Ontology-based Urban Data Exploration

Booma Sowkarthiga Balasubramani[1], Vivek R. Shivaprabhu[1], Smitha Krishnamurthy[1], Isabel F. Cruz[1], and Tanu Malik[2]

[1]ADVIS Lab, Department of Computer Science, University of Illinois at Chicago

[2]School of Computing, DePaul University

[1]{*bbalas3,vrevan2,skrish27,ifcruz*}@uic.edu; [2]*tanu*@cdm.depaul.edu

ABSTRACT

Cities are actively creating open data portals to enable predictive analytics of urban data. However, the large number of observable patterns that can be extracted as rules by techniques such as Association Rule Mining (ARM) makes the task of sifting through patterns a tedious and time-consuming task. In this paper, we explore the use of domain ontologies to: (i) filter and prune rules that are variations of a more general concept in the ontology, and (ii) replace groups of rules by a single general rule with the intent of downsizing the number of initial rules while preserving the semantics. We show how the combination of several methods reduces significantly the number of rules thus effectively allowing city administrators to use open data to generate patterns, use them for decision making, and better direct limited government resources.

1. INTRODUCTION

There is a tremendous explosion in the volume of urban data made available by open data initiatives around the world. In the US, the City of Chicago is a pioneer in this front. OpenGrid [1] was recently launched to enable the navigation of urban datasets, with its code available as open source to the programming community. The availability of data has opened up opportunities for data-driven analysis and decision making. Using data-driven analysis, policy makers and city administrators everywhere can help governments increase the efficiency of public services and in general improve the lives of citizens. However, cities are complex systems and analyzing the data they generate is a challenging task.

In the data-driven approach, a typical workflow consists of splitting data into subsets, aggregating, or mining data for analysis, and producing visual summaries of the resulting data. By iterating over this workflow, decision makers will selectively subset, aggregate, and mine input data to study how the outcome changes. For example, they may iterate by aggregating crime events over different regions of Chicago, producing a histogram at each step to acquire knowledge that burglary was a high-occurring event in the 9^{th} district during morning hours, whereas for another district, the unlawful possession of a handgun was also frequent but in the evening hours. There are cases where this process is ineffective, for example, when the number of distinct values associated with each attribute is large, thus aggregation and visualization could easily overwhelm analysts by presenting them with too many results.

Domain ontologies provide a hierarchy of concepts that are often not integrated with the databases but if available (or created) can serve as a useful means for the aggregation and the semantic exploration of data. In this paper, we describe how domain ontologies can be used for the semantic aggregation of predictive rules as obtained by Association Rule Mining (ARM) [5]. The use of ARM in urban data warehouses, in comparison to simple aggregation, is increasingly common as associative rules can potentially predict interesting spatial and temporal characteristics about urban behavior [18].

ARM generates rules of the form $X \rightarrow Y$ where X, the antecedent, and Y, the consequent, are sets of database attributes (e.g, *district, time of day, crime event*) that are distinct from each other, where the occurrence of X implies the occurrence of Y. ARM algorithms aim to mine rules that have high support (frequency of occurrence) and that have high confidence (ratio of the support of $X \cup Y$ over the support of X). In a large database, even though ARM algorithms prune rules that have low confidence and support there is the opportunity to perform further pruning. One way is by semantically aggregating the rules, for example those with either the same consequent and different antecedents or different consequents and the same antecedent. Ontologies can be particularly useful in this aggregation process because they encode generalization/specialization relationships among the database attributes, thus allowing for the determination of which rules are more general in comparison with other rules. Specialized rules that can be replaced by a more general rule are pruned.

We have developed an ontology-based explorer in the context of Plenario [9], which is a database and middleware supporting OpenGrid. Plenario's geospatial data warehouse for urban datasets integrates Socrata datasets [22] into a common spatial and temporal frame. The OpenGrid interface uses Plenario to support advanced queries for subsetting, aggregating, and visualizing incidents across the city. Through a Weka plugin, users can mine for association rules on query results. Our ontology-based exploration improves

UrbanGIS 16, October 31-November 03 2016, Burlingame, CA, USA

DOI: http://dx.doi.org/10.1145/3007540.3007550

upon the quality and number of rules obtained by association rule mining and is especially targeted toward decision makers who appreciate semantic exploration of the rule space. Adding this kind of functionality to Plenario follows the vision of GIVA [13], a semantic framework for geospatial and temporal data integration, visualization, and analytics.

The paper is organized as follows. Section 2 describes the preliminaries of Association Rule Mining and discusses ontologies for Urban Science. Section 3 describes the creation of the domain ontology for the crime dataset and the two categories of pruning methods: pre-processing and ontology-based. Section 4 describes the overall architecture and how our approach is integrated with the City of Chicago's urban database and visual interface provided by Plenario and OpenGrid, respectively. Section 5 contains the experimental results in terms of the number and percentage of the rules that are pruned, considering the different methods in isolation and then combined with one another. Section 6 describes related work. Finally, Section 7 points to directions for future work.

2. PRELIMINARIES

2.1 Association Rule Mining

The association rule mining (ARM) techniques are applied over databases described as $D = \{I, T\}$, where $I = \{I_1, I_2, \ldots, , I_p\}$ is a set of attributes (called items) and $T = \{t_1, t_2, \ldots, t_n\}$ is the transaction set. Each transaction $t_i = \{I_1, I_2, \ldots, I_m\}$ is a set of items, where $t_i \subset I$. An association rule R is an implication $X \rightarrow Y$ where X, the antecedent, and Y, the consequent, are two itemsets (that is, X, $Y \subset I$) and $X \cap Y = \emptyset$.

