteristic of ontologies developed by different parties for the same (or similar) domains. The heterogeneous semantics may occur in two ways. (1) Different ontologies could use different terminologies to describe the same conceptual model. That is, different terms could be used for the same concept, or an identical term could be adopted for different concepts. (2) Even if two ontologies use the same name for a concept, the associated properties and the relationships with other concepts are most likely to be different.

Ontology Matching is also known as "Ontology Alignment," or "Ontology Mapping." It is the process of determining correspondences between concepts from heterogeneous ontologies (often designed by distributed parties). Such correspondences include many relationships, for example, *equivalentWith*, *subClassOf*, *superClassOf*, and *siblings*.

Machine Learning is a scientific discipline that is concerned with the design and development of algorithms that allow computers to change behavior based on available data (a.k.a. "training data"). A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data.

Artificial Neural Network (ANN) is a mathematical or computational model that tries to simulate the structure and/or functional aspects of biological neural networks. Consisting of an interconnected group of artificial neurons, an ANN processes information via a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Generally speaking, ANNs are non-linear statistical data modeling tools, and can be used to model complex relationships between inputs and outputs, or to find patterns in data.

Agglomerative Clustering is a special category of clustering, which is the assignment of a set of observations into subsets (i.e., clusters) so that observations in the same cluster are similar in some sense. Agglomerative clustering belongs to the hierarchical clustering that creates a hierarchy of clusters represented in a tree structure. Different from divisive clustering, agglomerative clustering starts at the leaves and successively merges clusters together.

B. Related Work in Digital Forensics

S. Raghavan, A. Clark, and G. Mohay [21] motivate the need to formalize the process of analyzing digital evidence from multiple sources simultaneously. The authors present the forensic integration architecture (FIA) that provides a framework for abstracting the evidence source and storage format information from digital evidence. The FIA architecture identifies evidence information from multiple sources that enables an investigator to build theories to reconstruct the past. FIA is hierarchically composed of multiple layers and adopts a technology independent approach.

R. Gong, K. Chan, and M. Gaertner [11] propose a a highly automatic and efficient method to bind computer intelligence to the current computer forensic framework. In particular, the authors place emphasis on the data analysis phase. A high level concept, Case-Relevance, is defined to measure the importance of any information to a given case. The proposed framework demonstrates the benefits of computer intelligence technologies: automatic evidence extraction and knowledge reusability.

[3] discusses the requirements of digital image forensics that underpin the design of the forensic image mining system proposed by R.A. Brown, B.L. Pham, and O.Y. De Vel. The system can be trained by a hierarchical Support Vector Machine (SVM) to detect objects and scenes that are made up of components under spatial or non-spatial constraints. In addition, the authors propose to use a Bayesian Networks approach to deal with information uncertainties inherent in forensic work. An analysis of the performance of the first prototype of the system is also provided.

G. Mohay discusses some technical challenges and future directions for digital forensics in [18]. The digital evidence may be manifest in various forms. It may be manifest on digital electronic devices or computers that are simply passive repositories of evidence that documents the activity, or it may consist of information or meta-information resident on the devices or computers that have been used to actually facilitate the activity, or that have been targeted by the activity. In each of these three cases, the author has recorded digital evidence of the activity.

C. Related Work in Ontology Matching

According to the classification in [8], most schema-matching techniques [9] can be divided into two categories: rule-based approaches and learning-based approaches.

1) Rule-Based Approaches: In [19], N.F. Noy and M.A. Musen describe PROMPT, a semiautomatic approach to ontology alignment. By performing some tasks automatically and guiding the user in performing other tasks for which intervention is required, PROMPT helps in understanding ontologies covering overlapping domains.

Similarity Flooding (SF) [17] is a matching algorithm based on a fixpoint computation that is usable across different scenarios. SF takes two graphs as input, and produces as output a mapping between corresponding nodes. This work defines several filtering strategies for pruning the immediate result of the fixpoint computation.

Cupid [15] is an algorithm for generic schema matching outside of any particular data model or application. It discovers mappings between schema elements based on their names, data types, constraints, and schema structure. Cupid has a bias toward leaf structures where much of the schema content resides.

S-Match [10] views match as an operator that takes two graph-like structures and produces a mapping between the nodes of the graphs. Mappings are discovered by computing semantic relations, which are determined by analyzing the meaning that is codified in the elements and the structures of the schemas.

B. Hu et al. [13] explore the ontology matching in a dynamic and distributed environment where on-the-fly alignments are needed. Their approach exploits imperfect consensuses among heterogeneous data holders by collaborating the logic formalisms with Web repositories.

