IMT Institute for Advanced Studies, Lucca

Lucca, Italy

Generalized discrimination discovery on semi-structured data supported by ontology

PhD Program in Computer Science

XXIII Cycle

By

Luong Thanh Binh

2011

The dissertation of Luong Thanh Binh is approved.

Program Coordinator: Prof. Ugo Montanari, IMT Institute for Advanced Studies, Lucca - Lucca. Dipartimento di Informatica, Università di Pisa.

Supervisor: Prof. Franco Turini, Dipartimento di Informatica, Università di Pisa.

Supervisor: Prof. Salvatore Ruggieri, Dipartimento di Informatica, Università di Pisa.

Tutor: Prof. Leonardo Badia, IMT Institute for Advanced Studies, Lucca - Lucca.

The dissertation of Luong Thanh Binh has been reviewed by:

Assistant Prof. Toon Calders, Faculty of Math and Computer Science, Eindhoven University of Technology.

Prof. Stan Matwin, School of Electrical Engineering and Computer Science, University of Ottawa.

IMT Institute for Advanced Studies, Lucca

2011

To my daddy, Luong Manh Ba

Contents

List of Figures

List of Tables

Acknowledgements

It is never sufficient to express my gratitude to my supervisor, Prof. Franco Turini, for his limitless enthusiasm, kindness, patience, as well as his academic experience. Throughout my research period, he provided me a lot of invaluable ideas, and insightful criticisms, which aided the growth of my research ability in innumerable ways. Moreover, he showed an admirable forbearance when revising the English in my manuscript. I would have been lost in the "engineering style" without him. I would also like to kindly thank Prof. Salvatore Ruggieri for his brilliant ideas, and for his acceptance of sharing ideas in my thesis, particularly in the kNN Situation Testing part.

I am indebted to many friends of mine for providing a friendly and warm environment in which I learn and grow. Barbara Furletti, and Bellandi Andrea patiently listened to my confusing presentations, gave me precious discussions, and shared with me their experiences. Moreover, thank you for helping me adapt to the new environment. To Davide Bacciu, I greatly appreciate our talk at the very beginning of my PhD, it assisted me much in specifying my approach.

A big thank you to "bimba" Rebecca Ong, it is so nice of her to be with me and offer help whenever I need.

It is impossible to not mention Michele Budinich, my true friend at IMT, for helping me get through many difficult times, and for all the "expensive" advices, fun, and "shipping" he provided.

Further, I would like to extend my thanks to the secretaries and librarians of the IMT Lucca for their untiring efforts in

supporting me in document works. Serena Argentieri, and Barbara Iacobino deserve special mention. I know that I took a lot of time from them and gave back not a few disturbances.

Especially, I wish to thank my parents and my brother, who followed and endlessly encouraged me in all my ways. From whom I was inspired to achieve further educations and to approach new cultures.

Lastly, my husband Duong is the one who always stands by me. He is the source of my strength, my trust, my joy, and my serene sky. I am nothing without him, and I thank him.

Vita

September 27, 1982	Hanoi, Vietnam
2005	Bachelor of Engineering in Information Technology Grade: Excellent Hanoi University of Technology, Vietnam
2007	Master in Information Technology Hanoi University of Technology Hanoi, Vietnam

Publications

1. Binh T. Luong, F. Turini. "Association analysis of semi-structured data for discrimination discovery in business," DMIN 2010. Las Vegas, Nevada, USA, 2010.
2. Binh T. Luong, F. Turini, S. Ruggieri. "k-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention," Technical report. Dipartimento di Informatica, Universit di Pisa., 2011.
3. Binh T. Luong, F. Turini. "Discrimination Association Analysis on Semi-structured Data," European Conference on Data Mining (ECDM 2011) - IADIS Multi Conference on Computer Science and Information Systems (MCCSIS 2011), Rome, Italy, 2011.
4. Binh T. Luong, F. Turini, S. Ruggieri. "k-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention," KDD2011, San Diego, CA, USA, 2011.

Presentations

1. Binh T. Luong, "Discrimination in the HTS," at *University of Pisa*, Pisa, Italy, 2009.
2. Binh T. Luong, "Extract terms from the HTS," at *University of Applied Sciences Cologne*, Koln, Germany, 2009.
3. Binh T. Luong, "Finding discrimination at different levels of generality," at *University of Pisa*, Pisa, Italy, 2010.
4. Binh T. Luong, "Ontology-based discrimination discovery," at *University of Bologna*, Bologna, Italy, 2010.
5. Binh T. Luong, "Discrimination Association Analysis on Semi-structured Data," European Conference on Data Mining (ECDM 2011) - IADIS Multi Conference on Computer Science and Information Systems (MCCSIS 2011), Rome, Italy, 2011.

Abstract

Recently, data mining has been deemed to be an effective means for disclosing evidences and hidden causes of discrimination. If data mining succeeds in finding associations proving the fact that discriminatory treatments has strong relations with sensitive attributes, discrimination is obviously irrefutable. In this thesis, I propose a modified approach of the traditional data mining process to unveil and represent discrimination in a "rich semantic" form for semi-structured business data with multiple-valued treatments based on support from ontology. First, input data are preprocessed to be well-structured with semantic relations, which considerably support discrimination exploration later. The framework then seeks possibly discriminatory relations between the unequal treatments and protected-by-law attributes, e.g., race, religion, sex. These discriminatory relations will be represented in the form of association rules through the notion of matching pairs of itemsets with different sensitive attributes and equal non-sensitive ones that are subject to different treatments. By combining data mining and reasoning service over the ontology, the achieved rules are semantically enriched by object properties between classes (concepts). Thus, they are more valuable and interesting than the flat association rules. In order to address the drawback of local knowledge, the solution of "kNN as Situation Testing" is provided. Besides, a number of measures of discrimination are provided for the purpose of quantifying the level of discrimination to obtain a precise vision of how different sensitive attributes negatively affect the decision and even on each other. Experimental results confirm the potential and flexibility of the approach.

Chapter 1

Introduction

Huge datasets are nowadays produced in commerce, medicine, science and a wide diversity of scientific disciplines at an enormous speed (possibly gigabytes per hour) due to advances in data collection, processing, management and storage technologies. For example, data warehouses store data of banking transactions, sales and purchases, personal health records, etc; genomic databases contain detailed information of structures and functions of genomic sequences; terabytes of images and other data from telescopes and satellites over large-area sky; surveys in optical, infrared and radio wavelengths in astronomy researches. Traditional data analysis tools and techniques are no more suitable for processing these data because of the massive size and high dimensionality of data as well as the heterogeneous, distributed nature of data. Moreover, users in most of situations do not expect trivial statistical information but hidden useful knowledge achieved by human - thought like processing. Thus, new methodologies need to be developed for discovering and representing interesting knowledge extracted from giant repositories of data, which is significantly important in high level and long term decision making systems.

Data Mining or Knowledge Discovery in Databases (KDD) generally is referred as a process which automatically or semi-automatically finds and extracts novel and valuable information or knowledge which is not

explicit in gigantic volumes of data. Actually, there are numerous definitions for Data Mining such as definitions proposed by Fayyad, Piatetsky, Shapiro and Smyth (FPSS96), (HMS01) or Benoît (Ger02). However, they all agree on the purpose of the data mining process that is the useful information extracted from the real-world datasets which discovers unknown interesting knowledge about characteristics of data, reveals new phenomena or enhances our understanding about known phenomena, and provides helpful predictions for further operations.

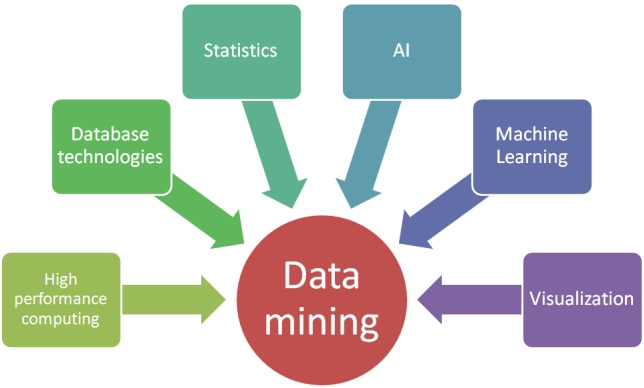


Figure 1: Data Mining backgrounds.

Data mining reflects novel directions within data processing which goes beyond simple data analysis. This is a multi-disciplinary approach which combines database technology, statistics, machine learning, artificial intelligence (AI), visualization and high performance computing (JM01) as illustrated in Fig. 1. In particular, database systems supply efficient data storage, indexing and query processing. Statistics play a critical role in the heart of data mining tools and techniques by concepts such as standard distribution, standard deviation, standard variation, regression analysis, discriminatory analysis and cluster analysis. They are significantly helpful to study data and relationships among data. Machine learning techniques are used to build computer programs which are able

to learn from the data, and then make appropriate decisions based on qualities of the studied data. Pieces of software use statistics for fundamental concepts, and possibly need adding varied AI heuristics and high performance computing in addressing the vast amount of data. Finally, an interface is required to visualize the structure, characteristics of data, and relationships among data items or data attributes as well as the new achieved knowledge.

The term KDD was coined in the 1989 IJCAI Workshop on Knowledge Discovery in Databases at Detroit, USA and since then researches in data mining have proliferated in around last twenty years. Yet foundation of present data mining technologies dated back to the 1950s when AI and machine learning were born by the combination of works of mathematicians, computer scientists and logicians (Buc06). Subsequently, main data mining technologies such as regression analysis, maximum likelihood estimates, bias reduction, neural networks, and linear models for classification were developed in 1960s (Dun03). At the same time, clustering techniques and similarity measures appeared and served for the area of Information Retrieval (IR). During the 1970s, 1980s, and 1990s, strong development of related disciplines (AI, IR, statistics, and database systems) supported by the appearance of fast microcomputers, new programming languages and new computing techniques. Especially, VSM model, genetic algorithms, EM algorithms, K-Means clustering, and decision tree algorithms (Dun03) are really useful in studying trends of data. All of these achievements built a firm foundation for the field of data mining later.