The typical example in association rule mining is of transactions representing all the items in shopping carts in a store. There are two important basic measures for association rules: support and confidence. Since the number of possible purchased items is large (corresponding to the store inventory) the focus is on those items that are often purchased (hence in the shopping cart) and especially those that are purchased at the same time, as captured by the mined rules. In particular we are interested in those items Y that are present in the shopping cart when items X are in there, denoted by the implication ($\rightarrow$).

The *support* of an association rule is defined as the percentage/fraction of transactions that contain $X \cup Y$ over the total number of transactions in the database. The count for each item is increased by one every time $X \cup Y$ is encountered in a different transaction. The *confidence* of an association rule is defined as the percentage/fraction of the number of transactions that contain $X \cup Y$ over the total number of records that contain X. Thresholds of support and confidence are set to drop those rules that are not so interesting or useful. The two thresholds are called minimal support and minimal confidence, respectively.

The purpose of association rule mining is to find those association rules that satisfy the predefined minimum support and confidence from a given database [5]. The problem is usually decomposed into two subproblems. One is the frequent itemset problem, that is, to find those itemsets whose occurrences exceed a predefined threshold in the database. The second problem is to generate association rules from those large itemsets that satisfy the set thresholds. The Apriori algorithm solves the two sub-problems by generating candidate itemsets and scanning the database to check the actual support count of the corresponding itemsets. In each pass only those candidate itemsets that include the same specified number of items are generated and checked. The candidate k-itemsets are generated after the (k-1)th passes over the database by joining the frequent k-1 itemsets. All the candidate k-itemsets are pruned out after checking their sub (k-1)-itemsets, if any of its sub (k-1)-itemsets is not in the list of frequent (k-1)-itemsets. The Apriori property states that every sub (k-1)-itemsets of the frequent k-itemsets must be frequent.

2.2 Ontologies for Urban Science

An ontology contains a set of concepts (classes or properties) and a set of relations defined over those concepts. The most important relation that defines a directed acyclic graph is subsumption (or specialization/generalization or is-a).

While there are several ontologies that have been developed in the spatial context for urban science [19], the existence of ontologies that combine spatial and urban concepts such as crime, 911/311 calls, license permits, and finance are not well-developed. A possibility is to construct ontologies specific for each urban concept manually by relying on expert knowledge, by inspection of the dataset, or by reusing previously defined ontologies (or taxonomies). In the next section we describe the process of creating an ontology for the crime dataset, but we are in the process of building ontologies for other urban datasets as well.

3. ONTOLOGY-BASED ASSOCIATION RULE PRUNING

In this section we present how ontologies can be used to prune rules. While our pruning methods and consideration of ontologies is general, to illustrate the policies for pruning, we focus on the crime dataset and associated ontologies. While the former is available from the City of Chicago, the crime ontology is not, therefore we first show how to construct it.

Our focus on crime is based on the fact that Chicago has a crime rate that is significantly higher than the national average. At 41 crimes per one thousand residents, reducing crime in Chicago is a top priority for the city. On average there is a 1 in 25 chance of becoming a victim of crime in Chicago; however, the number of crime events vary by neighborhood and thus the probability of being a victim of crime is non-uniform across neighborhoods. Our objective is to use ontology-based association rule mining to discover types of crimes in different neighborhoods and the co-occurrence of related events.

3.1 Crime Ontology

To create our crime ontology, we utilize two existing classifications of crime, one as available from the Chicago Police Department (CPD), and the other as available from the Federal Bureau of Investigation (FBI). First we mapped the two classification schemes to establish correspondences between them and then enhanced the result with categories that were not classified but are included in the description of a crime event. This helps us to create a multi-level hierarchy of concepts. A fragment of this multi-level crime ontology is shown in Figure 1.

In the crime dataset, the CPD [2] classifies 401 types of crimes, under 34 different categories. Each of the crime type

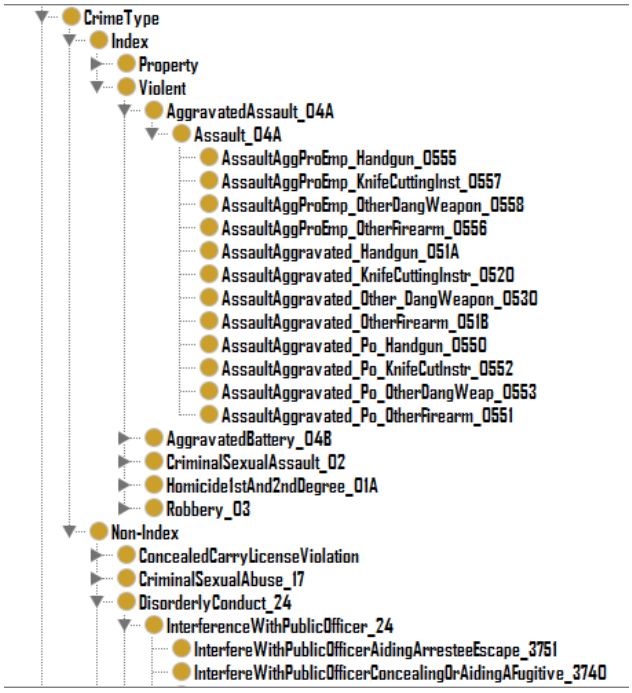

Figure 1: A fragment of the crime ontology.