In [22], the authors design a procedure for mapping hierarchical ontologies populated with properly classified text documents. Through the combination of structural and instance-based approaches, the procedure reduces the terminological and conceptual ontology heterogeneity, and yields certain granularity and instantiation judgments about the inputs.

2) Learning-Based Approaches: In [7], A. Doan et al. describe GLUE that employs machine learning techniques to find semantic mappings between ontologies. A Metalearner is used to combine the predictions from both Content Learner and Name Learner; a similarity matrix is then build; and common knowledge and domain constraints are incorporated through a Relaxation Labeler. In addition, GLUE has been extended to find complex mappings.

[1] presents a general method for agents using ontologies to teach each other concepts to improve their communication and thus cooperation abilities. An agent gets both positive and negative examples for a concept from other agents; it then makes use of one of its known concept learning methods to learn the concept in question, involving other agents again by taking votes in case of knowledge conflicts.

J. Madhavan et al. [16] use a corpus of schemas and mappings to augment the evidence about the schemas being matched. The algorithm exploits a corpus in two ways. It first increases the evidence about each element by including evidence from similar elements in the corpus; then it learns statistics about elements and their relationships and use them to infer constraints to prune candidate mappings.

In order to solve the problem of low precision resulted from ambiguous words, J. Gracia et al. [12] introduce techniques from Word Sense Disambiguation. They validate the mappings by exploring the semantics of the ontological terms involved in the matching process. They also discuss techniques to filter out mappings resulting from the incorrect anchoring of ambiguous terms.

P. Lambrix et al. [14] describe SVM-based algorithms to align ontologies using literature. The authors have discovered: (1) SVM-S and NB obtain similar results; (2) the combinations of TermWN with SVM-S and with SVM-P lead to a large gain in precision compared to TermWN and SVM-P.

III. Methodology

A. Overview

In order to obtain “deep” knowledge out of a wealth of digital forensic evidence from heterogeneous resources, the DIEAOM framework aims to reconcile independently designed ontologies from distributed groups, so that to satisfy criminal investigators’ preference to keep their original conceptual models’ integrity, while obtaining the ability to understand others’ models. This goal will be accomplished through a creative learning-based algorithm to match ontologies from different resources. The uniqueness and innovativeness of our approach is that **our learning process depends on ontology schema information alone**. The following section explains in detail why it is challenging for the learning to rely on schema information only.

B. Challenges and Solution

The semantics of an ontology concept is determined by three aspects: (1) concept name, (2) concept properties, and (3) concept relationships. These three features together specify a conceptual model for each concept from the viewpoint of an ontology designer. Any ontology-matching algorithm, either rule-based or learning-based, needs to handle some or all of these three semantic aspects, by different rules or machine learning techniques.

1) Problems with Existing Matching Algorithms: The rule-based algorithms usually have the advantage of relatively fast running speed, but share the disadvantage of ignoring the additional information from instance data. In addition, there is a more serious concern for this type of algorithms. In order to obtain a helpful matching result from any algorithms, more than one of three semantic aspects should be considered. It is then unavoidable to determine the corresponding weights for different aspects, reflecting their different importance (or contributions) in ontology matching. Many existing rule-based algorithms make use of human heuristics and/or domain knowledge to predefine these weights.

While taking advantage of extra clues contained in instance data, the learning-based algorithms are likely to be slower. Moreover, the difficulty in getting enough good-quality data is also a potential problem. On the other hand, it is very challenging for machines to learn to match ontologies by only providing with schema information. The most critical challenge is that, because ontologies reflect their designers' conceptual views, they exhibit a great deal of diversity. Identical terms can be used to describe different concepts, or vice versa, different terms can be assigned to the same concept. A more complicated situation is, even if the same set of terms are adopted, which is almost impossible in the real life, different designers can still create different relationships for the same concept, corresponding to their different conceptual views for this concept. Compared with schemas, instance data usually have a lot less varieties. Therefore, **existing learning-based algorithms make use of instance data, more or less**.

2) Our Solution - the DIEAOM framework: The main idea is that machine learning techniques are essential in ontology matching; however, at the same time, it is preferable to avoid the problem of lacking instance data, either in quality or in quantity, which is *very common* for real-world ontologies. Our learning process is therefore carried out at the schema level, instead of the instance level.

C. Details of DIEAOM

First, *superconcept* is defined as a set of original concepts. Within each superconcept, all *components*, i.e., original concepts, are from different ontologies; at the same time, they are equivalent to each other. DIEAOM tries to find all superconcepts between two ontologies in question. Because the number of superconcepts is not known in advance, an agglomerative clustering algorithm fits our needs.

