

Evaluating ontology mapping techniques: An experiment in public safety information sharing

Siddharth Kaza*, Hsinchun Chen

Department of Management Information Systems, University of Arizona, United States

Available online 15 December 2007

Abstract

The public safety community in the United States consists of thousands of local, state, and federal agencies, each with its own information system. In the past few years, there has been a thrust on the seamless interoperability of systems in these agencies. Ontology-based interoperability approaches in the public safety domain need to rely on mapping between ontologies as each agency has its own representation of information. However, there has been little study of ontology mapping techniques in this domain. We evaluate current mapping techniques with real-world data representations from law-enforcement and public safety data sources. In addition, we implement an information theory based tool called MIMapper that uses WordNet and mutual information between data instances to map ontologies. We find that three tools: PROMPT, Chimaera, and LOM, have average F-measures of 0.46, 0.49, and 0.68 when matching pairs of ontologies with the number of classes ranging from 13–73. MIMapper performs better with an average F-measure of 0.84 in performing the same task. We conclude that the tools that use secondary sources (like WordNet) and data instances to establish mappings between ontologies are likely to perform better in this application domain. © 2007 Elsevier B.V. All rights reserved.

Keywords: Ontology mapping; Public safety information sharing; Mutual information; Intelligence and security informatics

1. Introduction

The public safety community in the United States comprises of several thousand state and local police agencies, hundreds of federal agencies ranging from the FBI to the Government Printing Office police, and thousands of courts, prosecutor's offices, and probation departments. Each of these agencies has one or more

internal information system as well as links to national information systems. In the past few years, there has been a thrust on the interoperability and information sharing between systems in agencies responsible for local law-enforcement and national security. However, a majority of these agencies have records management systems that are incapable of sharing information with each other in an efficient manner because they are not interoperable [4].

The interoperability problem exists because there is a need to connect information systems that are heterogeneous and incompatible. This problem has been a major focus of the research and practitioner communities alike. There is a large body of work in the database and information integration fields that ranges from matching database schemas to answering queries from multiple

* Corresponding author.

E-mail addresses: sidd@u.arizona.edu (S. Kaza),
hchen@eller.arizona.edu (H. Chen).

systems. Recently, research has focused on the use of ontologies as they might play a key role in achieving seamless connectivity between systems [2]. However, there has not been much ontology related research in the public safety domain due to the lack of standard ontologies and the sensitive nature of information.

Many ontology-based information sharing approaches rely on mapping between ontologies from different sources. Mapping tools use different techniques to suggest matches between ontology elements, and vary in input requirements, output formats, and modes of interaction with the user. Due to their diversity, there has been little work on the comparative evaluation of mapping techniques in the information integration literature [18]. Thus, there is a lack of understanding of their pitfalls with real-world data. Mapping between ontologies is especially important in the public safety domain since each agency has a different representation for information. Comparative evaluation of appropriate mapping techniques with real ontologies from the public safety community allows us to study them in detail and take a few more steps to the goal of seamless connectivity between information sources.

In this study, we survey public safety information sharing initiatives and discuss the ontological implications of the systems and standards. We also perform a survey of ontology mapping and evaluate commonly used tools that use various techniques to map ontologies. In addition, we design a mapping tool that uses WordNet and mutual information between data instances and compare its performance to other techniques. In doing this, the study has two primary aims: firstly, address important issues on the use of ontology-based information integration techniques with public safety and law-enforcement data, and secondly, identify the set of techniques (and not necessarily the tools themselves) that perform well with real law-enforcement data representations.

The study attempts to answer the following questions:

- How can we ontologically define important law-enforcement and other public safety data elements?
- How do the common ontology-mapping techniques perform with real law-enforcement data representations?
- How can we use data instances to identify mappings between similar ontology classes?

The next section discusses relevant previous work in this area. Section 3 presents the research design, testbed, and the design of the new mapping tool used in our work. The experiment results are presented and discussed in Section 4. Section 5 concludes and presents the possible future directions for this study.

2. Related work

2.1. Challenges in information sharing and interoperability

The problem of bringing together heterogeneous and distributed computer systems is known as the interoperability problem [46]. Data and database heterogeneity is classified into structural (data is in different structures) and semantic (data in different sources has different meaning) heterogeneity. Even though structural conflicts pose major problems in information integration, they have been addressed to a great extent in the schema matching literature (see Rahm et al. [40] for a comprehensive survey). Semantic heterogeneity and conflicts have been identified as one of the most important and pressing challenges in information sharing [26]. The conflicts have been categorized into many different taxonomies by previous studies including data level and schema level conflicts [39]; confounding, scaling, and naming conflicts [14]; and data entry and representation conflicts [21]. Semantic conflicts in the public safety information domain include:

- Data representation/expression conflicts: Data are recorded in different formats, e.g., different representations of date ('050798' vs. '7-May-98').
- Data precision conflicts: The domain of data values differ, e.g., five levels of security alerts (DHS advisory system) vs. four level systems (many police agency alert systems).
- Data unit conflicts: the units of data differ, e.g., measurement of height in inches vs. in feet.
- Naming conflicts: The labels of elements differ, e.g., an element might be called 'Person' in one data source whereas it might be called 'Individual' in another.
- Aggregation conflicts: An aggregation is used in one data source to identify a set of entities in another source, e.g., date is represented as month, date, and year as separate attributes in one source and as combined attributes in another source.

Another important challenge in sharing public safety related data is addressing privacy, security, and policy concerns. Federal, state, and local regulations require that agreements between agencies within their respective jurisdictions receive advance approval from their governing hierarchy. This precludes informal information sharing agreements between those agencies. The requirements vary from agency to agency according to

the statutes by which they were governed. Sharing information among these agencies requires many months of negotiation along with formal agreements. Though extremely important, these concerns are not the primary focus of this paper.

2.2. Public safety information sharing

The domain of public safety information sharing has not been adequately explored in the research community. This stems from the sensitivity and policy issues associated with obtaining this information for research purposes. However, there have been some successful national, state, and local efforts to share data among public safety agencies. Table 1 lists a few of the major public safety information sharing efforts along with their potential ontological contributions. Selected initiatives are also discussed in detail in the following sub-sections.

2.2.1. National level sharing initiatives

Most national level initiatives are championed by the Government. Perhaps the most important recent one is a data model standardization effort by the U.S. Department of Justice, called the Global Justice XML Data Model (GJXDM, <http://it.ojp.gov/jxdm>). The GJXDM is an XML based standard intended to be a data reference

model for the exchange of information within the justice and public safety communities. It defines the semantics and the structure of common data elements that are used in the community. The data elements are based on approximately thirty-five different data dictionaries, XML schema documents, and data models used in the justice and public safety communities. The purpose of the GJXDM is to provide a consistent, extensible, maintainable XML schema reference specification for data elements and types that represent the data requirements for a majority of the community. It is intended to allow database schema developers to base the components and structures of the schema elements on a universally accepted standard. From the ontological perspective, the GJXDM is an important step in the right direction. It provides a comprehensive collection of data classes in addition to a taxonomy outlining relationships among them. This will allow future public-safety ontology engineers to base their ontology on a commonly used standard, thus enabling easy interoperability.