is associated with the crime classification codes known as Illinois Uniform Crime Reporting (IUCR) Codes. The crime categories are associated with the FBI National Incident-Based Reporting System (NIBRS) code [4]. Apart from the 34 categories of crime described by CPD, we have also added other categories such as Aggravated, Assault, Attempt, Computer Related, Deceptive Practice, Kidnapping, Non-Aggravated, Offense, Simple, Theft, and Weapon. This categorization enabled us to identify similar types of crimes from the 34 different categories drafted by the CPD. For example, the crime type *DECEPTIVE PRACTICE COMPUTER FRAUD* under the category *Fraud*, and the crime type *STALKING CYBERSTALKING* under the category *Simple Assault*, will be grouped under the class *ComputerRelated*. This is implemented in the ontology using the property restrictions functionality, as shown in Figure 2. The description of the crime types is appended with the respective IUCR code to uniquely identify the crime types, as there are differences in the classification between the IUCR dataset and the CPD Clearmap Crime Summary.

The crime ontology also includes spatial and temporal entities, to model information such as the address, location description, and the time at which the crime occurred. The temporal entity is further subdivided into morning, noon, evening, and night, so that the based on the timestamp of crime occurrence in the database, the crime incident can be categorized into any of these groups. This helps in identifying the kind of crimes occurring at specific times of the day. For example, rules of the form [*Time of Day=MORNING*] $\rightarrow$ [*Assault=FALSE*] can be obtained from this categorization. The spatial and temporal entities also enable us to roll up or down to obtain different granularities.

3.2 Pre-processing Association Rules

To be able to prune rules based on ontologies, we must first pre-process the association rule list to remove combinatorial variations that lead to unnecessary passes during the semantic grouping phase. This pre-processing is based purely on the syntactic aspect of the association rules generated using the Apriori algorithm. The pre-processing of association rules is based on the following four cases.

Same Antecedent Condition. Given two or more rules with the same antecedent, but with different consequents, rules can be grouped together based on the antecedent. For example, consider the rules:

R_1: [Description=FORGERY] $\rightarrow$ [Aggravated=FALSE]
R_2: [Description=FORGERY] $\rightarrow$ [Assault=FALSE]
R_3: [Description=FORGERY] $\rightarrow$ [Weapon=NONE]

In all these three rules, the antecedent is the same. Thus, there is some potential to group these rules based on the antecedent. In this case R_1, R_2, and R_3 can be replaced by the following refined rule R_4.

R_4: [Description=FORGERY] $\rightarrow$ [Aggravated=FALSE, Assault=FALSE, Weapon=NONE]

Specialization Condition. If two rules contain the same consequent, and the antecedent of the second rule contains the antecedent of the first rule plus additional items from the dataset, then the second rule can be eliminated. This stems from the second rule being the specialization of the first rule and therefore deemed redundant. In general, for rules R_1 and R_2, where $R_1 : A \wedge B \rightarrow C$ and $R_2 : A \wedge B \wedge X \rightarrow C$, rule R_2 is redundant. For example:

R_1: [Month=7, Time of Day=LATE, Description= OVER $500] $\rightarrow$ [Aggravated=false]
R_2: [Month=7, Time of Day=LATE, Description= OVER $500, Assault=FALSE, Weapon=NONE] $\rightarrow$ [Aggravated=FALSE]

R_2 can be removed since it is a specialized version of R_1.

Generalization Condition. If two rules have the same antecedent, and the consequent of the second rule contains the consequent of the first rule plus additional items from the dataset, then the first rule can be eliminated. This is based on the fact that the second rule contains more information because of the additional items. Consider two rules R_1 and R_2, where $R_1 : A \rightarrow B$ and $R_2 : A \rightarrow B \wedge C$, then rule R_1 can be eliminated. For example:

R_1: [Month=4, Description=TO VEHICLE] $\rightarrow$ [Aggravated=FALSE]
R_2: [Month=4, Description=TO VEHICLE] $\rightarrow$ [Aggravated=FALSE, Assault=FALSE]

In this example, R_1 can be removed because R_2 provides more information than R_1.

Implication Condition. For each attribute in the dataset, the Apriori algorithm takes a Cartesian product of the distinct set of values that satisfy the filter criteria. This results in a large number of rules with different combinations of the attribute and values in the antecedent and consequent of the rule. Such rules can be refined to form a single rule. Consider the rules $R_1, R_2, \ldots, R_9$, where $R_1 : A \wedge X \rightarrow Y$, $R_2 : A \wedge X \rightarrow Z$, $R_3 : A \wedge Y \rightarrow X$, $R_4 : A \wedge Y \rightarrow Z$, $R_5 : A \wedge Z \rightarrow X$, $R_6 : A \wedge Z \rightarrow Y$, $R_7 : A \rightarrow X \wedge Y$, $R_8 : A \rightarrow Y \wedge Z$, and $R_9 : A \rightarrow X \wedge Z$. These rules can be replaced by a single rule $R_{10} : A \rightarrow X \wedge Y \wedge Z$. For example:

R_1: [Month=7, Time of Day=LATE, Description= OVER $500, Assault=FALSE] $\rightarrow$ [Aggravated=FALSE]
R_2: [Month=7, Time of Day=LATE, Description= OVER $500, Assault=FALSE] $\rightarrow$ [Weapon=NONE]
R_3: [Month=7, Time of Day=LATE, Description=

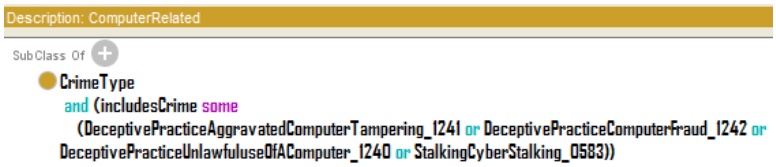

Figure 2: Example of additional crime categories.