1) Phase I - Learn Weights: We first calculate similarities in concept name, concept properties, and concept relationships (s_1, s_2, and s_3) between two concepts, C_1 and C_2; the overall similarity value, s, is then obtained as the weighted sum of s_1, s_2, and s_3:

$$s = \sum_{i=1}^{3}(w_i s_i), \tag{1}$$

where $\sum_{i=1}^{3} w_i = 1$, and $s \in [0, 1]$. For two ontologies being matched, $\mathcal{O}_1$ and $\mathcal{O}_2$, the similarity values are calculated for pairwise concepts (one from $\mathcal{O}_1$, the other from $\mathcal{O}_2$, considering all combinations). Then a $n_1 \times n_2$ matrix $\mathcal{M}$ is built to record all values calculated, where n_i is the number of concepts in $\mathcal{O}_i$. The cell $[i, j]$ in $\mathcal{M}$ stores the similarity value between the i^{th} concept in $\mathcal{O}_1$ and the j^{th} concept in $\mathcal{O}_2$. A two-layer 3 × 1 network (Figure 1) is adopted in DIEAOM. Initially, a concept similarity matrix $\mathcal{M}$ is obtained for $\mathcal{O}_1$ and $\mathcal{O}_2$, with w_i's being initialized randomly. A set of concepts are randomly picked up from $\mathcal{O}_1$, and their equivalent concepts are found in $\mathcal{O}_2$ by a manual matching. Each of such manually matched pairs will be processed by DIEAOM, and the similarity values in name, properties, and ancestors for these two concepts are calculated and used as a training example to the network. Note that the correctness of Formula 1 can be verified in the experiments: (1) three weights converge to certain values; and (2) when applying the learned weights to other concepts, the matching result has a good performance.

2) Phase II - Cluster Concepts: Upon obtaining the learned weights for three semantic aspects (name, properties, and relationships), the similarity matrix is recalculated. An agglomerative clustering algorithm is then adopted to form a set of superconcepts.

Input:

- Ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$
- $\mathcal{M}$ (the recalculated similarity matrix between $\mathcal{O}_1$ and $\mathcal{O}_2$)

Output:

- A set of superconcepts

```
begin
  1) Each original concept forms a $singleton$ cluster
```

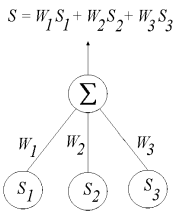

Figure 1. Neural Network Structure

```
    2) Find a pair of clusters, $(a)$ and $(b)$, such that their similarity $s[(a),(b)]$ = max $(s[(m),(n)])$
    3) If $s[(a),(b)] > similarity\ threshold$, go to step 4, otherwise go to step 7
    4) Merge $(a)$ and $(b)$ into a new cluster $(a+b)$
    5) Update the matrix $\mathcal{M}$ by deleting the row and column corresponding to $(a)$ and $(b)$
    6) Repeat steps 2 and 3
    7) Output current clusters as the set of superconcepts
end
```

Pseudocode for Agglomerative Clustering

IV. Conclusion

The Digital Forensics field is facing a data volume explosion, and major challenges are to maintain the integrity of the evidence found by different parties, to establish the evidence's accurate interpretation, and to ensure a trustworthy conclusion drawn. Researchers have proposed to adopt ontological techniques to handle these challenges. In order to tackle the inherent heterogeneity among ontologies from different parties, we propose a Digital Investigation Evidence Acquisition Model Based on Ontology Matching (DIEAOM) to facilitate knowledge collection from the digital evidence, knowledge sharing and reuse, and decision support for criminal investigators. DIEAOM's goal is to synthesize vast amounts of evidence from different parties by matching conceptual models to permit data mining and knowledge discovery.

The uniqueness of our approach is that it depends on ontology schemas alone. To the best of our knowledge, this is the first learning-based ontology-matching algorithm that does not rely on instance data. We are now working on refining our algorithm and evaluating it on real-world ontologies. Besides, we plan to explore the issue of *Evidence Annotation*, i.e., to enrich the original evidence with proper and meaningful semantics. Important features specifically for the forensic evidence need to be encoded into conceptual models (ontologies), e.g., evidence provenance, evidence integrity, etc.

To sum up, by incorporating computer intelligence, especially the techniques available through the pervasive cyberinfrastructure, into the Digital Forensics domain, our proposed DIEAOM framework aims to benefit the current criminal investigation procedure with higher automation, enhanced effectiveness, and better knowledge sharing and reuse.

References

[1] M. Afsharchi, B.H. Far, and J. Denzinger, "Ontology-Guided Learning to Improve Communication between Groups of Agents," Proc. the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 06), Hakodate, Japan, May 2006.