Typically, it is considered that "knowledge discovery process" is composed of a sequence of transformation steps (as depicted in Fig. 2). It includes all the data preparation and post-processing meanwhile "data mining" step is used in the sense of applying data mining techniques to clean data (FPSS96), (Dun03), (JM01), (TSK06):

- *Data Selection*: the real dataset is generally large and it is difficult to process the whole data. Hence, analysts and end-users must decide which data or which part of data is appropriate to the problem. In order to choose the suitable data, the problem as well as the scope

of the studied area followed by the solution should be primarily defined.

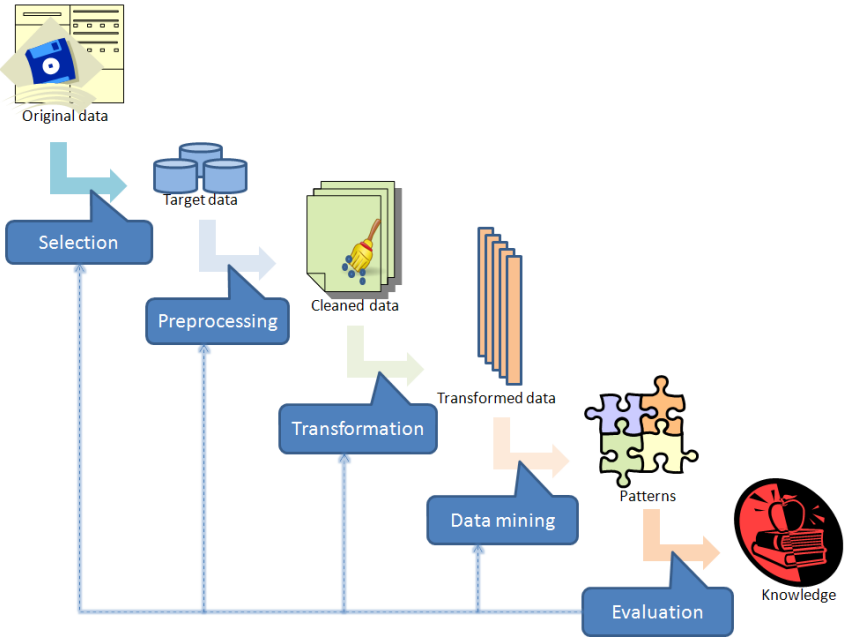


Figure 2: Data Mining Process.

- *Data Preprocessing*: Beside the massive volume, original data are probably inconsistent, weakly structured or ambiguous due to the heterogeneity, noise or mistakes of users, etc. Thus, it is required to clean the chosen data before being mined, for instance, fill in a default constant or an estimated value to the noisy or missing data.

- *Data Transformation*: due to (WIZT04), "data mining methods expect a highly structured format for data, necessitating extensive data preparation". It means the original data must be either supplied in a highly structured format or transformed into a suitable form. Actually, the entry data may exist in a number of formats (flat files, spread sheets, or relational tables), which may be complex for

processing stages. Aim of this phase is to adapt data for a corresponding form to the defined problem. It is composed of smoothing, aggregation, generalization, normalization, and attributes construction. Data reduction may be needed to implement the analysis process in a feasible, manageable and cost-efficient way.

- *Data Mining*: in this step, data mining algorithms are realized over clean and transformed data to extract novel and valuable information (patterns) or interesting relationships among data. The level of "interestingness" is measured by a variety of parameters related to the novelty, usefulness, simplicity, ... of pattern values (GH06). These patterns are able to provide predictions for further observations about trends or characteristics of the data. In general, classification and clustering algorithms are primarily applied and followed by association rules.

- *Interpretation and Evaluation*: the outcome of the data mining step must be evaluated by end-users or domain experts. Therefore, tools and algorithms of computer graphics are used to visually present the mined result for analyzing revealed characteristics of data or relationships among them. Subsequently, discovered knowledge should be applied in combining with existing models and knowledge in specific areas or simply reported to parties who are interested in.

It may be iteratively deployed between any two continuous stages or in the whole process to retrieve more refined or more precise results.

Data mining techniques are currently applied in various domains such as consumer analytics, finance, banking, medicine, genomics, and astronomy (MBGV07), (Han01), (Tho00), (Tho00), (VK06), etc, on a wide range of data types including databases, text, spatial data, temporal data, images, and other complex data. KDnuggets has conducted a survey on the application of data mining on several areas as depicted in Fig. 3 (KDn).

Nevertheless, among those applications, data mining techniques or more precisely classification models, which are often used in social anal-

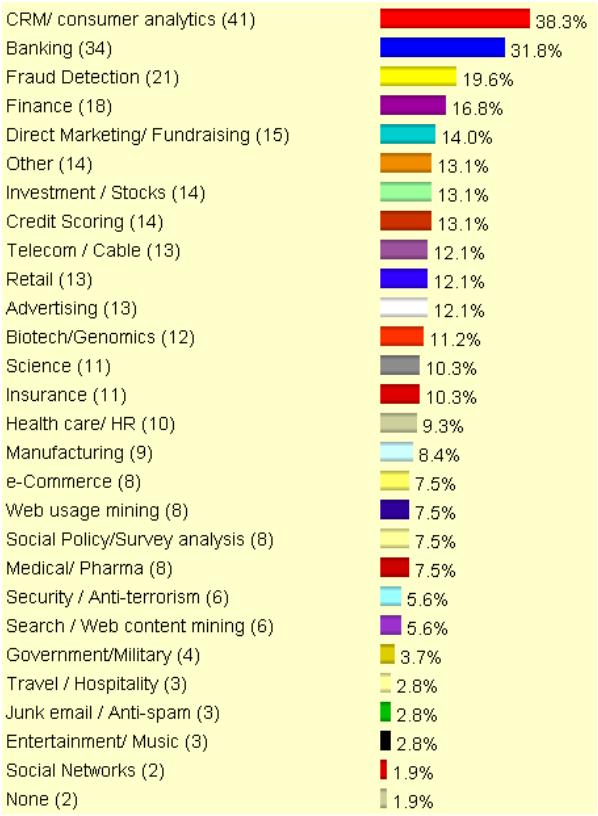


Figure 3: Data Mining Applications survey in 2008 by KDnuggets.

ysis, have been recently concerned of causing unfair treatments on the bases of protected-by-law attributes such as race, color, marital status, religion, disability, sex, age, etc. Starting from research of (Cli03), it was followed by (HH97), (Tho00), (Han01) in the field of credit scoring, (PRT08), (PRT09b) experimented in the area of loan application, (KC09) for discrimination prevention. Unfair treatment or discrimination herein is comprehended that members of a minority are treated unequally (less favorably) than the members of a majority without regarding to individual merits. These discriminatory behaviors are basically outlawed as

almost all the nations over the world promulgate laws against discrimination. For example, US federal legislative system has Equal Credit Opportunity Act, Fair Housing Act, and Equal Pay Act, etc., (U.S10) to forbid bias treatments in finance, housing, and credit; the European Union also enacts acts as well as policies and directives prohibiting discrimination, i.e., (Ell05), (BCP07), (Mak07), (Mak06), (Eur10), (El), (U.K10). Actually, the principles of equality and non-discrimination have been running as a red thread throughout all human rights treaties and declarations (Mak07), (Uni08), (Uni10).

The critical *status quo* of discrimination shows an urgent need for finding evidence and the hidden causes of unlawful discriminatory treatments in social fields, especially employment, health care, business, education, housing, ... This problem has been surveyed for a long time by economists, sociologists and legislators in human beings societies. For instance, there are studies in racial profiling and redlining (SO01), (Squ03), personnel selection (Gas84), (HN04), (M.J91), (PW99), (AC77) and mortgage granting (LL99), (Dym06), market place (RR02), (Sac04), (J.07), (Mic07a), etc. Despite a multitude of works on discrimination analysis, it is difficult to prove discrimination in practice. There are numerous challenges in building successful cases against discrimination, e.g., social customs, insufficient and/or inconsistent legal systems, and especially in lack of proofs because discrimination is often covert or indirect which requires a complicated process of discovery and inference. Usually the victims of discrimination are unable to provide persuasive evidence for the case since they do not have either the rights to access the source of information. Additionally, there are not sufficiently powerful means (theoretically and technically) to expose discrimination. Fortunately, the European legislative entities have recently established directives in order to instruct how to build and provide proofs of discrimination, for instance, the Directive $2000/43/EC$, the Directive $2000/78/EC$, the Race and Framework Equality Directives, etc. The common content of those acts and directives are establishing cases of forbidden discrimination, remedies, and enforcement, particularly enforcement of rights, burden of proof, and sanctions. Particularly, the burden of proof undoubtedly

takes the most significant role in convincing the courts that victims of discrimination have been objected to illegal treatments. Generally, the burden of proof must carry clear and powerful evidences showing that these victims face less favorable treatments than others in a comparable situation on the ground of prohibited discrimination. And it is probable that the most popular used method is *situation testing* (also situation test, auditing, pair-comparison testing, scientific testing, research testing). It uses a pair of people in experiments, one of them possessing a specific characteristic whereas the others do not. If distorted treatments are observed in a comparable case between the two people, it can be said that there is bias on the ground of the different characteristic. This method permits the disclosure of direct discrimination "on the spot", hence, is highly appreciated by scientists. "Situation testing has unique potential for studying the behaviour of actual employers in real workplaces while maintaining the methodological rigour of a laboratory-like scientific experiment. It is therefore appropriate to define the technique in a way emphasising its links to rigorous empirical research traditions in the social and behavioural sciences. In this spirit, (...) situation testing [is defined] as a systematic research procedure for creating controlled experiments analysing employers candid responses to employees personal characteristics." as remark of Marc Bendick, a researcher who has been working on the problem of discrimination in the USA for more than three decades (Ben07b). This method has been proving its strength in multiple cases of discrimination, e.g., access to employment in Belgium (AFN98), in the USA (Ben07b), in Czech (Aut08), (TMRMB09), entrance at restaurants and night clubs in Malmo (cou08), etc.