OVER $500, Aggravated=FALSE] → [Assault=FALSE]
R_4: [Month=7, Time of Day=LATE, Description=
OVER $500, Aggravated=FALSE] → [Weapon=NONE]
R_5: [Month=7, Time of Day=LATE, Description=
OVER $500, Weapon=NONE] → [Assault=FALSE]
R_6: [Month=7, Time of Day=LATE, Description=
OVER $500, Weapon=NONE] → [Aggravated=FALSE]
R_7: [Month=7, Time of Day=LATE, Description=
OVER $500] → [Aggravated=FALSE]
R_8: [Month=7, Time of Day=LATE, Description=
OVER $500] → [Aggravated=FALSE, Weapon=NONE]
R_9: [Month=7, Time of Day=LATE, Description=
OVER $500] → [Assault=FALSE, Weapon=NONE]
R_{10}: [Month=7, Time of Day=LATE, Description=
OVER $500] → [Aggravated=FALSE, Assault=FALSE,
Weapon=NONE].

3.3 Ontology-based Pruning

Ontology-based pruning considers the domain ontology and the pre-processed association rules as inputs. There are two kinds of pruning that we can achieve with ontologies: (i) identification of facts, (ii) identification of parent-child relationships. We describe the two associated pruning cases.

Fact-based Pruning. This condition identifies the rules that contain only one item in the antecedent and one item in the consequent, and looks up the ontology for a parent-child relationship between the item in the antecedent and the item in the consequent. If the antecedent is a parent or child of the consequent, then this rule will be eliminated based on the fact that this is a direct inference from the knowledge base. For example, consider the rule:

R_1: [Description=ARMED: HANDGUN] → [Weapon=
HANDGUN]

In the domain ontology, the class *HANDGUN* is a child of class *Weapon* and the crime type *ARMED: HANDGUN* belongs to the class *HANDGUN*. Since this subclass relationship exists in the knowledge base, this rule can be pruned.

We also prune rules with only one item in the consequent, and more than one item in the antecedent, one of which is a parent or child (in the domain ontology) of the consequent. In general, rules of the form $R_1 : A \wedge X' \rightarrow X$, where item X is the parent (or child) of item X' in the domain ontology can be pruned. For example:

R_1: [Time of Day=MORNING, Description=
ARMED: HANDGUN] → [Weapon=HANDGUN]

will be pruned, because the class *ARMED: HANDGUN* is a child class of *HANDGUN*, which is again a child class of *Weapon*.

Parent-Child Subsumption. If P is a class in the domain ontology, and $C_1, C_2, \ldots, C_n$, where $n > 1$, are the subclasses of P, the set of n rules which have the same consequent and whose antecedents contain one of the subclasses C_k (where $k = 1, \ldots, n$) of P, can be refined. That is, the set of rules of the form $R_1 : A \wedge C_1 \rightarrow X$, $R_2 : A \wedge C_2 \rightarrow X, \ldots, R_n : A \wedge C_n \rightarrow X$, can be refined to the higher-level rule $A \wedge P \rightarrow X$. Here the primary issue is to determine the *coverage*, that is if we require for all n children to be present in rules to be refined as a higher-level rule, or if only a majority of those rules with high support is necessary. We describe a sound statistical procedure to identify when sub-rules can be refined to a higher level rule based on high support of rules that provide coverage and a low standard deviation.

We consider a general rule, $R_i : C \rightarrow Y$, and a set, S_i, of specialized rules, $r_{i,j}$ of the form $C_i \rightarrow Y$, then we define the coverage of the general rule by the support of the antecedent of the specialized rules as follows [17]:

$$CRg(R_i, S_i) = \frac{\sum_{j=1,\ldots,n} support(ant(r_{ij}))}{support(ant(R_i))} \quad (1)$$

The larger the value of $CRg(R_i, S_i)$, the larger the representability of the instances covered by the more specialized rules in relation to those instances covered by the more general rule. We use also consider another measure [17]:

$$TRg(R_i, S_i) = \frac{\mu}{\sigma} \quad (2)$$

where x_j is the value of the support of rule r_{ij} and $\mu = \frac{\sum_n^{j=1} x_j}{n}$ and $\sigma = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j - \mu)^2}$. The larger the value of $TRg(R_i, S_i)$, the closer are the values x_j (support of the specialized rules) to their average μ.

We consider two threshold values CRg_{min} and TRg_{min} that are specified by the domain expert, against which the CRg and TRg values are compared. Thus, when $CRg \geq CRg_{min}$ the specialize rules can be discarded because they are represented by the more general rule. Otherwise, the discarded rule is the more general rule. A similar comparison is made between TRg and TRg_{min}. Both values are considered together, in the sense that a group of rules is replaced by a more general rule only if $CRg \geq CRg_{min}$ and $TRg \geq TRg_{min}$.

We present next two scenarios that show the pruning of specialized and general rules. Based on domain knowledge, CRg_{min} and TRg_{min} are set to 2.0 and 0.7, respectively.