An established national level data sharing initiative is the Regional Information Sharing Systems (RISS, <http://www.iir.com/riiss/>) program that links law-enforcement agencies throughout the nation, providing secure communications, information sharing resources, and investigative support. Existing information systems in agencies

Table 1
National and regional public safety information sharing efforts

	Type	Agency participation	Ontological contributions	
National	GJXDM (DOJ)	Data standard	40 justice related federal, state, and local agencies (April 2005), but likely to expand soon	Classes, properties, and comprehensive taxonomy for the justice community
	NIEM (DHS,DOJ) www.niem.gov	Data standard	13 federal, state, and local agencies (May 2006)	Classes and properties for homeland and border security agencies
	RISS (BJA)	Data standards and communication (comm.) network	485 state agencies and 920 federal agencies (May 2006)	Message exchange ontologies and classes for law-enforcement data
	N-Dex (FBI)	Data warehouse	20 pilot federal, state, and local agencies (March 2006)	Classes for incident reports
	NLETS (multiple agencies)	Comm. network	Several hundred federal, state, local, and some international agencies	Message exchange ontologies
	NCIC (FBI)	Data warehouse and comm. network	Several hundred federal, state, and local agencies	Primarily individual and vehicle related classes
Regional	COPLINK® (multiple regions including AZ and CA)	Data warehouse and analysis suite	Over a 300 state and local agencies and a few federal agencies	Comprehensive classes and properties for law-enforcement data and analysis results
	ARJIS (San Diego, CA region)	Distributed query and communication network	50 state and local agencies and a few federal agencies	Law-enforcement message exchange ontologies
	CAPWIN (Washington, D.C.)	Distributed query and communication network	40 state and local agencies and 10 federal agencies	First responder (police, fire, EMS) data and message exchange ontologies
	FINDER (Florida) finder.ucf.edu	Distributed query and communication network	121 local agencies	Message exchange ontologies for individual, vehicle, and pawn records
	LINX (VA and others) www.hrlinx.com	Data warehouse	27 state and local agencies	Incident report classes

connect as nodes to the RISS net-work and use standardized XML data exchange specifications to exchange information on individuals, vehicles, and incident reports.

In addition to data standardization efforts, other national level information sharing efforts include the National Law Enforcement Telecommunications System (NLETS) and the FBI's National Crime Information Center (NCIC) network. Both of these systems are telecommunications networks that enable agencies to submit and access public safety information. Information in these systems is entered manually and most local and state law-enforcement information is not available through them. Even though the systems are in wide use, their narrow scope and the lack of efficient methods of mapping information to them limit their use.

2.2.2. Regional level sharing initiatives

The initiatives described in this sub-section have attempted to address the problem of interoperability among the state and local public safety agencies. Many of these systems are capable of scaling up to the national level and adapting to larger scale public safety related applications. One of the most successful efforts in regional data sharing is the COPLINK[®] [4] system. COPLINK[®] is based on mapping multiple database schemas into one standard schema. It allows diverse police departments to share data seamlessly through an easy-to-use interface that integrates different data sources including legacy records management systems. During the writing of this paper, the system was being used by over a 300 law-enforcement agencies in more than twenty states.

Another regional sharing effort is the Automated Regional Justice Information System (ARJIS, www.arjis.org) adapted by over fifty agencies in the San Diego, C.A. region. The ARJIS system uses a query translation approach to send distributed queries to various agencies on the network. A third notable regional sharing effort is Capital Wireless Integrated Network (CapWIN, www.capwin.org) for first responders in the Washington, D.C. region.

As can be seen in Table 1, from an ontological perspective, multiple agencies in the community can contribute to different aspects of a public safety ontology. The survey also shows the presence of multiple information sharing efforts, each with its own representation of information. This increases the need to map between these various representations for systems to interoperate.

2.3. The role of ontologies

The term 'ontology' originated in philosophy (meaning subject of existence) and acquired a different

meaning in the context of information sharing. In the artificial intelligence and information integration literature, ontology is "a formal, explicit specification of a shared conceptualization [15]." In simpler terms, an ontology is a formal description of concepts (known as classes) and relationships (known as properties) that exist between them. There are several languages to represent ontologies. The most prominent ones include: OWL (Web Ontology Language, www.w3.org), DAML (DARPA Agent Markup Language, www.daml.org), and RDF (Resource Description Language, www.w3.org). Four roles of ontologies can be identified in the information integration literature [45,46]:

- Neutral authoring: Modeling all the data in an enterprise to enhance maintainability and long term knowledge retention.
- Interoperability: Explicating content to allow multiple systems to interoperate.
- Ontology-based specification: Building software systems based on predefined ontologies.
- Ontology-based query models: Using ontologies as a global query schema, sub-queries are reformulated based on a global ontology.

Ontology-based interoperability techniques have been used in many domains including e-commerce [12,36], health [3], finance [11], electronics [8], bioinformatics [42] and manufacturing [16]. Some of these studies have used the multiple local ontology approach [3,11,12,36] while others have preferred the hybrid approach (a common vocabulary is used to generate multiple local ontologies) [8,16,42]. Many times the selection of an approach depends upon the availability of standard vocabularies [30]. This is an important challenge in ontology-based approaches since most communities that need to interoperate do not share a single ontology [45] or have source-specific disparate ontologies already defined for them [35]. This is especially true in the public-safety information domain where different sources use different representations of information (some of which may be proprietary). Ontology based information sharing between such sources requires mapping between the disparate ontologies. Ontology mapping is a difficult and time consuming task and has been extensively studied in the literature.

2.4. Ontology mapping

Identifying semantic correspondences (mappings) between ontologies and database schemas has been the

Table 2
Representative set of ontology mapping tools

		Mapping	Input type–output type	Uses data instances	Level of automation	Availability
Primarily NLP based	Prompt, Anchor– Prompt, PromptDiff	NLP methods, heuristics based on structure	<i>Input:</i> OWL, RDF <i>Output:</i> merged ontology with suggestion of similar concepts to guide the user	Yes, Not necessary for functioning	Semi-automated. Guides user through merging steps	Available as parts of the Protégé ontology development environment
	Chimaera	NLP Methods Structural heuristics (only within the Ontolingua environment)	<i>Input:</i> Ontologies in multiple representations. <i>Output:</i> merged ontology with suggestions of similar concepts	No	Semi-automated. Guides user through merging steps	Available as part of the Ontolingua environment and as a stand-alone online tool
	FCA-merge	NLP Methods	<i>Input:</i> documents containing domain concepts <i>Output:</i> pairs of similar concepts	Yes. Necessary for functioning	Relies on human expert to construct ontology from similar concepts	Not available for public use
Structure based	GLUE	Probabilities definitions fo several similarity measures. Uses relaxation labeling to incorporae domain constraints	<i>Input:</i> taxonomies not ontologies <i>Output:</i> mappings between concepts	Yes. Necessary for functioning	Automated. No human necessary. Probably will need experts to verify and correct mappings	Not available
	OMEN	Probabilities definitions fo several similarity measures. Different from GLUE in that this uses the semantics of ontology relations	<i>Input:</i> RDF <i>Output:</i> mappings between similar concepts	No	Automated. No human necessary. Probably will need experts to verify and correct mappings	Not available
External sources	IF-Map	Logic Isomorphism – Generates subclass relationships between local ontologies and reference ontology	<i>Input:</i> multiple — Ontology libraries, editors, Web, ontology representations. <i>Output:</i> merged ontology	Yes. Necessary for functioning	Semi-automated. Needs expert	Not available
	LOM	NLP Methods. In addition, uses Wordnet Synsets and SUMO types to match concepts	<i>Input:</i> DAML/RDF <i>Output:</i> mappings between similar concepts	No	Automated. Gives a probability for each match	Prolog demo implementation available for download