The idea of situation testing is considerably similar to association analysis of data mining in the aspect of discovering trends of groups of objects sharing a numbers of properties where transposition of trends basically depends on changes in background properties. Thus, a potential approach of discrimination discovery inspired from situation testing can be built with support of data mining, especially association analysis methodologies. Follow this trend, data mining can explore associations between different treatments and sensitive attributes such as gender, eth-

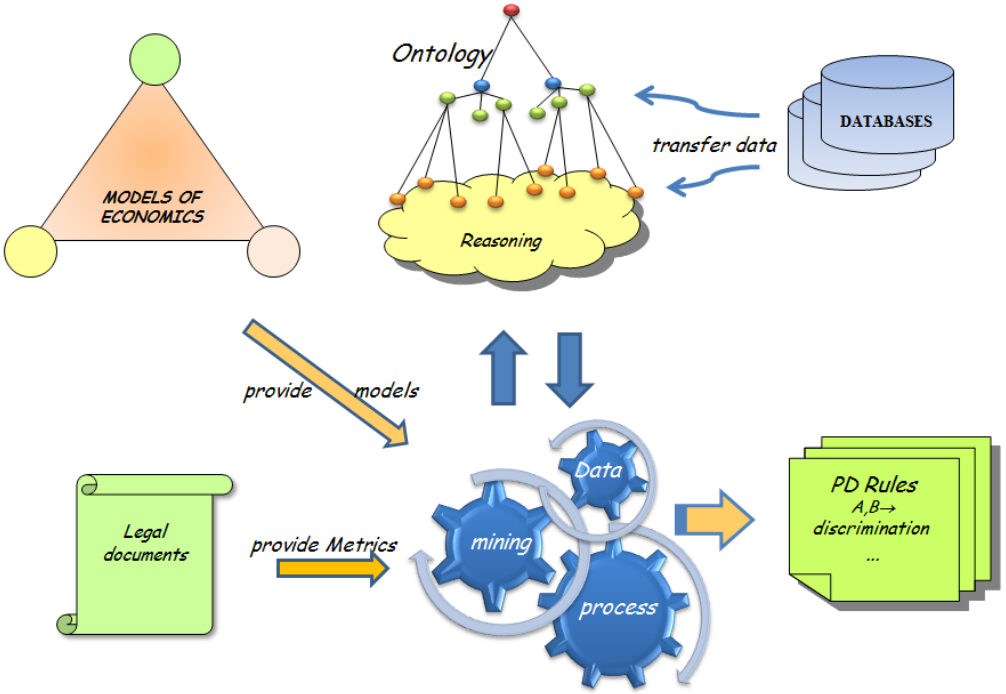


Figure 4: Proposed approach for discrimination mining.

nic origin, age, marital status, religion, etc, and represent them in the form of rules. Subsequently, the fact that different treatments are based on prohibited grounds are brought to light, therefore, illegal discrimination in treatment is irrefutable (with a few exceptions).

Despite the high risk of ethical and legal aspects of discrimination problem as well as the potential of the approach, this problem does not receive considerable attention from the computer science. Only a limited number of researches on this field have been published by applying data mining methods. On overall, there are probably three non-exclusive branches in data mining towards discrimination discovery and prevention. The first one is to adapt the preprocessing approaches of data sanitization (HISK05), (VEB$^+$04) and hierarchy-based generalization (Swe02), (WFY05). Along this line, (M.J91) adopts a controlled distortion of the training set. The second one is to post-process the traditional classification model, by implementing specified metrics on the obtained results. First works in this branch include (PRT09a), (PRT08), (PRT10), in which they propose a confidence-altering approach for classification rules inferred by the CPAR algorithm of (YH03). The third one is to modify the classification learning algorithm by integrating specialized steps and measures for discrimination analysis. Our research is inspired by the third branch.

In this research, I propose a framework as well as corresponding measures, and definitions to expose and show discrimination in a "rich semantic" form for business data using data mining methodologies with support from ontology as depicted in Fig. 4. The framework can find possibly discriminatory relations in the form of discriminatory association rules at multiple levels based on a general ontology for discrimination discovery as the common sense knowledge base. Formulae and their threshold for quantifying relationships between the different treatments, background attributes, and prohibited-by-law attributes are built from articles and legal documents in international and national law. It also seeks supports from economics models to collate the treatments achieved by advantageous and disadvantageous groups of the prohibited-by-law group. The methodology was then applied to the alternative datasets

showing potential results.

The thesis is structured as follows. Chapter 1 gives a brief idea about the scope of the thesis. Chapter 2 contains preliminaries of the background knowledge used in the proposed methodology. Details of the problem of discrimination in social sciences and early achievements in data mining are described in Chapter 3. Chapter 4 defines the problem and expounds of the discriminatory association analysis framework at multiple levels on business data for discrimination discovery. The architecture of the experimental systems is illustrated in Chapter 5. It is then followed by numerous experiments on varied merchandise databases shown in Chapter 6. To reduce the drawback of the proposed solution, a further approach using kNN is provided in Chapter 7; first results are also presented. Finally, conclusion on the approach in comparisons with previous work and future development are discussed in Chapter 8.

Chapter 2

Background knowledge and related works

The work of this thesis is mainly based on association analysis, and ontology engineering, and partly text mining techniques, thus, a review of this background knowledge is necessary. This chapter will provide the theoretical matters of computer science closely related to the scope of the problem addressed in the thesis.

2.1 Association analysis

2.1.1 General ideas

Association analysis or association rules mining is a basic task in data mining whose purpose is to discover events which often occur in tandem in the context of transaction databases or relational databases. It is motivated by the field of *market basket analysis* where the analysis of transactions of customers is important in predictions about the trends of market from which business policies or marketing strategies are issued. For example, it is observed that customers who purchase tea are likely to buy brown sugar and lemon as well. Hence, a discount on brown sugar may entice customer into purchasing brown sugar and also tea, therefore

Trans_id	Items
1	Bread, tea, lemon, brown_sugar, sausage
2	Fish, milk, bread, tea
3	Tea, brown_sugar, lemon
4	Fish, pepperoncino
5	Tea, brown_sugar

(a). Normal market-basket data example

Trans_id	Bread	Tea	Lemon	Brown_sugar	Sausage	Fish	Milk	Pepperoncino
1	1	1	1	1	1	0	0	0
2	1	1	0	0	0	1	1	0
3	0	1	1	1	0	0	0	0
4	0	0	0	0	0	1	0	1
5	0	1	0	1	0	0	0	0

(b). Binary transformation of market-basket transactions

Figure 5: An example of market-basket data.

sales of both brown sugar and tea are increased.

In the context of transaction databases, we are provided with a finite set of items $\mathcal{I} = \{a_1, \ldots, a_n\}$. In the relational databases, $\mathcal{I}$ is built from a relation $\mathcal{R}$ by considering items of the form $a = v$, where a is an attribute of $\mathcal{R}$ and v belongs to the domain of values of a. An example of ordinary transactions is provided in Fig. 5a. Also, it is needed to have a simpler form of data representation for measuring the relevance of data objects, then the binary form is created as illustrated in Fig. 5b, which is regularly used in data mining. An itemset $I \subseteq \mathcal{I}$ is a set of items. As usual in the literature, it can be written I, J for the itemset $I \bigcup J$. A transaction is a pair $t = (id, I)$, where id is a transaction identifier and I is an itemset. In relational databases, transactions reduce to tuples in the relation. We write $J \subseteq t$ as a shorthand for $J \subseteq I$, and t_{id} as a shorthand for id. A database of transactions, denoted by $\mathcal{D}$, is a set of transactions. The absolute support of an itemset I is the number of transactions in $\mathcal{D}$ covering I: $asupp(I) = |\{\ t_{id} \in \mathcal{D} \mid I \subseteq t\ \}|$, where $|\ |$ is the cardinality operator.