Scenario 1: Pruning Specialized Rules.

R_1: [Time of Day=LATE, Description=AGGRAVATED: HAND-
GUN] → [Assault=FALSE]
R_2: [Time of Day=LATE, Description=ARMED: HANDGUN] →
[Assault=FALSE]
R_3: [Time of Day=LATE, Description=UNLAWFUL POSS OF
HANDGUN] → [Assault=FALSE]
R_4: [Time of Day=LATE, Weapon=HANDGUN] →
[Assault=FALSE]

R_1, R_2, and R_3 are specialized rules and R_4 is the general rule. All the rules in this example have the same consequent.

By looking up the ontology, we conclude that items *AGGRAVATED: HANDGUN*, *ARMED: HANDGUN*, and *UNLAWFUL POSS OF HANDGUN* are children of (*HANDGUN*, which is a subclass of *Weapon*). *TRg* for this set of rules is 11.9765, and *CRg* is 0.9503. Since, $TRg > TRg_{min}$, and $CRg > CRg_{min}$, we prune the specialized rules R_1, R_2, and R_3, and keep the general rule R_4.

Scenario 2: Pruning General Rules.

R_5: [Time of Day=LATE, Description= POSS: CANNABIS 30GMS OR LESS, Arrest=TRUE, Ward=24] → [Domestic=FALSE, Assault=FALSE]

R_6: [Time of Day=LATE, Description= POSS: HEROIN(WHITE), Arrest=TRUE, Ward=24] → [Domestic=FALSE, Assault=FALSE]

R_7: [Time of Day=LATE, Arrest=TRUE, Ward=24, Aggravated=FALSE] → [Domestic=FALSE, Assault=FALSE]

R_5, and R_6 are specialized rules and R_7 is the general rule and all have the same consequent. By looking up the ontology, we conclude that items *POSS: CANNABIS 30GMS OR LESS*, and *POSS: HEROIN(WHITE)* belong to the same parent (*NON-AGGRAVATED*, which is nothing but *Aggravated=FALSE*). *TRg* for this set of rules is 5.0373, and *CRg* is 0.3946. In this case, $TRg > TRg_{min}$, but $CRg < CRg_{min}$. Therefore, we prune the general rule R_7, and keep the specialized rules R_5 and R_6.

4. SYSTEM ARCHITECTURE

The overall architecture of our system is shown in Figure 3. Spatial datasets in any format (tabular, shapefiles, KML, CSV, JSON, and geoJSON) are input into Plenario which is essentially a PostGIS database with an efficient ingest system and middleware to support complex geospatial queries. The database is available through a RESTful interface for spatial querying. Initially spatial data is selected from the database based on a region of interest. Given the datasets selected, the corresponding domain ontology is also selected. Each row in the selected dataset is considered as a transaction for the ARM module. For ARM, we use the Apriori algorithm from Weka. The post-processing module is implemented in Java and takes as input a domain ontology and user specified filtering thresholds as described in Section 3.

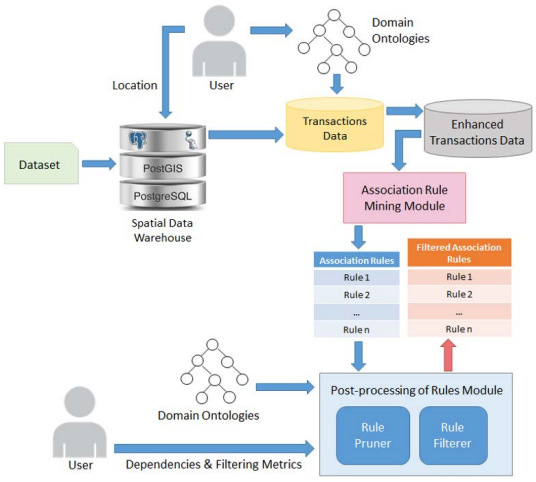

Figure 3: Ontology-based ARM architecture.

5. EXPERIMENTAL RESULTS

In this section, we analyze the Chicago crime data [3] for the year 2014 to obtain rules using the visual interface in Figure 4.

The following 10 features have been extracted from the dataset: *Month*, *Time of Day*, *Street*, *Description*, *Arrest*, *Domestic*, *Ward*, *Aggravated*, *Assault*, *Weapon*. All the features except *Weapon* have a fill rate of 100%. *Weapon* is populated in 7% of the instances. Five of these are features that are used as is. The features *Month* and *Time of Day* are derived from the *Date Timestamp* in the instances. ARM rules are generated using the Weka API. Since the ARM algorithm works only with nominal attributes, all numerical attributes are transformed into nominal attributes using the domain ontologies by checking the corresponding class to which the data value belongs, and transforming the numerical data into the class name. For example, if the subclass *early morning* is defined as 12am-6am in the ontology, then any time in that interval will be assigned *early morning*. A total of 64137 rules were generated from 274322 records in the Crime dataset. These rules were pruned using the four pruning methods (also listed in Table 1) discussed in Section 3. The ontology, as a knowledge repository, maps the *Description* to True/False values for *Aggravated* and *Assault*, and as nominal values for *Weapon*. The ontology enables parent-child relationships from *Assault*, *Aggravated* and *Weapon* (parents) to *Description* (child). These maps are used in Methods 3 and 4 to identify related features.