[2] D. Bianculli, R. Jurca, W. Binder, C. Ghezzi, and B. Faltings, "Automated dynamic maintenance of composite services based on service reputation," LNCS 4749, pp. 449-455, 2007.

[3] R.A. Brown, B.L. Pham, and O.Y. De Vel, "Design of a Digital Forensics Image Mining System," Proc. The Workshop on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP05), Melbourne, Australia, September, 2005.

[4] DFRWS CDESF Working Group, http://www.dfrws.org/CDESF/index.shtml, December 2009.

[5] Z. Ding, Y. Peng, and R. Pan, "A Bayesian Approach to Uncertainty Modeling in OWL Ontology," Proc. the International Conference on Advances in Intelligent Systems - Theory and Applications, Luxembourg, November 2004.

[6] Z. Ding, Y. Peng, R. Pan, and Y. Yu, "A Bayesian Methodology towards Automatic Ontology Mapping," Proc. Technical Report WS-05-01 of AAAI Workshop on Contexts and Ontologies: Theory, Practice, and Applications, Pittsburgh, PA, July 2005.

[7] A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, and A. Halevy, "Learning to match ontologies on the Semantic Web," The VLDB Journal, vol.12, no.4, pp. 303-319, 2003, Springer-Verlag, New York, NY, USA.

[8] A. Doan and A.Y. Halevy, “Semantic Integration Research in the Database Community: A Brief Survey,” AI Magazine, vol.26, no.1, pp. 83-94, Spring 2005.

[9] J. Euzenat and P. Shvaiko, *Ontology Matching*, Springer-Verlag, Berlin Heidelberg (DE), 2007.

[10] F. Giunchiglia, P. Shvaiko, and M. Yatskevich, “Semantic matching,” Encyclopedia of Database Systems, pp. 2561-2566, Springer, 2009.

[11] R. Gong, K. Chan, and M. Gaertner, “Case-relevance information investigation: binding computer intelligence to the current computer forensic framework,” International Journal of Digital Evidence, vol.4, no.1, 2005, Springer.

[12] J. Gracia, V. Lopez, M. Aquin, M. Sabou, E. Motta, and E. Mena, “Solving semantic ambiguity to improve Semantic Web based ontology matching,” Proc. the Second International Workshop on Ontology Matching (OM 07), Busan, Korea, November 2007.

[13] B. Hu, S. Dasmahapatra, and P. Lewis, “Emerging consensus in-situ,” Proc. the Second International Workshop on Ontology Matching (OM 07), Busan, Korea, November 2007.

[14] P. Lambrix, H. Tan, and W. Xu, “Literature-based alignment of ontologies,” Proc. the Third International Workshop on Ontology Matching (OM 08), Karlsruhe, Germany, October 2008.

[15] J. Madhavan, P.A. Bernstein, and E. Rahm, “Generic Schema Matching with Cupid,” Proc. the Twenty-seventh VLDB Conference, Roma, Italy, 2001.

[16] J. Madhavan, P.A. Bernstein, A. Doan, and A. Halevy, “Corpus-based Schema Matching,” Proc. the Twenty-first International Conference on Data Engineering (ICDE 05), Tokyo, Japan, April 2005.

[17] S. Melnik, H. Garcia-Molina, and E. Rahm, “Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching,” Proc. the Eighteenth International Conference on Data Engineering (ICDE 02), San Jose, CA, 2002.

[18] G. Mohay, “Technical challenges and directions for digital forensics,” Proc. the First International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE 05), Taipei, Taiwan, November 7-9, 2005.

[19] N.F. Noy and M.A. Musen, “PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment,” Proc. the 17th National Conference on Artificial Intelligence (AAAI 00), AAAI Press, Menlo Park, CA, USA, 2000.

[20] R. Pan, Z. Ding, Y. Yu, and Y. Peng, “A Bayesian Network Approach to Ontology Mapping,” Proc. the International Semantic Web Conference (ISWC 05), Galway, Ireland, November 2005.

[21] S. Raghavan, A. Clark, and G. Mohay, “FIA: An Open Forensic Integration Architecture for Composing Digital Evidence ,” Proc. the 2nd International ICST Conference on Forensic Applications and Techniques in Telecommunications, Information, and Multimedia (eForensics 10), LNICS 1867-8211, Adelaide, Australia, January 19-21, pp. 83-94, 2009.

[22] K. Todorov and P. Geibel, “Ontology mapping via structural and instance-based similarity measures,” Proc. the Third International Workshop on Ontology Matching (OM 08), Karlsruhe, Germany, October 2008.