The (relative) support of I is the ratio:

$$supp(I) = asupp(I)/|\mathcal{D}|. \tag{2.1}$$

The associations among itemsets, subsequences, substructures, or generally called *patterns* are represented in the form of *probabilistic rules*. They are expressions $I \rightarrow J$ with a probability p where both left-hand side and right-hand side of the rules are Boolean (true or false in value) propositions of itemsets, with $I \bigcap J = \emptyset$. I is called the *premise* (or the *body*) and J is called the *consequence* (or the *head*) of the association rule. The interpretation of the rule is that the consequence is true with a probability p given that the premise is true. It means that there is a *causal* relationship between the body and the head of the rule: the existence of the former will lead to the existence of the latter with the accuracy p. For instance, an association rule is:

$$tea \rightarrow brown\_sugar \text{ [support = } 30\%\text{, confidence = } 75\%\text{]}$$

The above association rule means that 30% of all the client transactions include tea and sugar, and 75% of the clients who purchase tea also purchase brown sugar. The two probabilities used herein are *support* and *confidence* which are likely the most typical measures among metrics of rule interestingness; they are indicators to specify the level of usefulness and certainty of the discovered rule. A rule containing k items is called a k-itemset. The support of $X \rightarrow Y$ is defined as:

$$supp(I \rightarrow J) = supp(I, J) = P(I \bigcup J) \tag{2.2}$$

The coverage of $I \rightarrow J$ is:

$$cov(I \rightarrow J) = supp(I). \tag{2.3}$$

The confidence, defined when $supp(I) > 0$, is:

$$conf(I \rightarrow J) = supp(I, J)/supp(I) = P(J|I) \tag{2.4}$$

Support, coverage and confidence range over $[0, 1]$. If the discovered rule satisfies certain thresholds of minimum support *min_supp* and minimum confidence *min_conf*, then it is considered to be interesting or strong

Apriori Pseudocode

```
Apriori(T, ε)
    L_1 ← {large 1-itemsets occur in more than ε transactions}
    k ← 2
        while L_{k-1} ≠ 0
            C_k ← Generate(L_{k-1})
            for transaction t ∈ T
                C_k ← Subset(C_k, t)
                for transaction c ∈ C_t
                    count[c] ← count[c] + 1
            L_k ← {c ∈ C_k | count[c] ≥ ∈}
            k ← k + 1
    return ∪L_k
```

Figure 6: Apriori algorithm.

and can be used in further analysis. These thresholds are often established by domain experts.

Practically, association rules mining is composed of two steps:

- *Find all frequent itemsets*: generate itemsets whose occurrences exceed the pre-specified threshold of minimum support, *min_supp*.

- *Generate strong association rules*: generate strong rules from the frequent itemsets where each rule is a binary partioning of a frequent itemset by checking the condition of satisfying the minimum support *min_supp* and minimum confidence *min_conf*.

It is obvious that the first step requires much effort than the second, thus the overall performance of rules mining significantly depends on this step. Reducing the size of the candidate frequent sets then is the most challenging task in association rules discovery. A popular algorithm that can relatively well solve this matter is the Apriori which is based on the

Apriori principle: if an itemset is frequent, then all of its subsets must be frequent as well. Recent improved algorithms in association rules mining are based on this algorithm. Detail of this algorithm is illustrated in Fig. 6 (TSK06).

Since the seminal papers introducing association rules (ATA93), (AS94), many well explored algorithms have been designed in order to extract association rules with a user-specified minimum support, and minimum confidence. A survey on frequent pattern mining is due to (HCXY07), and a summary of interestingness measures for association rules is reported by (GH06). Also, explanation of implementations can be found in (Goe10).

However, primary results of studies on association mining are solely at single conceptual level, not yet represented at varied degrees of abstraction, i.e. represent attributes in a more refined data such as in a hierarchical structure. These *flat* association rules may lead to the limitation in expressivity caused by specific and concrete knowledge. This restriction can be tackled by *generalized association rules* which better disclose critical knowledge about trends of a broad domain. For instance, customers who purchase dairy products tend to buy bread also, or outwear products of company A is consumed better than that one of company B. This kind of information is truly precious in decision-making process in business. Hence, researches on generalized association rules mining have been conducted recently due to the need of discovering rules at different levels of granularity. The next session will represent this matter in more detail.

2.1.2 Generalized association analysis

Motivated by the need of finding associations between items at different levels of abstraction, the association mining process was adapted for the discovery of multiple-level knowledge data. Typically, *generalized association rules* extend flat association rules by exploiting an *is-a* hierarchy of concept taxonomy over items. The set $\mathcal{I}$ herein includes items appearing in transactions as well as their ancestors in the hierarchy. Premise

Figure 7: An example of a taxonomy.

and consequence in an association rule can now include items at different levels of the hierarchy, and rules can be compared on the basis of the levels in which items appear. It is said that an itemset $\hat{I}$ is an *ancestor* of an itemset I if $\hat{I} \neq I$ and for every $b \in \hat{I}$, either $b \in I$ or b is an ancestor of some $a \in I$. The rules $\hat{I} \rightarrow J$, $\hat{I} \rightarrow \hat{J}$, and $I \rightarrow \hat{J}$ are called *ancestors* or *generalized association rules* of the rule $I \rightarrow J$.

The idea of generalized association rules mining sprang from the study of (SA94), and then (HY95); it gradually evolved to be a main trend of research on association mining by works of (SS98), (TL01), (TL04), (WH08). In comparison to the classic association rules mining, this branch of association mining has been adjusted in the two following aspects:

- The taxonomy represents the structures of concepts in term of *classes*. Typically, it can be modeled as a DAG (Directed Acyclic Graph). An example of taxonomy of *Apparel Goods* is provided in Fig. 7. This domain is classified into categories such as *Outerwear*, *Footwear*, which are further classified into more specific items.

- Since the number of possible combinations as candidates of frequent itemsets are considerably increased, new algorithms are required to efficiently traverse the taxonomies.

In practice, the taxonomies can be either stored in the database, or

defined by an XML format for their structure. In the former case, taxonomies can be directly retrieved from the database and then used in generating candidates, e.g. replicating original transactions by replacing leaf nodes by their ancestors in the taxonomies. In the latter case, taxonomies are represented as a separate XML file or any formal structure. Among the available solutions, the newly developed ontology engineering appears to be a good suggestion. The second approach is obviously more organized, flexible and easy to access than the first. Thus, we decided to use ontology structure to represent the hierarchy of general association rules in discrimination discovery. This direction will be presented in more details in next chapter.

2.2 k-NN Classification

In the field of classification, objects are classified by their similarity which means objects belonging to a group A will possess almost the same characteristics or in the other hand they are "*closer*" to each other than objects belonging to other classes. Hence, one of common questions is: "*Given a data set and a query point in an m-dimensional metric space, find the item of the dataset that is closest to the query point*", which is called the *Nearest Neighbor* (NN) problem in classification as illustrated in Fig. 8. And k-Nearest Neighbors Algorithm might be one of the most popular classification algorithms due to its simplicity. It is an *instance-based learning*, or *lazy learning* algorithm as it:

- Defers data processing until it receives a request to classify an unlabeled query point.
- Replies to a request for information by combining its stored training data set.
- Discards the constructed answer and any intermediate results.

Given a data item r (the query point) to be classified and a set of N labeled items (the training set), kNN will: i)find the k items in the training set that are closest to r with regard to a distance measure d, i.e., its

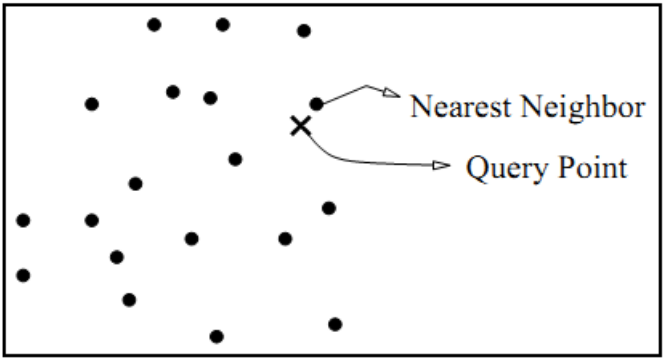


Figure 8: Query point and its nearest neighbor.

k-nearest neighbors; ii)assign r to the class value that appears most frequently among its k-nearest neighbors.

Depending on the type of data as well as applications, there are different kinds of distance metric to be chosen. Let us investigate these metrics. Let r and s be the two tuples having the same size of n. A distance metric $d()$ measures the dissimilarity between r and s. $d()$ is a non-negative real value, which will reach 0 if the two tuples are almost the same.

For *interval-scaled variables*, standardization of the data would be implemented first in order to avoid the dependence on measurement units using z-score:

$$z_i(x) = (x - m_i)/s_i \tag{2.5}$$

where m_i is the mean value, s_i is the mean absolute deviation of the domain of the ith attribute. Distance between x and y is then measured by the absolute difference of their z-scores. The most well-known distance metric is the *Euclidean distance*, which is defined as:

$$d_i(x;y) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{in} - x_{jn})^2} \tag{2.6}$$

Another popular metric is the *Manhattan distance* or the city block, defined as:

$$d_i(x;y) = \mid x_{i1} - x_{j1} \mid + \mid x_{i2} - x_{j2} \mid + ... + \mid x_{in} - x_{jn} \mid \tag{2.7}$$

These two metrics are particular instances of the *Minkowski distance*:

$$d_i(x;y) = (\mid x_{i1} - x_{j1} \mid^p + \mid x_{i2} - x_{j2} \mid^p + ... + \mid x_{in} - x_{jn} \mid^p)^{1/p} \quad (2.8)$$

For *binary variables*, their dissimilarity represents the number of different 1-digits. If the two binary variables are *symmetric* which means their states are equally valuable and carry the same weight; that is, there is no preference on which outcome should be coded as 0 or 1. Then, dissimilarity that is based on symmetric binary variables is called *symmetric binary dissimilarity*:

$$d(x,y) = (r+s)/(q+r+s+t) \quad (2.9)$$

where q is the number of attributes that equal 1 for both variables, r is the number of attributes that equal 1 for variable x but that are 0 for variable y, s is the number of attributes that equal 0 for variables x but equal 1 for variables y, and t is the number of attributes that equal 0 for both variables. The total number of attributes is p, where $p = q + r + s + t$.