Table 1 shows the pruning results when each methods executed independently of one another. That is, each method is executed on the initial set of 64137 rules, independently of the others. The table lists the method identification numbers, the principle behind that method, total number of rules pruned by applying that method, and the pruning percentage. The pruning percentage gives the ratio between the number of rules pruned and the total number of rules generated. In this scenario, Method 1 gives the highest pruning percentage. This is because of the combinatorial explosion in the items present in the rules generated using ARM.

It is also to be observed that pruning based on ontologies is not significant percentage, however, the objective is to not compare methods but to use a combination of ARM-based and ontology-based methods to get a better pruning percentage/result. The ontology-based methods are also dataset specific. Crime, even though an important dataset to consider as a first, has very little correlation between its features (for instance, Methods 3 and 4 do not seem as dramatic as Method 1). However, if there are several fields with hierarchical relationships, as in payroll datasets, we will get higher pruning percentages.

The right blue frame in Figure 5 shows the visual interface that enables user interaction to choose ontological concepts and observe how the rules are pruned based on the thresholds they specify. Then by choosing one or more rules, users will be able to see various crime scenarios as displayed on the interface map (left map frame in Figure 5).

Figure 6 shows the pruning percentage and the number of rules pruned by each method, when they are executed in different orders. The rule pruning methods were executed in all 24 (that is, 4!) orders. Though each method had pruned different number of rules when executed in different order, more than one order of the methods resulted in the same pruning percentage. For simplicity, we chose to depict only

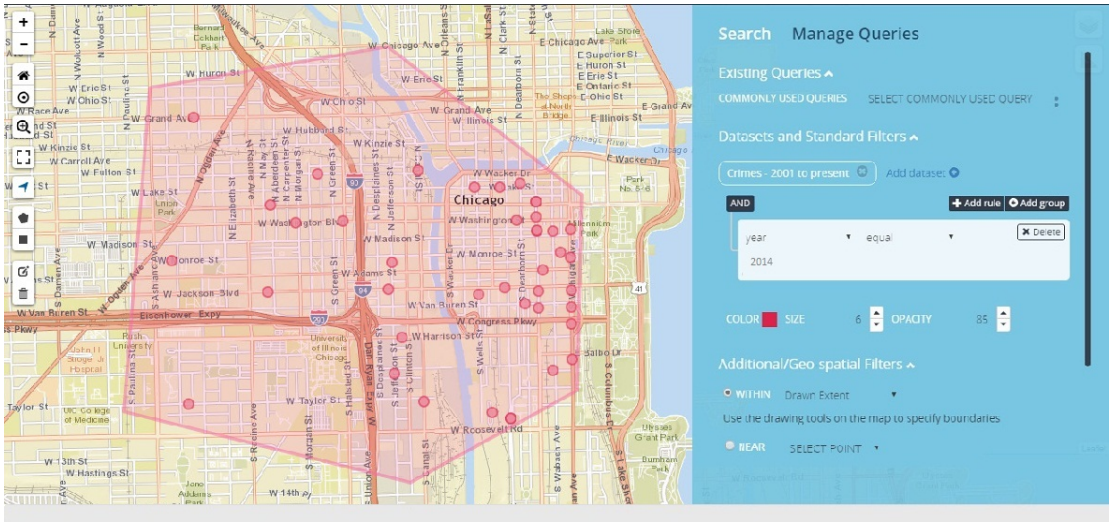

Figure 4: Crime data for a specified region in 2014.

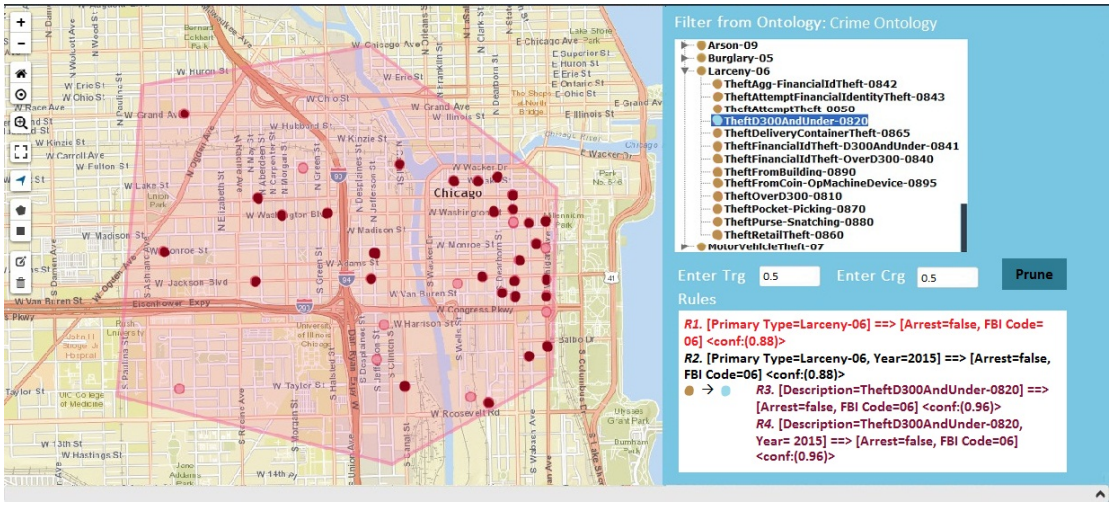

Figure 5: Choosing ontological concepts for rule pruning in crime scenarios.

once each of the distinct pruning percentage obtained, even if several different orders gave rise to those values. From Figure 6, it is evident that when Method 1 is executed first, it prunes a larger number of rules from the initial set of rules, and hence the pruning percentage is higher. Also, for this dataset, when Method 1 is executed before Method 3, Method 1 covers all the rules that are to be removed by Method 3. Hence, no rules are pruned by Method 3. Similar observations can be made for the other orders in which the methods are executed.