The tools include Prompt [31], Anchor-Prompt [32], Prompt-Diff [34], Chimaera [27], FCA-Merge [43], Glue [6], OMEN [28], IF-Map [17], and LOM [23].

focus of many works from diverse communities [18,40]. There are two major approaches for discovering mappings between ontologies [30]. If the ontologies share the same upper model, then this common grounding can be used to establish mappings. There are several upper ontologies that include SUMO [29], DOLCE [13], SENSUS [1], and Cyc [22]. The second set of approaches characteristics of ontologies like names (NLP based) and definitions of classes, structure, and instances of classes to establish mappings. These are similar to schema matching techniques but sometimes use automated reasoning to identify hierarchies. Some tools also use other external reference ontologies to establish mappings. The second approach is more suitable in the public safety domain as there is still a lack of consensus on middle and upper level ontologies. Table 2 presents a representative set of these tools and their characteristics. The tools are broadly divided based on the core techniques, which are their use of NLP, structural properties, use of data instances, and use of secondary sources for mapping.

As can be seen from Table 2, ontology mapping tools do not have common input requirements. Many tools do not take standard input in the form of OWL, RDF or DAML files and derive ontological concepts directly from documents or from the Web. The output formats vary from fully merged ontologies to class-to-class mappings. The tools also provide different levels of support to users and require varying amounts of human input. Many of the tools merge ontologies without human intervention; others (like PROMPT [31]) suggest matches and require an expert to verify them at each step in the merging/mapping process. This is usually considered better for most mapping scenarios. Most of the tools are also not available for public use that limits their use to address real-world ontology mapping tasks. In terms of evaluation, the accuracy of matches identified by GLUE [6], a matching tool that uses structural properties and data instances, ranged from 66% to 97% in different datasets ranging from 34–176 classes. A

precision of 71% and a recall of 57% in matching concepts between Cyc and SENSUS ontologies are reported for LOM [23], a tool using secondary sources. Mitra et al. [28] report almost a 100% precision and recalls as high as 80% using OMEN, a probabilistic tool using structural properties, on two ontologies containing 19 classes. Due to their diversity, there have been few studies that compare ontology mapping techniques [18]. The recent I3CON (<http://www.atl.lmco.com/projects/ontology/i3con.html>) and Ontology Alignment Evaluation Initiative (OAEI, <http://oaei.ontologymatching.org>) are steps in the right direction. These and other conferences aim to provide a standard platform to compare mapping tools. However, even though the previous competitions compare the tools head-to-head, the results don't always provide insights on how the tools will perform in real-world scenarios. In addition, it is not always clear which techniques (NLP, structural, use of secondary sources, and use of data instances) are best suited for mapping between domain ontologies made from real data representations. In this article, we compare popular ontology mapping tools PROMPT [31], Chimaera [27], and LOM [23] using multiple ontologies derived from real schemas used by law-enforcement agencies. All three of these tools use a different combination of techniques to match ontologies. A detailed description of the functioning and the reason for the selection of these tools is presented in Section 3.2. In addition, we also design and evaluate a new tool that uses NLP techniques and mutual information between data instances. The mutual information measure offers certain advantages (these are discussed in Section 3.3) compared to other measures used in previous data-driven techniques.

3. Research design and testbed

Our research design is based on the multiple ontology approach where ontologies are defined for individual sources that provide the structure and

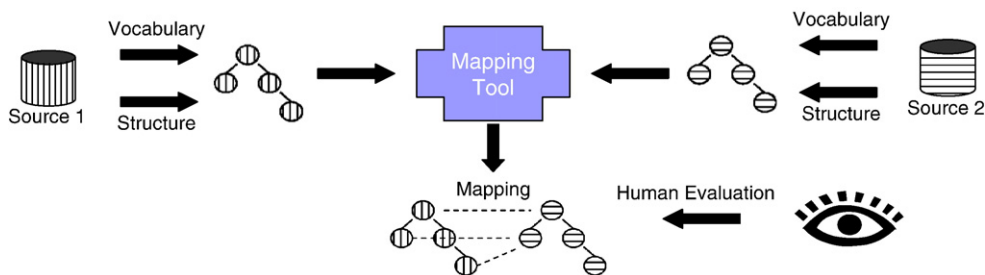


Fig. 1. Research design: Evaluating ontology mapping tools.

vocabulary for the ontology (Fig. 1). A mapping tool is used to suggest the mappings between two source ontologies and these are compared to a benchmark. The benchmark mappings are provided by a domain expert who has intricate knowledge of the structure and semantics of the data sources. In this study, the benchmark mappings are identified with the help of a police detective with over 30 years of public safety and information technology experience at a mid-size city police department.

We use the precision and recall metrics as used by previous evaluative studies of mapping tools. The F-measure represented as

$$F(\beta) = \frac{(\beta^2 + 1) * P * R}{\beta^2 * P + R}$$

is used to combine precision (P) and recall (R). The value of β can be modified to change the relative importance of precision and recall. Setting β to '1' in the above equation makes the F-measure an average of the two measures. In this study, the β was set to 0.5 to make precision twice as important as recall. We believe that even though the tools may not be used without human verification, precision may be more important than recall in this and other sensitive domains.

3.1. Data sources

Three data sources are used to create six ontologies for this study. The first two data sources are database schemas used by many police agencies in the U.S. The third data source is the Global JXDM (JX) by the U.S. Department of Justice. The first police department data source (hereafter referred to as PD1) contains 3 million police incident records with information on 1.8 million individuals and 700,000 vehicles recorded over 10 years. The second police department data source (hereafter referred to as PD2) contains 2.1 million police incident records with information on 1.3 million individuals and 500,000 vehicles recorded over 10 years. These data sources are representative of those in major police agencies and thus mapping between them is of vital importance to the public safety community. For testing, we extracted subsets of schema elements that are used to record information on individuals and vehicles. We selected these as they are one of the most important elements in any law-enforcement/national security investigation. The subsets were used to design person and vehicle ontologies (using OWL and RDF) for each of the data sources. These ontologies are described in later sections.