A binary variable is *asymmetric* if the outcomes of the states are not equally important, such as the positive and negative outcomes of a disease test. Given two asymmetric binary variables, the agreement of two 1s (a positive match) is then considered more significant than that of two 0s (a negative match). Therefore, such binary variables are deemed "monary" (as if having one state). The dissimilarity based on such variables is called *asymmetric binary dissimilarity*:

$$d(x,y) = (r+s)/(q+r+s) \quad (2.10)$$

For *categorical variables*, which are often considered as a generalization of *binary variables* where the number of possible states can exceed two. The dissimilarity between two variables x and y then can be computed based on the ratio of mismatches:

$$d(x,y) = (p-m)/p \quad (2.11)$$

where m is the number of matches (i.e., the number of variables for which x and y are in the same state), and p is the total number of attributes.

For *ordinal variables* whose states are ranked in an order, distance calculation is similar to the case of *interval-scaled variables*.

For *ratio-scaled variables* which do not follow a linear scale, there are three possible ways to compute their dissimilarity:

- Apply the same method of *interval-scaled variables* on ratio-scaled variables.
- Apply logarithmic transformation to a ratio-scaled variable by the formula $y_{if} = log(x_{if})$. The achieved value y_{if} can be treated as interval- valued.
- Apply the metric of *interval-valued variables*.

Finally, distance between the two tuples is defined as:

$$d(r; s) = \sum_{i=0}^{n} d_i(r_i; s_i)/n \tag{2.12}$$

2.3 Text mining

From data warehouses, Internet to individual data stores, text is likely the most popular kind of data nowadays. However, its nature of unstructured, free form especially internal semantic dependences among lexical units make it difficult or even infeasible to automatic computer processing. Therefore, text mining was born to analyze and then structure the input text to extract its (potentially useful) patterns which are suitable for computing algorithm for a particular purpose (JM00). To serve this aim, text mining is based on Information Retrieval, Web mining, Statistics, Data Mining, and Computational Linguistics and Natural language processing. Thus, it is closely related to data mining, it is even called *text data mining*, or *text analysis*. A typical process of text mining shown in Fig. 9 illustrates this fact:

- Usually, the input data are massive, they must be filtered selectively to find the suitable textual data for the system. Generally,

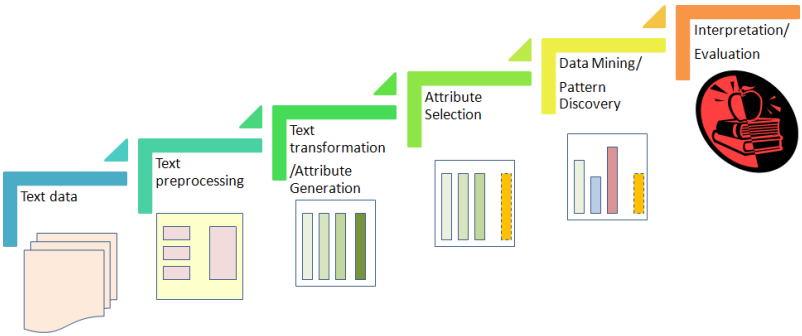


Figure 9: Text mining process.

document clustering is implemented to choose data, the most common methods are K-Means clustering, and Agglomerative hierarchical clustering.

- Text Preprocessing: As textual data are unstructured, they need cleaning before the automatic computation. To remove noise, and ambiguity whilst maintaining semantic dependence, numerous methods are realized: text cleanup, tokenization, Part-of-Speech (POS) tagging (Vou95a), word sense disambiguation, Semantic Structure.

- Text Transformation: After the preprocessing, textual data are represented in the form of words (*features*) and their occurrences, normally either by "*Bag of words*" or by "*Vector Space*" method. But a document can contain a critically high number of features, and processing all of these features is not effective, or even unnecessary. A subset of features will be selected to represent the given document. First, *stop words*, e.g., "*the*", "*or*", "*at*", "*which*", "*in*", "*on*", ... are removed; then stemming is performed to attain the root of a word. The preferred algorithms are Porter (Por80) or KSTEM (Kro93). Finally, a classifier is used to automatically generate labels (attributes) from the selective features.

- Attribute Selection: Not all features are relevant to the given prob-

lem, besides, the number of features sometimes is still considerably high. Hence, further selection is necessary to reduce the dimensionality. At the end of this step, the original unstructured data become well-structured with a reasonable number of representative features.

- Data Mining: Data mining algorithms are applied on the well-structured data to extract novel, useful, and interesting knowledge.
- Evaluation: Obtained results from the mining step can be used iteratively to derive a well-defined outcome, or as entry for further analysis.

Text mining has been a significant field of life; it is applied in various domains of application, for instance marketing, sentiment analysis, decision management, spam filter, security, biomedical, etc.

2.4 Ontology engineering

2.4.1 Ontology - A Description Logic based conceptualization

Ontology engineering is an emerging knowledge engineering process about conceptualization in computer science, but it was initially used in metaphysics by Aristotle and his students to describe the characteristics of entities. Traditionally, ontology is a branch of philosophy studying the nature of beings, existence or reality as such. It is possible to comprehend how entities are categorized and related given similarities and differences of their basic characteristics.

In computer science, ontologies are members of the family of class (concept) based knowledge representation formalisms-Description Logics (DLs). DLs, also known as *terminological logics*, *concept languages*, describe the domain of interest by *concepts* (classes) and *roles* (binary relations between concepts), particularly *subsumption* (is-a-kind-of) relationships. Complex concepts and roles can be built from the atomic ones by constructors. The strength of DLs is the central service of *reasoning*,

which allows one to infer covert knowledge from the overt one that is contained in the knowledge base. Let us recall the general formal definition of Description Logic (HT00):

DEFINITION (Description Logic) Let L be a *DL* based on infinite sets of atomic concepts NC and atomic roles NR. We will identify L with the sets of its well-formed concepts and require L to be closed under boolean operations and sub-concepts.

An interpretation is a pair $I = (\Delta^I; \cdot^I)$, where Δ^I is a non-empty set, called the domain of I, and $\cdot^I$ is a function mapping NC to 2^{Δ^I} and NR to $2^{\Delta^I \times \Delta^I}$. With each *DL* L we associate a set $Int(L)$ of admissible interpretations for L. $Int(L)$ must be closed under isomorphism, and, for any two interpretations I and I' that agree on NR, L must satisfy $I \in Int(L) \iff I' \in Int(L)$. Additionally, it is assumed that each *DL* L comes with a semantics that allows any interpretation $I \in Int(L)$ to be extended to each concept $C \in L$ such that it satisfies the following conditions:

- it maps the boolean combination of concepts to the corresponding boolean combination of their interpretations, and
- the interpretation C^I of a compound concept $C \in L$ depends only on the interpretation of those atomic concepts and roles that appear syntactically in C.

The basic constructors used in DLs are:

- conjunction ($\sqcap$), disjunction ($\sqcup$), negation ($\neg$).
- restricted forms of quantification: existence ($\exists$), universal($\forall$).

Their fundamental semantics are provided by the interpretation I as represented in Fig. 10.

Besides, a number of diversified constructors have been also investigated such as depicted in Fig. 11: Different DLs are diversified due to the constructors they use.

Distinguished characteristics of DLs sufficiently satisfy the requirements of ontologies languages (DAML-OIL, OWL-Lite, OWL) (HT00), (HS01), (SC03), (MSS04):

Semantics given by **interpretation** $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$

Constructor	Syntax	Example	Semantics
atomic concept	A	Human	$A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$
atomic role	R	has-child	$R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$
and for C, D concepts and R a role name			
conjunction	$C \sqcap D$	Human ⊓ Male	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$
disjunction	$C \sqcup D$	Doctor ⊔ Lawyer	$C^{\mathcal{I}} \cup D^{\mathcal{I}}$
negation	$\neg C$	¬Male	$\Delta^{\mathcal{I}} \setminus C$
exists restr.	$\exists R.C$	∃has-child.Male	$\{x \mid \exists y.\langle x, y\rangle \in R^{\mathcal{I}} \wedge y \in C^{\mathcal{I}}\}$
value restr.	$\forall R.C$	∀has-child.Doctor	$\{x \mid \forall y.\langle x, y\rangle \in R^{\mathcal{I}} \implies y \in C^{\mathcal{I}}\}$

Figure 10: DLs semantics and syntax.

Constructor	Syntax	Example	Semantics
number restr.	$\geqslant nR$	$\geqslant 3$ has-child	$\{x \mid \lvert\{y.\langle x, y\rangle \in R^{\mathcal{I}}\}\rvert \geqslant n\}$
	$\leqslant nR$	$\leqslant 1$ has-mother	$\{x \mid \lvert\{y.\langle x, y\rangle \in R^{\mathcal{I}}\}\rvert \leqslant n\}$
inverse role	R^-	has-child$^-$	$\{\langle x, y\rangle \mid \langle y, x\rangle \in R^{\mathcal{I}}\}$
trans. role	R^*	has-child*	$(R^{\mathcal{I}})^*$
concrete domain	$f_1, \ldots, f_n.P$	earns spends $<$	$\{x \mid P(f_1^{\mathcal{I}}, \ldots, f_n^{\mathcal{I}})\}$
⋮			

Figure 11: Extended DLs semantics and syntax.

- Its well formally defined semantics (as they are logics) make it possible for structured objects to be reasoned with.
- The key inference problem (subsumption reasoning) is decidable. The key inference herein is used in the classification of concepts and individuals, which specifies the subsumption relationships among concepts or determines if a given individual is an instance of a certain concept;
- The practical decision procedures and efficient implementation systems are available by optimization techniques.