6. RELATED WORK

While ontologies have been used extensively for data integration [16], for data mining and in particular for association rule mining, their use has been much less frequent. Association rule mining considers the frequency of co-occurrence of items in transactions to extract patterns. However, the meaning of each item nor the semantic relationship to other items are taken into consideration. In this work, we have considered the introduction of semantic content by using ontologies to improve the quality of the data mining results.

Approaches that introduce semantic content to improve the quality of data mining results can be divided into two broad areas of work: one, using ontologies to guide the space of data mining results [10,20,23] and the other using ontologies to improve the interestingness of rules [8,11,17,21]. Our work is in the latter category as we would like to reduce the search space of the rules that decision makers have to navigate. Ontologies can be used to reduce this search space either by pruning the rules and generalizing them [17, 21] or by abstracting them using orthogonal measures as correlation mappings and their usage between abstracted terms and refined terms [11]. In this paper, we have focused on pruning.

In a previously mentioned pruning approach [17], the redundancy between the rules is categorized into four types, depending on the existence of determinant attributes, which give rise to aggregating attributes, as is the case with *Date* and dependent attributes as is the case of *Month* in the *Date*

Table 1: Pruning results for the 64137 initial rules run independently.

Method	Principle	# Rules Pruned	% Rules Pruned
1	Remove $A \wedge C \rightarrow B$, if $A \rightarrow B$ exists.	61700	96.20%
2	Remove $A \rightarrow B$, if $A \rightarrow B \wedge C$ exists.	35088	54.71%
3	Remove $A \wedge Child \rightarrow Parent$.	17124	26.70%
4	Remove based on CRg and TRg.	2769	4.32%

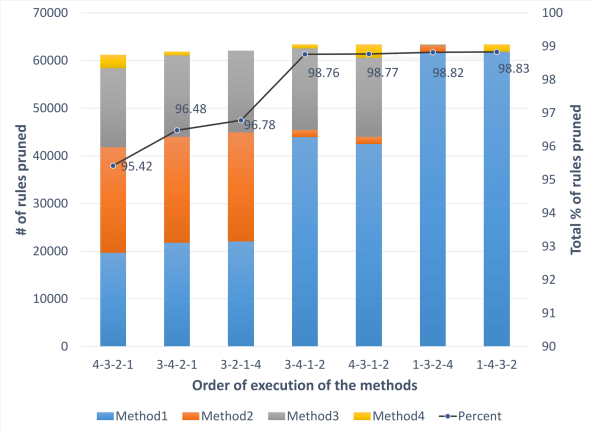

Figure 6: Pruning percentage when the methods are executed in different orders.

attribute in the antecedent or in the consequent of the rules. They use ontologies to post-process the rules to eliminate those redundancies. Our work addresses two of those redundancy types by eliminating them using Association Rule Mining (ARM) based pruning as mentioned in Section 3. We eliminate the rules causing such redundancies using either a domain ontology or using other pruning principles that do not need to use the domain ontology because our principles apply even if the items in the antecedent and the consequent are not related to each other through a parent-child relationship in the ontology. The introduction of these principles increases the pruning percentage of the rules.

Another related approach [21] uses domain ontologies to strengthen the integration of user knowledge and to increase the interestingness of the rules in the post-processing phase. This is done by having two hierarchies in the ontology, which is basically the regrouping of all concepts created using necessary and sufficient conditions over other concepts. In our work, each crime type is classified under various categories based on the Crime database, NIBRS code, and our customized classes in the ontology. This classification is purely based only on the description of the crime. For example, the crime type *Aggravated: Handgun* is classified under *Battery* according to the Crime database, under *Violent* crime based on the NIBRS code, and under the *Aggravated* category and *Weapons* category according to our customized classes.

7. CONCLUSIONS AND FUTURE WORK

We demonstrated the use of domain ontologies to reduce the number of rules generated by ARM. However, it is also the case that one needs a domain ontology for that domain. Therefore, we developed a crime ontology because an ontology was not available.

Our experimental evaluation demonstrates the reduction of the number of rules, which facilitates the identification of predictive patterns in the City of Chicago datasets. We explored the use of subsumption but other inference mechanisms can be explored in the future such as creating new classes from different branches of the ontology, for example, a class of all the crimes committed using a *weapon* or of those that involve a *child*. Most crimes co-occur with *Assault* being *False*. An exception involves a *child*, for example, *sex assault of child family member* is linked to *Assault* being *True*. However, due to its scarcity, this knowledge will be likely filtered out due to the minimal support value chosen. Capturing rare items with high confidence may be needed.

With this work being in the context of OpenGrid and Plenario, which are supported by the City of Chicago, any such determinations can be made in consultation with the policy makers and city administrators. Another point where consultation would be important is in the determination of CRg_{min} and of TRg_{min}. For example, for the latter they should be able to adjust the value by looking at "representative" groups of rules to determine a value that makes sense for all the groups of rules.

Human-in-the-loop computation is a current trend, which may combine the opinion of one or more experts based on representative output samples [7, 14, 15]. As soon as humans are involved in the decision process, the presentation and visualization of the results is paramount, including the implementation of advanced database visualization techniques [12] and the support for a visual analytics process [6].