3.2. Mapping tools for evaluation

For evaluation, we selected PROMPT [31], Chimaera [27], and LOM [23] based on the following criteria:

- *Input type*: We needed tools that take OWL/RDF based ontologies as input.
- *Output type*: We needed tools that suggest mappings as pairs of related classes.
- *Interaction*: We needed tools that suggest initial mappings without human intervention.
- *Representation*: We needed a set of tools that cover the range of techniques (NLP, structural, use of secondary sources, and use of data instances) in the literature.
- *Availability*: We needed tools that are publicly available, stand-alone or as a part of an ontology development environment.

Chimaera [27] suggests potential matches between classes of two ontologies based on the class names. The tool depends on the user to decide final mappings. Within the Ontolingua [27] environment the tool also suggests similarity scores between classes based on structural comparisons of the ontologies. In this study, since we use Chimaera outside the Ontolingua environment, only matches on the names of classes (an NLP technique) are used as a heuristic for mapping. PROMPT [31] is a tool that leads a user through the ontology merging process by identifying possible points of integration, and making suggestions for operations that should be done next [33]. The tool compares names of classes, relations among them, and constraints on slot values to make its suggestions. Thus, as shown in Table 2, the tool uses both NLP techniques and structural properties of ontologies. LOM [23] uses four methods to match classes from two ontologies. They are (1) exact name matches, (2) partial matches on names, (3) Synset matching: uses WordNet [10] to identify synonyms, and (4) type matching: exploring the ontological category of each class using SUMO (the Suggested Upper Merged Ontology, <http://ontology.tekknowledge.com>). Thus, LOM is a technique that utilizes NLP methods and refers to secondary sources to identify matches (see Table 2). The above three tools use various combinations of the common techniques reflected in literature. In addition, they are also publicly available and commonly used (especially PROMPT and Chimaera) for ontology development and mapping. Few tools that use data instances were found to be publicly available.

We evaluated these three tools and compared their performance to the MIMapper that uses NLP techniques

in addition to data instances from ontologies to identify matches.

3.3. MIMapper: ontology mapping using WordNet and mutual information

The MIMapper tool determines matching classes between different ontologies using a two step process. The first step is lexicon driven, which uses the names of classes to identify matches. As done in some previous techniques, classes with the same or partially overlapping names are considered similar to each other. For instance, the class ‘Street’ is considered similar to the class ‘StreetName’ since there is a partial string match in the name. No other similarity metrics (e.g., edit-distance, token-based metrics) for the names of classes were used since these usually require experiments to determine a threshold for the similarity value between strings.

In addition to matches on the given class names, MIMapper also incorporates the use of the WordNet 2.0 [10] ontology. WordNet Synsets are used to determine if the names of the classes are synonymous. Classes with synonymous names are considered similar to each other. Once the synonyms of class names are found, the partial string match is used again. For instance, consider two class names: ‘Gender’ and ‘SexText.’ The partial string match routine cannot find a similarity between them, since the strings do not match. However, on using WordNet, the strings can be considered as ‘Gender’ and ‘GenderText’ since ‘gender’ and ‘sex’ are synonymous. Thus, in the second iteration, the partial string match routine identifies these class names as similar. Some popular tools like PROMPT and Chimaera do not use WordNet to aid in finding similar classes. On the other hand, LOM utilizes WordNet and the capability of MIMapper and LOM are very similar in terms of NLP techniques.

The second step is data driven, which uses data instances from the ontology to establish mappings between them. This approach uses point-wise mutual information (an information-theoretic concept) to identify the most informative data instances for each class in an ontology. The earliest definitions of point-wise mutual information (MI) were given by Fano [9]. It was defined as the amount of information provided by the occurrence of an event (y) about the occurrence of another event (x) and formulated as

$$I(x;y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

Intuitively, this concept measures if the co-occurrence of x and y ($P(x,y)$) is more likely than their independent occurrences ($P(x)*P(y)$). This concept and the procedure

Table 3
Sample feature vector for *First Name*

Instance	Frequency
Jose	2
Ben	1
John	3

outlined below has been used by previous studies [37,38] to match database columns from different data sources. The studies found that the MI measure can be used to find associations between schemas with a large amount of data. MI has also been previously used [5,19] to find associations between various seemingly unrelated objects. In this study, it is assumed that instances of similar or equivalent classes from disparate ontologies are likely to share common informative instances. For instance, consider the two classes *First Name* and *Given Name* from two different person ontologies. The two classes are likely to share more common informative instances (unusual first names) with each other rather than with other classes like *Last Name* or *City*. In order to calculate the mutual information, each instantiated class (a class with data instances) is represented by a feature vector. The feature vector contains each distinct instance along with a frequency count of the number of times it occurs in that class. Table 3 is an example of a feature vector for the *First Name* class.

Then, for each class c , a feature count vector ($F(c)$) is constructed:

$$F(c) = (f_{c1}, f_{c2}, \dots, f_{cm})$$

where m is the total number of features (over all classes) and f_{ct} is the frequency of occurrence of feature t in class c . For instance, for class $c=First\ Name$ and feature $t=John$, $f_{ct}=3$ (from Table 3). Using this procedure we obtain a large frequency count matrix with n rows for classes and m columns for features with each cell containing the frequency count f_{ct} . This matrix can be very large and sparse depending on the number of classes to be matched and the number of data instances provided.

With the frequency count vectors, the MI vector is constructed for each class as:

$$MI(c) = (mi_{c1}, mi_{c2}, \dots, mi_{cm})$$

where mi_{ct} is the point-wise mutual information between class c and feature f that is calculated as:

$$mi_{ct} = \log \frac{\frac{f_{ct}}{N}}{\sum_{i=1}^n \frac{f_{it}}{N} * \sum_{j=1}^m \frac{f_{cj}}{N}}$$

where n is the number of classes and $N = \sum_{i=1}^n \sum_{j=1}^m f_{ij}$ is the total frequency count of all the features in all classes. Intuitively, this formulation calculates the information content of a feature (t) with respect to a class (c) by taking into consideration the number of classes that contain the feature (first denominator term) and the number of features that are contained in the class (second denominator term). The division by N is used to generate probabilities as shown in the MI formulation (Eq. (1)). The use of MI gives MIMapper certain advantages in the use of data instances over previous data-driven techniques. For instance, in FCA-Merge [43] the first step is to identify common documents that describe similar ontological concepts. However, at this step no attempt is made to identify the most informative documents for a particular concept. If a large number of documents are available, then MI can be used to identify a subset of documents that provide the highest information content and reduce the computational complexity. GLUE [6], another data driven tool, provides a framework that utilizes multiple similarity measures to estimate similarity between classes. The measures proposed in [6] include the Jaccard similarity and the ‘most-specific-parent’ similarity. The use of a specific similarity measure is contingent upon many factors, and different measures may perform differently in a given situation. A recent study [44] showed that the MI and Jaccard measures returned very similar results in an association analysis task. Another study [20] showed that Jaccard performed better than MI in query expansion, but MI was found to be more efficient. Since GLUE provides a framework where multiple similarity measures may be used, in future work, the MI measure can be used in the GLUE framework to complement it.