Based on Description Logics, ontologies appear to be an effective means of conceptual modeling of an application domain. They realize and manipulate on formally descriptive categories of beings and their relations via a hierarchical vocabulary of terms (for concepts) along with

formal specifications of each term. They bolster the development of the World Wide Web nowadays by their key role in the *Semantic Web* through the annotation of web resources contained in meta-data. Ontologies show their potential via numerous aspects:

- Its advantages of easy sharing, maintaining and reusing domain knowledge among multiple parties are extremely suitable to the nature of de-centralization of existing resources of knowledge.
- Ontology reasoning provides answers to questions not easily solved by normal SQL queries about classes and instances of domain data.
- Ontologies focus on the structural aspect of domain knowledge instead of operational aspect. This separation provides a better view on domain knowledge and more convenient for management.

In ontology, concepts are represented as classes mainly via the *subsumption* relationships representing the hierarchical structure of the domain; they can also connect each other through *object properties*; internal attributes can be shown by *datatype* properties as the example illustrated in Fig. 12. Suppose there are eight individuals (*Italy, Switzerland, USA*, and (*IMT, Pisa university*), (*Ann, Leo, Marc*). They are grouped in three classes (*STUDENT, COUNTRY*, and *UNIVERSITY*). These groups (and also individuals) are connected by means of three types of properties (*hasHomeland, studies_at*, and *located_in*, e.g., *Leo* is a *STUDENT*, he *studies at Pisa university*, but his *homeland* is the *USA*. There may be constraints specified on these properties to restrict their domain to limited values. Therefore, it is very likely that ontology can be used in association mining at different levels of granularity due to the hierarchical structure of concepts. Indeed, this potential approach can be implemented in a flexible and effective manner using the *subsumption* relationship. In addition, due to advantages of ontologies, it appears to be more intuitive, and convenient to use ontologies than hierarchical databases.

The exact definition of an ontology is debated (FLMCO04), but the mostly quoted one is that an *ontology* is an explicit specification of a conceptualization(Gru93) in the form of a knowledge representation system based on Description Logics which comprises two components: a

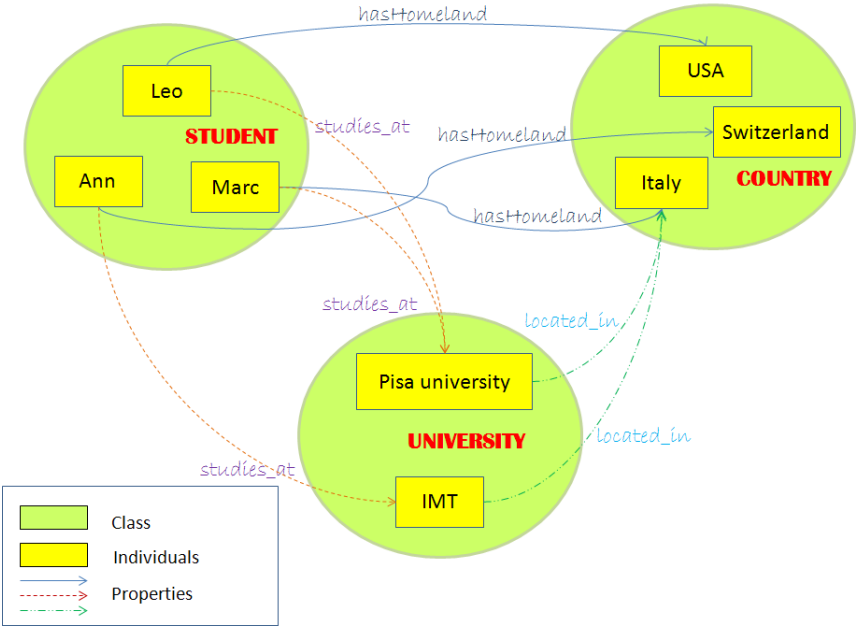


Figure 12: An example of a simple ontology.

TBox and an ABox. Hence, it has the classical architecture of DLs as depicted in Fig. 13 and typical characteristics of DLs as introduced above.

An interpretation $I = (\Delta^I; \cdot^I)$ is a model of the TBox T (Terminological Box), where Δ^I is a non-empty set, called the domain of I, and $\cdot^I$ is an interpretation function mapping every concept to a subset of Δ^I and every role to a subset of $\Delta^I \times \Delta^I$. The TBox describes concept hierarchies by means of a tree representation of *concepts* $C_1, C_2, \ldots, C_p$. It denotes a set of axioms asserting the structure of the application domain in the form of subsumption or equality relationships between (possibly complex) concepts:

- *inclusions* between either concepts or roles: $C_i \sqsubseteq C_j$ (C_i is said to be subsumed by C_j) if and only if $C_i^I \subseteq C_j^I$, denoting that C_i is an

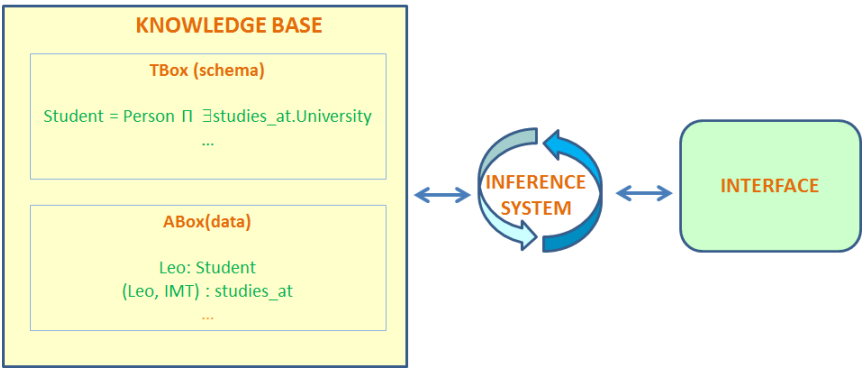


Figure 13: DL system architecture.

instance of C_j, i.e., every individual belongs to C_i also belongs to C_j; and similarly for roles, $R_i \sqsubseteq R_j$, e.g., Sandal $\sqsubseteq$ Footwear.

- *equalities*: $C_i \equiv C_j$ if and only if $C_i^I = C_j^I$, stating that C_i and C_j specify the same concept, and similarly for roles $R_i \equiv R_j$, e.g., Father $\equiv$ Male $\sqcap \exists$hasChild.Human.

The ABox (Assertional Box) contains knowledge about concrete situations. It is composed of axioms of assertions about named individuals according to terminologies of the concepts and roles provided in the TBox. These axioms have the forms of:

- *concept assertions*: $C(a)$ states that the individual a belongs to concept C, e.g., Sandal(Product123);

- *role assertions*: $R(a_1, a_2)$ states that a_2 is the filler of the role R for a_1, e.g., Tim $\equiv$ Male $\sqcap \exists$hasChild.Human.

The ontologies offer not only terminologies and assertions about an application domain but also reasoning services over these resources to unveil implicit knowledge of concepts and individuals in the given domain. The next section will concentrate on the key reasoning service of *subsumption* reasoning.

2.4.2 Reasoning on ontology

There are multiple basic inference problems in ontologies (HS01), (SC03), (BMNPS03):

- *Satisfiability* (i.e. non-contradiction) is the problem of checking whether a concept expression does not necessarily denote an empty concept. A concept C is *satisfiable* with respect to T if there exists a model I of T such that C^I is non-empty. In this case it is also said that I is a model of C.

- *Subsumption* is related to the structure of the knowledge domain; it is the problem of checking whether one concept is more general than another. Subsumption with regard to a terminology T: concept C is said to *subsume* concept D, it is written $C \sqsubseteq_T D$ or $T \models C \sqsubseteq D$ if and only if its extension is a superset of D's extension, i.e., $D^I \subseteq C^I$ for every possible interpretation I. This allows structuring the terminology of the domain as a subsumption hierarchy, which provides precious information on the relationships among concepts. And also subsumption with regard to individuals: an individual a is said to be an instance of a certain concept C if and only if $a \in C^I$ for all interpretations I. Hence, information on the properties of the given individual is revealed.

- *Consistency* checks if there is a model (terminology) and assertions that a particular individual is an instance of a certain concept. Or given a concept C, it is said to be *consistent* with regard to the terminology T if there is an interpretation I so that C^I is not empty (C^I is meaningful).

- *Equivalence*: Two concepts C and D are considered to be equivalent with respect to T if $C^I = D^I$ for every model I of T . In this case we write $C \equiv_T D$ or $T \models C \equiv D$.

- *Disjointness*: Two concepts C and D are disjoint with respect to T if $C^I \cap D^I = \emptyset$ for every model I of T.

The purpose of the above reasoning services is to determine if the given knowledge base is meaningful. Subsumption checking herein is the pivotal task in reasoning over ontologies as well as DLs since it is needed that the knowledge representation systems do not fail in verifying subsumption (BMNPS03). However, they all can be reduced to *satisfiability*. In practice, concept satisfiability is a case of subsumption, in which the empty concept is the subsumer. Generally, satisfiability checking is not performed directly. Practically, the tableaux algorithms use negation to reduce subsumption to (un)satisfiability of concept descriptions, e.g., $C \sqsubseteq D$ if $C \sqcap \neg D$ is unsatisfiable. The normal diagram of tableaux algorithm is:

- Build a model of concept C.
- Represent this model in the form of a tree T, in which:
  - Each node in T corresponds to a set of individuals in the model.
  - Nodes are labeled with sets of subconcepts of C.
  - Edges are labeled with role names in C.
- Start from the root node labeled C
- Apply expansion rules to node labels until
  - Expansion is completed (tree represents valid model). Or
  - Contradictions prove there is no model.
- Non-deterministic expansion satisfies the search (e.g., $C \cup D$)
- Blocking ensures termination (with expressive DLs).