8. ACKNOWLEDGEMENTS

We thank Tom Schenk (CDO of the City of Chicago) and his team for many discussions. Work at UIC was par-

tially supported by a grant from the Bloomberg Philanthropies and by NSF awards CNS–1646395, III–1618126, CCF–1331800, and III–1213013. Research at DePaul University is funded by a seed grant from the City of Chicago.

9. REFERENCES

[1] OpenGrid, (Accessed January 20, 2016). http://opengrid.io.

[2] Chicago Police Department - Illinois Uniform Crime Reporting (IUCR) Codes, (Accessed September 4, 2016). https://data.cityofchicago.org/Public-Safety/Chicago-Police-Department-Illinois-Uniform-Crime-R/c7ck-438e/data.

[3] City of Chicago - Crime Dataset, (Accessed September 4, 2016). https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data.

[4] Crime Type Categories: Definition & Description, (Accessed September 4, 2016). http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html.

[5] Agrawal, R., Imieliński, T., and Swami, A. Mining Association Rules Between Sets of Items in Large Databases. In *ACM SIGMOD International Conference on Management of Data* (1993), pp. 207–216.

[6] Aurisano, J., Nanavaty, A., and Cruz, I. F. Visual Analytics for Ontology Matching Using Multi-Linked Views. In *ISWC International Workshop on Visualizations and User Interfaces for Ontologies and Linked Data (Voila!)* (2015), vol. 1456 of *CEUR Workshop Proceedings*, pp. 25–36.

[7] Balasubramani, B. S., Taheri, A., and Cruz, I. F. User Involvement in Ontology Matching Using an Online Active Learning Approach. In *ISWC International Workshop on Ontology Matching (OM)* (2015), vol. 1545 of *CEUR Workshop Proceedings*, pp. 45–49.

[8] Bayardo Jr., R. J., and Agrawal, R. Mining the Most Interesting Rules. In *SIGKDD International Conference on Knowledge Discovery and Data Mining* (1999), ACM, pp. 145–154.

[9] Catlett, C., Malik, T., Goldstein, B., Giuffrida, J., Shao, Y., Panella, A., Eder, D., van Zanten, E., Mitchum, R., Thaler, S., and Foster, I. T. Plenario: An Open Data Discovery and Exploration Platform for Urban Science. *IEEE Data Eng. Bull. 37*, 4 (2014), pp. 27–42.

[10] Charest, M., and Delisle, S. Ontology-guided Intelligent Data Mining Assistance: Combining Declarative and Procedural Knowledge. In *Artificial Intelligence and Soft Computing* (2006), pp. 9–14.

[11] Chen, P., Verma, R. M., Meininger, J. C., and Chan, W. Semantic Analysis of Association Rules. In *International Florida Artificial Intelligence Research Society Conference (FLAIRS)* (2008), pp. 270–275.

[12] Cruz, I. F., Averbuch, M., Lucas, W. T., Radzyminski, M., and Zhang, K. Delaunay: A Database Visualization System. In *ACM SIGMOD International Conference on Management of Data* (1997), pp. 510–513.

[13] Cruz, I. F., Ganesh, V. R., Caletti, C., and Reddy, P. GIVA: A Semantic Framework for Geospatial and Temporal Data Integration, Visualization, and Analytics. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2013), pp. 544–547.

[14] Cruz, I. F., Palmonari, M., Loprete, F., Stroe, C., and Taheri, A. Quality-Based Model for Effective and Robust Multi-User Pay-As-You-Go Ontology Matching. In *Semantic Web Journal*, vol. 7. IOS Press, 2016, pp. 463–479.

[15] Cruz, I. F., Stroe, C., and Palmonari, M. Interactive User Feedback in Ontology Matching Using Signature Vectors. In *IEEE International Conference on Data Engineering (ICDE)* (2012), pp. 1321–1324.

[16] Cruz, I. F., and Xiao, H. Ontology Driven Data Integration in Heterogeneous Networks. In *Complex Systems in Knowledge-based Environments*, A. Tolk and L. Jain, Eds. Springer, 2009, pp. 75–97.

[17] Ferraz, I. N., and Garcia, A. C. B. Ontology in Association Rules. *SpringerPlus 2*, 1 (2013), p. 452.

[18] Koperski, K., and Han, J. Discovery of Spatial Association Rules in Geographic Information Databases. In *International Symposium on Advances in Spatial Databases (SSD)* (1995), pp. 47–66.

[19] Laurini, R. Urban Ontologies. http://liris.cnrs.fr/robert.laurini/resact/urban-ontologies.pdf.

[20] Manda, P., McCarthy, F., and Bridges, S. M. Interestingness Measures and Strategies for Mining Multi-ontology Multi-level Association Rules from Gene Ontology Annotations for the Discovery of New GO Relationships. *Journal of Biomedical Informatics 46*, 5 (2013), pp. 849–856.

[21] Marinica, C., Guillet, F., and Briand, H. Post-processing of Discovered Association Rules Using Ontologies. In *Workshops of the International Conference on Data Mining (ICDM)* (2008), IEEE, pp. 126–133.

[22] Socrata Foundation. Socrata: The Open Data Platform for Digital Government. http://socrata.com.

[23] Srikant, R., and Agrawal, R. Mining Generalized Association Rules. In *International Conference on Very Large Data Bases (VLDB)* (1995), pp. 407–419.