MI scores calculated with the point-wise method (as above) are biased by features that occur very rarely [7]. In order to prevent this, the scores are multiplied with a discounting factor based on the frequency counts of the features [25]. We multiply mi_{ct} with the following discounting factor based on the previous study [38] to mitigate the effect:

$$\frac{f_{ct}}{f_{ct} + 1} * \frac{\min\left(\sum_{i=1}^n f_{ci}, \sum_{j=1}^m f_{jt}\right)}{\min\left(\sum_{i=1}^n f_{ci}, \sum_{j=1}^m f_{jt}\right) + 1}$$

Once the MI vectors for each class are obtained, the similarity between two classes is calculated using the cosine coefficient of their vectors [41]. The cosine

Table 4
Statistics of person and vehicle ontologies

	Police DB 1 (PD1)		Police DB 2 (PD2)		GJXDM (JX)	
	Person	Vehicle	Person	Vehicle	Person	Vehicle
	Number of classes	33	13	17	25	73

coefficient measures the cosine of the angle between the two vectors. A cosine of 0 indicates that there is no similarity (i.e., the vectors are orthogonal) and a cosine of 1 indicates that the vectors are identical. Thus, a cosine approaching 1 will indicate that the two given classes are similar to each other. The cosine metric is used because it gives less weight to the presence of zero values in the frequency count vectors [24]. This is an important property since the frequency count vectors are very sparse by nature. Thus, the cosine metric measures the similarity between classes based on the features they share (having frequency count >0) rather than by the features they don’t [38].

Such similarity scores were calculated between all pair of classes in the ontologies to be matched. A class $C1$ was considered similar (and thus mapped) to a class $C2$ if $C1$ had the highest similarity score with $C2$. For performance and efficiency reasons, only one thousand data instances were used for the experiments in this study. This number is similar to the amount of instances used by other data-driven techniques like GLUE [6].

4. Experiment results

4.1. Person and vehicle ontology characteristics

The statistics of the ontologies developed for individuals and vehicles from the three data sources are shown in Table 4.

The three sources differ in the amount of information they represent, with GJXDM containing a large number of classes since it is meant to be comprehensive enough to represent the entire public safety community. It contains elements such as ‘PersonFosterParent’ and ‘PersonNeighbor’ that are usually not included in police databases. Fig. 2 shows snapshots from the three person ontologies.

As can be seen, the ontologies differ in their structure and use of vocabulary. Class-subclass relationships were defined for classes where such relationships were appropriate or could be inferred (by a human) from the underlying data sources. In addition to class hierarchies, OWL object properties were created to

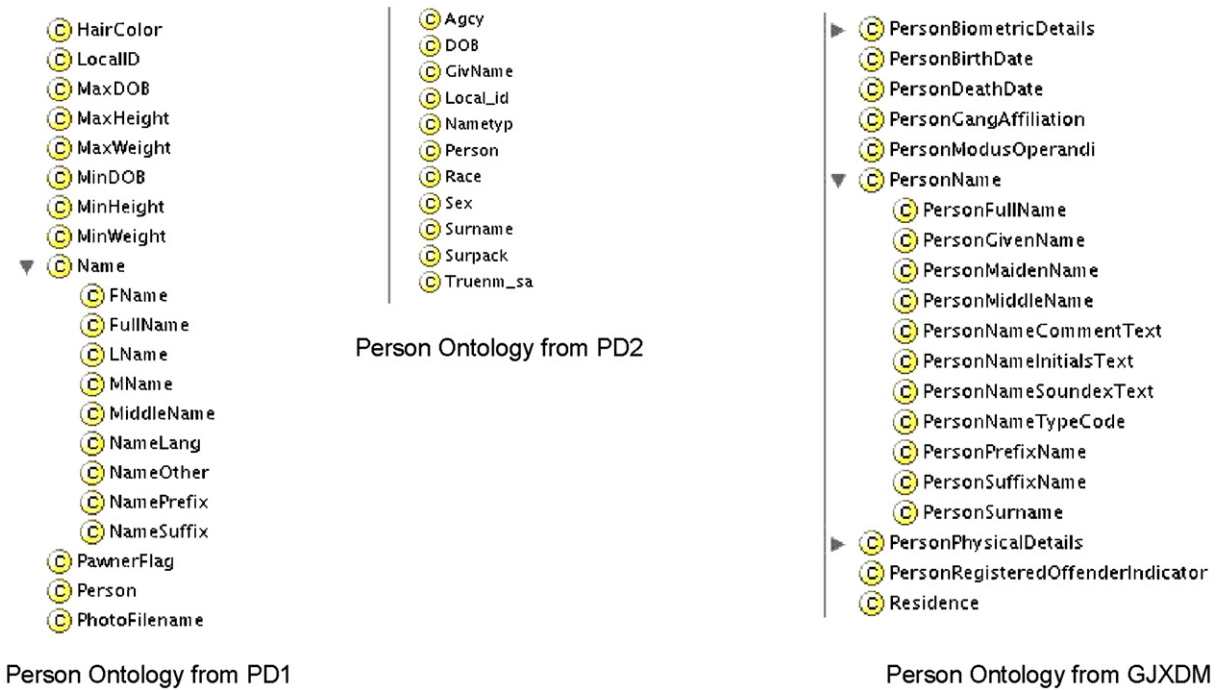


Fig. 2. Snapshots of person ontologies (screen shot from Protégé — <http://protege.stanford.edu>).

define relationships between various classes and appropriate domain and range classes were specified for properties. However, none of the properties were used to create restrictions. This was done as common restrictions (or integrity constraints) like ‘every person is required to have a SSN’ were not present in the underlying data sources. This is because police data is often incomplete and even basic information about a person or a vehicle may not be known. In such a scenario, ontologies need to be defined in a very general manner to accommodate all kinds of data instances. Table 5 shows the number of classes that match for each pair of person and vehicle ontologies. These numbers were used as benchmarks to calculate precision and recall.

4.2. Matching results for person ontologies

Fig. 3 shows the average recall, precision, and the F-measure achieved by the four ontology mapping tools in matching person ontologies.

Table 5
Matching classes in person and vehicle ontologies

		PD2	JX		PD2	JX
Person	PD1	13	18	Vehicle	PD1	9
	JX	8	–		JX	8

The first two coordinates in the triplet represent the recall and precision values. The third coordinate is the F-measure which is represented by the size of the bubble. The values were obtained by averaging the scores over the results of matching three pairs of person ontologies. Table 6 lists the individual precision and recall values achieved by the four tools in matching each pair of person ontologies.

As can be seen, both Chimaera and PROMPT achieved high precisions in all three evaluations. Chimaera relies on advanced NLP matching techniques and it found mappings between classes that had the same or similar names. PROMPT uses structural matching techniques and type constraints in addition to NLP and found accurate matches in addition to finding more matches as compared to Chimaera (as shown by the higher recall numbers). LOM uses secondary sources like WordNet and SUMO in addition to NLP techniques and it had a much higher recall (0.44) as compared to PROMPT (0.21) and Chimaera (0.16). However, overall MIMapper performed better than all the tools with an F-measure of 0.85. This was due to the use of data instances in addition to NLP techniques. Table 7 shows some representative examples of class matches that the tools identified in addition to techniques that were used to obtain those matches. A discussion of the matches that were not found by the tools is included in Section 4.4.