2.4.3 Semantic Web Rule Language

Semantic Web Rule Language (SWRL) is a proposal for a Semantic Web rules-language, submitted to the W3C in May 2004 (HPSB$^+$04). It combines the Rule Markup Language (Unary/Binary Datalog) with OWL

DL and OWL Lite, which are sublanguages of the OWL Web Ontology Language. Actually, they are Horn-like rules of a high-level abstract syntax combining with an OWL knowledge base. Hence, it can be considered as an extension to OWL with a kind of condition, i.e. the *if-then* rules that can be expressed in terms of OWL concepts to provide more powerful deductive reasoning capabilities than OWL alone. Semantically, SWRL is built on the same description logic foundation as OWL and provides similar strong formal guarantees when performing inference. A rule language is needed for OWL due to a variety of reasons (HPSB$^+$04), (PSG$^+$05):

- The existing rules sets should be reused.
- More expressivity can be added to OWL such as a relatively arbitrary combination of property and class expressions.
- Manipulating rules on ontology by a rule language is generally convenient.

Let C, R, U, D be a class, object property, datatype property and datatype respectively. Let i, j be SWRL object terms, and $v, v_1, ..., v_n$ be SWRL datatype terms and let p be a built-in name. Then, the set of SWRL atoms is defined by the following grammar:

$$Atom \longleftarrow C(i) \mid D(v) \mid R(i,j) | U(i,v) | builtIn(p, v_1, ..., v_n) | i = j | i \neq j$$

Let a and $b_1, ..., b_n$ be SWRL atoms. A SWRL rule r is an expression of the form of conjunctions of atoms of an antecedent (body) and those ones of a consequent (head):

$$a \longleftarrow b_1, ..., b_n$$

These rules are interpreted that if the conditions specified in the antecedent hold, then the conditions specified in the consequent must also hold. Both of them can consist of zero or more atoms. An empty antecedent is considered trivially true, i.e. it holds for all interpretation, thus the consequent must hold as well. Whereas an empty consequent is treated as trivially false, i.e. it is not satisfied by any interpretation

hence the antecedent cannot also be satisfied by any interpretation. For example, a SWRL rule is:

$$\mathsf{Publication}(?\mathsf{a}) \wedge \mathsf{hasAuthor}(?\mathsf{x}, ?\mathsf{y}) \wedge \mathsf{hasAuthor}(?\mathsf{x}, ?\mathsf{z}) \\ \wedge \mathsf{differentFrom}(?\mathsf{y}, ?\mathsf{z}) \longrightarrow \mathsf{cooperatedWith}(?\mathsf{y}, ?\mathsf{z})$$

SWRL provides numerous types of atoms:

- *Class Atoms*: $\mathsf{Student}(?\mathsf{x}), \mathsf{University}(?\mathsf{y})$. A simple example rule using class atoms to declare that all individual of type Student are also of type Person can be written:

  $$\mathsf{Student}(?\mathsf{x}) \longrightarrow \mathsf{Person}(?\mathsf{y})$$

- *Individual Property atoms*: An individual property atom consists of an OWL object property and two arguments representing OWL individuals, such as $\mathsf{studies\_at}(?\mathsf{x}, ?\mathsf{y}), \mathsf{located\_in}(?\mathsf{y}, ?\mathsf{z})$. For example, one could write a rule to classify an individual named Marc as a student as follows:

  $\mathsf{Person}(\mathsf{Marc}) \wedge \mathsf{studiesAt}(\mathsf{Marc}, \mathsf{IMT}) \longrightarrow \mathsf{Student}(\mathsf{Marc})$

- *Data Valued Property atoms*: A data valued property atom consists of an OWL data property and two arguments, the first represents an OWL individual, and the second represents a data value. For instance:

  $\mathsf{hasAge}(\mathsf{Ann}, ?\mathsf{a}), \mathsf{hasGender}(?\mathsf{x}, ?\mathsf{b})$

- *Different Individuals atoms*: A different individual atom consists of the differentFrom symbol and two arguments represent OWL individuals. For example:

  $\mathsf{differentFrom}(?\mathsf{x}, ?\mathsf{y}), \mathsf{differentFrom}(\mathsf{Ann}, \mathsf{Leo})$

- *Same Individual atoms*: A same individual atom consists of the sameAs symbol and two arguments represent OWL individuals such as:

  $\mathsf{sameAs}(?\mathsf{x}, ?\mathsf{y}), \mathsf{sameAs}(\mathsf{Leo}, \mathsf{Marc})$

- *Built-in atoms*

- *Data Range atoms*: A data range atom consists of a datatype name or a set of literals and a single argument representing a data value. For instance: $\mathsf{xsd} : \mathsf{int}(?\mathsf{x}), [1, 2, 3](?\mathsf{x})$

SWRL uses a set of built-ins. A built-in is a predicate that takes one or more arguments and evaluates them to true if the arguments satisfy the predicate. Built-ins assist in the interoperation of SWRL with other Web formalisms by offering an extensible, modular built-ins infrastructure for Semantic Web Languages, Web Services, and Web applications. These built-ins are based on the reuse of existing built-ins in XQuery and XPath. They are motivated by the potential extensions of SWRL within a (hierarchical) taxonomy in the future. For instance:

- $\mathsf{swrlb} : \mathsf{equal}$: Satisfied if and only if the first argument and the second argument are the same.

- $\mathsf{swrlb} : \mathsf{pow}$: Satisfied if and only if the first argument is equal to the result of the second argument raised to the third argument power.

- $\mathsf{swrlb} : \mathsf{sin}$: Satisfied if and only if the first argument is equal to the sine of the radian value the second argument.

Given $I = (\Delta^I, \Delta^D, \cdot^I, \cdot^D)$ an abstract interpretation, a binding $B(I)$ is an abstract OWL interpretation that extends the interpretation functions $\cdot^I, \cdot^D$ such that:

$$V_{IX} \longrightarrow P(\Delta^I); V_{DX} \longrightarrow P(\Delta^D)$$

where P is the powerset operator.

A Binding $B(I)$ satisfies the SWRL atoms according to Table. 1, where V_{IX}, V_{DX} be a set of object and datatype variables respectively and let $V_{Built-In}$ be a set of built-in names, C, R, U, D are an OWL class, object property, datatype property and datatype respectively, i, j are SWRL object terms, v is a SWRL datatype term and p is a built-in name.

Intuitively, it is more convenient, and easier to write, read, and understand SWRL in comparison with normal OWL DL. They match better

Table 1: SWRL atoms

SWRL atoms	Condition on Interpretation
$C(i)$	$i^I \in C^I$
$R(i,j)$	$(i^I, j^I) \in R^I$
$U(i,v)$	$(i^I, v^D) \in U^I$
$D(v)$	$v^D \in D^D$
$builtIn(p, v_1, ..., v_n)$	$({v_1}^D, ..., {v_n}^D \in p^D$
$i = j$	$i^I = j^I$
$i \neq j$	$i^I \neq j^I$

with the human being's way of thinking and reference in certain problems. Indeed, it provides a powerful expressivity to OWL, which is expected to satisfy the need of a comprehensive representation of knowledge in the problem of discrimination mining.

2.5 Ontology mining

Since ontology is a method of conceptual modeling, it is possible to provide a formal representation of an application domain via a multi-level hierarchical structure of concepts, in which semantic relationships among categories of beings are also formally defined. Therefore, ontology is used in a variety of knowledge mining systems in supporting more elaborated results especially in the semantic extent. It is very likely that (SL96) commenced this direction. It continues the trend of mining generalized association rules with the support of multi-level concept hierarchies, particularly focuses on designing the structure of the data domain. Indeed, they suggest an object-oriented implementation for an explicit representation of a "dynamic" hierarchy. Along this line, an adaptive encoding schema is also used to support the evolution of concept hierarchies during its cycle for more interesting and informative rules.

Subsequently, (RPA03) introduces an architecture combining natural language processing, machine learning techniques, and ontology engineering for automated ontology learning from domain text. Extracted terms from the input dataset are matched with appropriate concepts in a

general-purpose ontology. During this matching phase, taxonomic and other semantic relations among the concepts are also unveiled.

A main trend in ontology mining is the exploitation of semantics obtained from the user's search results in Semantic Web. These semantics are formalized by ontologies, which represent knowledge at multiple conceptual levels; hence mining process can provide more satisfied outcomes to users. There are numerous studies on this direction. (YN03), (YN04) focus on using ontology to represent information retrieved from users in Semantic Web for better mining results. Particularly, (YN03) applies ontology in the representation of user profiles in Web mining for a better satisfaction of the need of users, in which "part-of" relation is used to describe the relationship between classes. In order to unravel the relationships between facts and the existing concepts, set-valued relevance functions are deployed. (YN04) proposes an approach of generating and reasoning over an ontology to capture evolving patterns. The introduced framework is helpful for uncertain or frequently changed data.

Similarly, (GCP03), (TG04) assist the navigation of Internet users by modeling their profiles in the form of an ontology. (GCP03) suggests to frequently updating their interests through the organization of the user profiles as a weighted concept hierarchy. In essence, the hierarchy can be automatically generated from a reference ontology storing the browsing history of users. Thus it is possible to provide results well satisfying user's search. As the ontology generation is transparent to users, it is needed to develop investigation techniques to warrant the accuracy of the modeling. (TG04) modifies (GCP03) by using a ranking function, which considers the number of documents assigned to the profiles instead of accumulated weights. Besides, it is observed that the lower the concept on the hierarchy, the less precise of concept detection. This can be addressed via the inclusion of lower-level concepts. Another ontology-based approach of supporting personalized searches is found in (SMB07). The user interests are implicitly modeled by a spreading activation algorithm as interest scores, and then assigned to existing concepts in domain ontology.