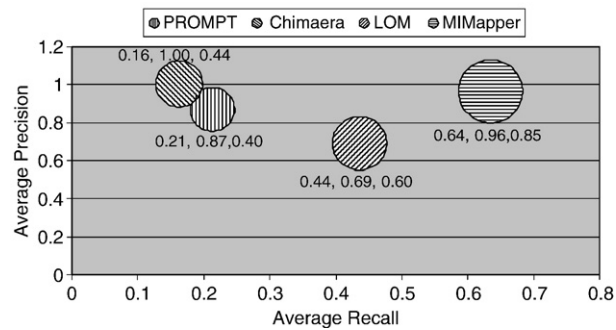


Fig. 3. Average recall and precision achieved by four ontology mapping tools in matching person ontologies. The size of the bubble indicates the F-measure.

4.3. Matching results for vehicle ontologies

Fig. 4 shows the average recall, precision, and the F-measure achieved by the four ontology mapping tools in matching vehicle ontologies. Overall, all tools had a better precision with vehicle ontologies as compared to person ontologies. This was probably because terms like ‘make,’ ‘model,’ ‘VIN,’ and ‘color’ were consistent across the three ontologies. PROMPT and MIMapper had high recalls in matching PD1 to PD2. On closer inspection, we saw that PD1 and PD2 had very similar terminologies for vehicle characteristics. The tools had a slightly higher F-score (LOM: 0.75, Chimaera: 0.54, PROMPT: 0.52, shown in Table 8) as compared to the results for person ontologies due to higher precision values. MIMapper performed better in matching vehicle ontologies as compared to other tools, however, the margin of improvement over the others decreased. This was due to the increase in the level of performance of other tools due to consistent nomenclature across ontologies. In addition, some characteristics of the data prevented the MIMapper to find mappings (these are explained in the next section).

4.4. Discussion and error analysis

The matches recommended (and those not recommended) by each tool were influenced by three factors:

tool-specific techniques, input ontologies and instances, and tool-specific parameters. The tools had a different level of sophistication in the set of techniques used. Chimaera primarily concentrated on NLP techniques and thus its performance was reduced by the fact that the six ontologies did not use a common vocabulary. In a hybrid ontology integration approach, Chimaera is likely to perform much better since the different ontologies share a common vocabulary. It appears that the NLP techniques of the tools also suffered due to GJXDM’s use of the ISO standard naming system for data elements (e.g., ‘PersonGivenName’ instead of ‘GivenName’). This may have prevented exact and partial matches on class names. Thus, in this scenario, it would be beneficial to use a common vocabulary (like GJXDM) to increase the precision and recall of NLP based tools.

The second factor that may have influenced the performance is the ontology creation approach. Most structure-based ontology mapping tools (like PROMPT) assume the presence of a large number of properties that link various classes (or individuals in those classes) to each other. They may also assume the presence of an upper ontology in addition to the domain ontologies which is typical in a top-down ontology construction approach. However, in this study, we performed a bottom-up construction of the ontologies that led to smaller size ontologies with no link to any upper

Table 6
Precision and recall in matching person ontologies

	Chimaera		PROMPT		LOM		MIMapper	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
PD1 — PD2	1.0	0.30	0.6	0.46	0.8	0.30	1.0	0.46
PD1 — JX	1.0	0.06	1.0	0.06	0.6	0.50	0.89	0.44
PD2 — JX	1.0	0.13	1.0	0.13	0.67	0.50	1.0	1.0
Average	1.0	0.16	0.87	0.21	0.69	0.44	0.96	0.64
F-measure	0.44		0.40		0.60		0.85	

Table 7
Examples of matches obtained by four ontology mapping tools

Chimaera	PROMPT	LOM	MIMapper
DOB — DOB (NLP: name match)	PersonFullName — FullName (Structural: Both are subclasses of Name)	Gender — Sex (WordNet)	Gender — PersonSexText (WordNet+name similarity)
LocalID — Local_ID (NLP: name similarity)	Person — Person (NLP)	SID — PersonStateID (SUMO: Same category) NamePrefix — PersonPrefixName (NLP: name similarity)	FName — PersonGivenName (Mutual Information) DOB — PersonBirthDate (Mutual information)

ontology model. Even though properties (like *is-a* and *has-a* relationships) existed between the classes of the person and vehicle ontologies, they may not have been sufficient for structure-based tools to infer matches. However, as discussed in Section 4.5, many domain ontologies may not have the necessary density of structure needed for the effectiveness of such tools. The structure of ontologies used in this study reflects a real-world scenario for the application of the tools. On the other hand, even though LOM and MIMapper performed well in this study, tools that only focus on secondary sources and data instances may not perform as well in large-scale ontology mapping scenarios.

The MIMapper, which was based on the assumption that data instances were available, also suffered due to the properties of the datasets. Since the tool was based on mutual information, it required that the matching classes share some instances of data between them. This made it easy for the technique to match classes such as *Fname* to *PersonGivenName* because of the limited domain of values that the class instances can take (common and informative first names would boost the mutual information score). However, for classes where the domain of values is too large (e.g., FBI IDs), the matching classes are unlikely to share informative instances and thus MIMapper was unable to match such classes. In addition, MIMapper also suggested wrong matches when classes with binary values were involved. For instance, it was not possible to differ-

entiate between classes like *PawnerFlag* and *PersonRegisteredOffenderIndicator* since both classes contain values of either 1 or 0. We plan to address these drawbacks and explore data type similarity and ontological category to aid in mapping such classes.

A third factor that affects the performance of the tools is the tool-specific parameters that were used in this study. As mentioned before, Chimaera was used outside the Ontolingua environment [27] which prevented it from using structural matches. So, it may have performed better if we had used the Ontolingua environment for this study. In addition, both Chimaera and LOM used user-defined thresholds in their algorithms and suggested matches that were above the threshold. In this study, we used the default threshold that Chimaera suggested. Based on some experimentation, we found that LOM performed the best at a threshold value of 0.4, so this value was used in this study. However, further tweaking the thresholds and other input parameters/methods in these tools might lead to better performance.

4.5. Lessons learned

The tools developed in the ontology mapping domain have focused on four primary methods to determine mappings between various elements of ontologies. These methods include using NLP techniques, structural techniques, the use of secondary sources, and the use of

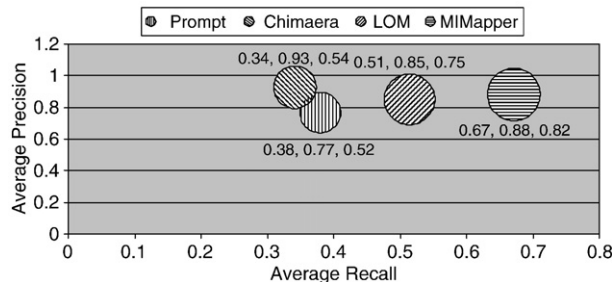


Fig. 4. Average recall and precision achieved by four ontology mapping tools in matching vehicle ontologies. The size of the bubble indicates the F-measure.

Table 8
Precision and recall in matching vehicle ontologies

	Chimaera		PROMPT		LOM		MIMapper	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
PD1 — PD2	0.78	0.78	0.8	0.89	1.0	0.67	1.0	0.89
PD1 — JX	1.0	0.13	1.0	0.13	0.75	0.38	0.8	0.5
PD2 — JX	1.0	0.13	0.5	0.13	0.8	0.5	0.83	0.63
Average	0.93	0.34	0.77	0.38	0.85	0.51	0.88	0.67
F-measure	0.54		0.52		0.75		0.82	

data instances. In this study, we aimed to establish which of these techniques (and not necessarily the tools themselves) is the most effective in real-world contexts. A difference between real-world ontology development and large-scale ontology development projects lies in the approach to ontology construction. Large scale ontology projects (e.g., SUMO [29], Cyc [22]) invest resources into the construction of an upper/middle ontology that provide structure to lower ontology elements. However, in many real-world scenarios, such a top-down approach to ontology construction may not be feasible. Most domain ontologies, such as those submitted to popular repositories like SemWeb-Central (<http://semwebcentral.org/>) and SchemaWeb (<http://www.schemaweb.info/>) have no link to upper ontologies and have limited structures. This is probably due to limited resources. In addition, it may be because small domain/problem specific ontologies are usually sufficient for most real-world applications. Even some of the largest domain ontologies (like Gene Ontology, <http://www.geneontology.org/> and Foundational Model of Anatomy, <http://sig.biostr.washington.edu/projects/fm/index.html>) do not have any link to upper ontologies, even though they have deep tree-like structures. However, development of such ontologies needs resources and time that may not be available in many real-world applications. Thus, tools that primarily use structural properties of ontologies may not perform well in such scenarios. Future development and use of mapping tools for such domain ontologies should focus on using other properties to determine mappings.

In addition, most real-world ontology development applications (including the one in this study) have data instances available. So, it would be beneficial if tools focus on using data instances in addition to other techniques to identify mappings. As can be seen from the experiments, data driven techniques are likely to perform better with smaller real-world ontologies. Thus, an important implication of this study is that the techniques used for ontology mapping are dependent upon the task. If the task is to map large domain ontologies with links to upper ontologies, then structural

and reasoning techniques would probably perform well. Whereas, if the task is to map smaller ontologies that have been developed with a bottom-up approach, then data driven techniques are likely to perform better due to the lack of structural elements.

In the public safety domain, since the U.S. Department of Justice is encouraging and requiring the use of GJXDM as a standard for information interchange, it would be beneficial to use GJXDM as a vocabulary to define public safety ontologies, so current mapping tools can achieve better accuracy and recall.

5. Conclusions and future directions

In this study, a survey of public safety information sharing initiatives was conducted and the ontological implications of the systems and standards were discussed. The performance of ontology mapping techniques was evaluated using real-world data representations used by public safety agencies. In addition, an information theory based technique called MIMapper was implemented that used WordNet and mutual information between data instances to map ontologies. It was found that three tools PROMPT, Chimaera, and LOM had average F-measures of 0.46, 0.49, and 0.68 when matching pairs of ontologies with the number of classes ranging from 13 to 73. MIMapper was found to perform better with an average F-measure of 0.84 in performing the same task. An analysis of the results showed that the tools that use secondary sources (like WordNet) and data instances to establish mappings between ontologies are likely to perform better in this application domain. In addition, since the U.S. Department of Justice is encouraging the use of GJXDM as a standard for information interchange, it would be beneficial to use GJXDM as a vocabulary to define public safety ontologies.

In the future, we plan to define ontologies for other important public safety related data elements and test mapping between them. We also plan to test MIMapper more thoroughly in other domains by using the standard platforms provided by ontology mapping conferences and comparing it against other tools that utilize data

instances. Enhancing MIMapper to include the use of structural and reasoning capabilities (by utilizing upper ontologies) for mapping is also planned.

Acknowledgement

This research was supported in part by: (1) the NSF DG program: “COPLINK Center: Information and Knowledge Management for Law Enforcement,” #9983304; (2) NSF KDD program: “COPLINK Border Safe Research and Testbed,” #9983304; (3) NSF ITR program: “COPLINK Center for Intelligence and Security Informatics Research- A Crime Data Mining Approach to Developing Border Safe Research,” #0326348; and (4) Department of Homeland Security (DHS) and Corporation for National Research Initiatives (CNRI) through the “BorderSafe” initiative, #2030002. We thank Tim Petersen and Daniel Casey of the Tucson Police Department for their contributions to this research.

References

- [1] J.L. Ambite, Y. Arens, E. Hovy, A. Philpot, L. Gravano, V. Hatzivassiloglou, J. Klavans, Simplifying data access: the energy data collection project, *IEEE Computer* 18 (1) (2001).
- [2] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web, *Scientific American* (2001).
- [3] A. Bouguettaya, M. Ouzzami, B. Medjahed, J. Cameron, Managing government databases, *Computer* 34 (2) (2001).
- [4] H. Chen, D. Zeng, H. Atabakhsh, W. Wyzga, J. Schroeder, COPLINK: managing law enforcement data and knowledge, *Communications of the ACM* 46 (1) (2003).
- [5] K.W. Chrucho, P. Hanks, Word association norms, mutual information, and lexicography, *Computational Linguistics* 16 (1) (1990).
- [6] A. Doan, J. Madhavan, P. Domingos, A.Y. Halevy, Learning to map between ontologies on the semantic web, Presented at 11th International World Wide Web Conference (WWW 2002), 2002.
- [7] T. Dunning, Accurate methods for the statistics of surprise and coincidence, *Computational Linguistics* 19 (1) (1993).
- [8] T. Edgington, B. Choi, K. Henson, T.S. Raghuram, A. Vinze, Adopting ontology to facilitate knowledge sharing, *Communications of the ACM* 47 (11) (2004).
- [9] R.M. Fano, *Transmission of Information*, MIT Press, Cambridge, MA, 1961.
- [10] C. Fellbaum, *WordNet: An Electronic Lexical Database*, The MIT Press, Cambridge, MA, 1998.
- [11] A. Firat, S. Madnick, Knowledge integration to ontological heterogeneity: challenges from financial information systems, Presented at Twenty-Third International Conference on Information Systems (ICIS), Barcelona, Spain, 2001.
- [12] A. Gal, G. Modica, H. Jamil, A. Eyal, Automatic ontology matching using application semantics, *AI Magazine* 26 (1) (2005).
- [13] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, Sweetening WordNet with DOLCE, *AI Magazine* 24 (3) (2003).
- [14] C.H. Goh, *Representing and Reasoning about Semantic Conflicts in Heterogeneous Information Sources*, MIT, 1997.
- [15] T.R. Gruber, A translation approach to portable ontologies, *Knowledge Acquisition* 5 (2) (1993).
- [16] M. Gruninger, J.B. Kopena, Semantic integration through invariants, *AI Magazine* 26 (1) (2005).
- [17] Y. Kalfoglou, M. Schorlemmer, Information-flow-based Ontology Mapping: On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE- Lecture Notes in Computer Science, Springer, 2002.
- [18] Y. Kalfoglou, M. Schorlemmer, Ontology mapping: the state of the art, *Knowledge Engineering Review* 18 (1) (2003).
- [19] S. Kaza, Y. Wang, H. Chen, Enhancing border security: mutual information analysis to identify suspect vehicles, *Decision Support Systems* (2006).
- [20] M.C. Kim, K.S. Choi, A comparison of collocation-based similarity measures in query expansion, *Information Processing and Management* 35 (1999).
- [21] W. Kim, J. Seo, Classifying schematic and data heterogeneity in multidatabase systems, *IEEE Computer* 24 (12) (1991).
- [22] D.B. Lenat, Cyc: a large-scale investment in knowledge infrastructure, *Communications of the ACM* 38 (11) (1995).
- [23] J. Li, LOM: a lexicon-based ontology mapping tool, Presented at Information Interpretation and Integration Conference (I3CON), 2004.
- [24] L. Loet, Similarity measures, author cocitation analysis, and information theory, *Journal of the American Society for Information Science and Technology* 56 (7) (2005).
- [25] D.M. Magerman, M.P. Marcus, Parsing a natural language using mutual information statistics, Presented at Eight National Conference on Artificial Intelligence, Menlo Park, CA, 1990.
- [26] S. March, A. Hevner, S. Ram, Research commentary: an agenda for information technology research in heterogeneous and distributed environments, *Information Systems Research* 11 (4) (2000).
- [27] D. McGuinness, R. Fikes, J. Rice, S. Wilder, An environment for merging and testing large ontologies, Presented at 17th International Conference on Principles of Knowledge Representation and Reasoning (KR 00), 2000.
- [28] P. Mitra, N.F. Noy, A.R. Jaiswal, OMEN: a probabilistic ontology mapping tool, Presented at 3rd International Conference on the Semantic Web, Hiroshima, Japan, 2004.
- [29] I. Niles, A. Pease, Towards a standard upper ontology, Presented at The Second International Conference on Formal Ontology in Information Systems, Ogunquit, Maine, 2001.
- [30] N.F. Noy, A. Doan, A.Y. Halevy, Semantic integration, *AI Magazine* 26 (1) (2005).
- [31] N.F. Noy, M.A. Musen, PROMPT: algorithm and tool for automated ontology merging and alignment, Presented at 17th national Conference on Artificial Intelligence (AAAI '00), 2000.
- [32] N.F. Noy, M.A. Musen, Anchor-PROMPT: using non-local context for semantic matching, Presented at Workshop on Ontologies and Information Sharing at the International Joint Conference on Artificial Intelligence (IJCAI), 2001.
- [33] N.F. Noy, M.A. Musen, Evaluating ontology-mapping tools: requirements and experience, Presented at 13th International Conference on Knowledge Engineering and Knowledge Management, Siguenza, Spain, 2002.
- [34] N.F. Noy, M.A. Musen, PROMTDIFF: a fixed-point algorithm for comparing ontology versions, Presented at 18th national Conference on Artificial Intelligence (AAAI '02), 2002.

- [35] D.E. O’Leary, Different firms, different ontologies, and no one best ontology, intelligent systems and their applications, *IEEE* 15 (5) (2000) [see also *IEEE Intelligent Systems*].
- [36] B. Omelayenko, Integration of product ontologies for B2B marketplaces: a preview, *SIGecom Exch.* 2 (1) (2001).
- [37] P. Pantel, A. Philpot, E. Hovy, Aligning database columns using mutual information, Presented at The 6th National Conference on Digital Government Research (dg.o), Atlanta, GA, 2005.
- [38] P. Pantel, A. Philpot, E. Hovy, Data alignment and integration [US government], *Computer* 38 (12) (2005).
- [39] J. Park, S. Ram, Information systems interoperability: what lies beneath? *ACM Transactions on Information Systems* 22 (4) (2004).
- [40] E. Rahm, P.A. Bernstein, A survey of approaches to automatic schema matching, *VLDB Journal* 10 (4) (2001).
- [41] G. Salton, M.J. McGill, *Introduction to Modern Information Retrieval*, McGraw Hill, 1983.
- [42] R. Stevens, P. Baker, S. Bechhofer, G. Ng, A. Jacoby, N.W. Paton, C.A. Goble, A. Brass, TAMBIS: transparent access to multiple bioinformatics information sources, *Bioinformatics* 16 (2) (2000).
- [43] G. Stumme, A. Maedche, Ontology merging for federated ontologies on the semantic web, Presented at International Workshop for Foundations of Models for Information Integration (FMII 2001), 2001.
- [44] P.-N. Tan, K. Vipin, J. Srivastava, Selecting the right objective measure for association analysis, *Information Systems* 29 (2004).
- [45] M. Uschold, M. Gruninger, Ontologies and semantics for seamless connectivity, *SIGMOD Record* 33 (4) (2004).
- [46] H. Wache, T. Voegelé, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, S. Hubner, Ontology-based integration of information — a survey of existing approaches, Presented at IJCAI Workshop: Ontologies and Information Sharing, Seattle, WA, 2001.



Siddharth Kaza is a doctoral student and research associate of the Artificial Intelligence Lab in the Department of Management Information Systems at the University of Arizona. He earned his M.S. in Computer Science from Central Michigan University and his B.Sc. in Mathematical Sciences from University of Delhi. His research interests include social network analysis, data mining and knowledge discovery, decision support systems, and security informatics.



Hsinchun Chen is McClelland Professor of Management Information Systems at the University of Arizona and Andersen Consulting Professor of the Year (1999). He received the BS degree from the National Chiao-Tung University in Taiwan, MBA from the SUNY Buffalo, and the PhD degree in Information Systems from New York University. Dr. Chen is a Fellow of IEEE and AAAS. He is author/editor of 13 books, 17 book chapters, 130 SCI journal articles, and 110 refereed conference articles covering Web computing, search engines, digital library, intelligence analysis, biomedical informatics, data/text/web mining, and knowledge management. He serves on ten editorial boards including: *ACM Transactions on Information Systems*, *ACM Journal on Educational Resources in Computing*, *IEEE Transactions on Intelligent Transportation Systems*, *IEEE Transactions on Systems, Man, and Cybernetics*, *Journal of the American Society for Information Science and Technology*, *Decision Support Systems*, and *International Journal on Digital Libraries*. Dr. Chen has served as an advisor for major NSF, DOJ, NLM, DOD, DHS, and other international research programs in digital library, digital government, medical informatics, and national security research. He is the founding director of the UA Artificial Intelligence Laboratory and Hoffman Ecommerce Laboratory.