(CZSG03) introduces the use of ontology as a concept hierarchy in

the support of association rule mining in order to address the problems of sparse data sets as well as their multi conceptual levels. The idea of the algorithm is raising the data according to the structure of the hierarchy. Particularly, interests within tuples that are very specific are replaced by more general interests retrieved from the ontology. Hence, there are many more tuples at a more general level, which better support rule mining process and that better represent the reality.

(TLN07) presents a computational model for ontology mining for the interpretation of the semantic contained in user's queries. In fact, the semantic relationships among concepts in the ontology are dynamically investigated by the concepts of exhaustivity and specicity.

(Fd07) follows the direction of clustering similar resources in Semantic Web research w.r.t. a semantic dissimilarity measure for discovering new concepts. Authors exploit a measure of dissimilarity, which is based on resource semantics w.r.t a group of features represented by concept descriptions. The approach is realized by an adaptation of the classic Bisecting k-Means to complex representations typical of the ontology. It is relatively flexible to be applicable to a wide range of ontology languages (RDF through OWL). (EdF10) continues this direction in fuzzy clustering semantically annotated resources in Semantic Web for tackling the inherent uncertainty of knowledge bases. It is especially applied for knowledge bases represented through multi-relational standard languages. Possible clusterings are represented in tuples of central elements, which are iteratively adapted due to the principles of fuzzy clustering. Also, (dFE08) proposes an extension of the k-Nearest Neighbor algorithm for OWL ontologies based on an epistemic distance measure.

(LO08), (LO10) focus on the application of ontology engineering in the field of e-learning. Particularly, (LO08) presents a framework for personalized search in an e-learning platform based on content and the user profiles. It is obvious that the learner's navigation logs can assist in providing more satisfied results to user's search by re-ranking the search results retrieved from the learner's browsing history. The approach generates an ontology for the semantics of learner's profile from navigation logs, which are then clustered to discover more refined sub-concepts.

The search results are re-ranked based on the concepts closest matching the user profile. It is a controlled approach of the extraction of a taxonomy from the user's profile, hence it is more convenient and effective for the learning platform to serve user's need. (LO10) extends (LO08) by augmenting the framework by external open-source resources. The augmented data are clustered to extract the Top_n keywords (descriptive terms), which are then added as semantic terms to the ontology.

Follow the above works, this thesis does not aim at building novel theories for the foundation of ontology engineering. Instead, it uses ontology as a representation for the domain of discrimination discovery (particularly in business) in supporting semantics to the mining results. Basically, discriminatory constraints among concepts are defined via a matching step, which investigates the similarity of the two data tuples. Because of these discriminatory constraints and other semantic relationships, the number of frequent candidates for mining is decreased in comparison with mining in relational databases. Besides, the results containing semantic relationships are more comprehensive to the end-users.

Chapter 3

Discrimination discovery

3.1 Discrimination analysis in social research

3.1.1 Anti-discrimination in legal system

In legal system, it is defined that discrimination occurs in situations when members of a minority are treated unequally or worse (less favorably) than that one of a majority group on a ground of mostly similar salient characteristics except sensitive attributes, i.e, sex, race, age, marital status, or any other attribute prohibited by law. The term *discrimination* springs from the Latin word "*discriminare*" which means to differentiate an object from another. Discriminatory treatments might be intentionally caused by prejudice, social customs based on membership to a group or a minority. Or it can be unintentionally created in policies, treatments, decision-making, or selection process without recognition of the writer or operator. For example, if there is a discrimination in recruitment policy, given equally qualified white and black candidates with similar eligibility for requirements, the rates of acceptance for the interview are not equal between the two groups. In other words, the representations of blacks and whites in the shortlist is not similar to their representation in the applicants pool.

Typically, discrimination is covert and widespread, causing strong social, economic, political, historical and cultural impacts on the society. It

violates prime rights of human, and devaluates human beings. Thus, all forms of discrimination should be absolutely forbidden in law except special cases benefiting people belonging to disadvantageous minority groups. In Vietnam, for example, people living in remote regions can get admission to university at lower criteria. Therefore, this problem has drawn great attention of the society. The United Nation and its member states have made a multitude of efforts to complete the legal system in order to struggle discrimination by establishing policies and acts against discrimination. The base for these laws is the Universal Declaration of Human Rights which proclaims that all human beings are born free and equal in dignity and rights and that everyone is entitled to all the rights and freedoms, without distinction of any kind, in particular as to race, color or national origin. It proclaims further that all are equal in the law and are entitled without any discrimination to equal protection of the law and that all are entitled to equal protection against any discrimination and against any incitement to such discrimination. Any doctrine of racial differentiation or superiority is scientifically false, morally condemnable, socially unfair and dangerous. There is no justification for racial discrimination either in theory or in practice according to the (Bol):

- No state, institution, group or individual shall make any discrimination whatsoever in matters of human rights and fundamental freedoms in the treatment of persons, groups of persons or institutions on the ground of race, color or ethnic origin.
- No state shall encourage, advocate or lend its support, through police action or otherwise, to any discrimination based on race, color or ethnic origin by any group, institution or individual.
- Particular efforts should be made to prevent discrimination based on race, color or ethnic origin, especially in the fields of civil rights, access to citizenship, education, religion, employment, occupation and housing.

And the UN also proclaimed the United Nations Declaration on the Elimination of All Forms of Racial Discrimination on November 20th 1963 (Uni08).

Along this line, US federal laws forbid all forms of discrimination based on race, color, religion, belief, nationality, gender, marital status, age and pregnancy, ... For instance, the US federal legislative system (oNS04), (U.S10) has enacted:

- *Equal Pay Act* of 1963: prohibits discrimination on the basis of sex with regard to the compensation paid to men and women for substantially equal work performed in the same establishment.
- *Voting Rights Act* of 1965: prohibits voting practices that discriminate on the basis of race, color, or membership in a language minority group. Specifically, the act prohibits the use of discriminatory redistricting plans or voter registration procedures and authorizes the use of federal voting observers to monitor elections.
- *Age Discrimination in Employment Act* of 1967 - prohibits discrimination in employment on the basis of age which protects individuals who are age 40 or older.
- *Pregnancy Discrimination Act* - enacted as an amendment to the sex discrimination provisions of Title VII, made it unlawful to discriminate on the basis of pregnancy, childbirth, or related medical conditions.
- *Immigration and Nationality Act* 1986 - prohibits discrimination in employment on the basis of national origin or citizenship status. It also protects individuals from unfair documentary practices relating to the employment eligibility verification process.
- *Americans with Disabilities Act* of 1990 - prohibits discrimination based on disability in employment, public services, public accommodations, transportation, and telecommunications.
- ...

Besides, US legal system also prevents discrimination in social services such as public accommodation, education, nursing homes, adoptions, and transportation, ...

The same move has been observed in the European community, the European countries enact a number of acts and policies prohibiting discrimination. The original EEC Treaty included provisions prohibiting discrimination based on nationality showing that non-discrimination is one of prerequisite principles of EU's law system. Over the years, more sufficient and concrete acts against discrimination have been established. Particularly, Article 13 EC (and articles 21, 23 and 33), introduced in 1997, provides that the EC may take actions against discrimination based on gender, racial or ethnic origin, religion or belief, disability, age or sexual orientation. Among parts mentioning the problem of discrimination in the EC legislation system, sex-discrimination has a great importance and is divided into three parts (El), (Ell05):

- *Equal pay*: Article 141 established the principle of equal pay for equal work between male and female followed by the Directive 2006/54, ...
- *Equal treatment in access to and conditions of employment*: Article 137 makes equal treatment of men and women in the labor market an area for "supportive" Community action. The Council is given the power to adopt directives such as Directive 2004/113 on Equal Treatment in Access to and Supply of Goods and Services, Directive 2000/78.
- *Social security*: directive 79/7 provides the principle of equal treatment for men and women in the field of social security and other elements of social protection.

These acts intertwine in a sole consolidating measure Directive 2006/54. Moreover, there is specific gender equality legislation on pregnancy and parental leave, which are respectively Directive 96/34 and 92/85 (El) and rules stating that direct discrimination on the grounds of nationality will be caught by Article 39. Basically, the current EU law forbids discrimination on the ground of nationality, sex, part-time and temporary employment, racial or ethnic origin, religion or belief, disability, age, and sexual orientation. Based on these grounds, four forms of outlawed discrimination are defined:

- *direct discrimination*: occurs in situations when "one person is treated less favourably than another is, has been or would be treated in a comparable situation" (Eur10) on the prohibited ground of discrimination. For instance, an advertisement of recruitment stating "No immigrant!" is an obvious case of direct discrimination. But due to recent legal constraints, it is not easy to find such clear evidence.

- *indirect discrimination*: situates in circumstances where "an apparently neutral provision, criterion or practice" would put people of a particular protected sensitive ground under a less favored decision unless it can be "objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary" (Eur10). By adding the requirement of native-speaking French employees to non-related-to-media positions, for example, a company has prevented immigrants from applying.

- *harassment*: behaviors aiming at breaching the dignity of a person, intimidating, degrading, humiliating or offending that person in a certain period of time (Eur10), (Ror09), e.g., an Asian is often bullied, underestimated, or separated at office.

- *instructions to discriminate*: behaviors that suggest, assist, or require a third party to cause discrimination on the forbidden ground. The action that a restaurant manager does not allow black people to enter is a case of this outlawed discrimination.

Nowadays, EU, US and many other nations have appointed numerous authorities to monitor discrimination compliances in society. For instance, Europol - the European Law Enforcement organization, EuroJus - European Institute for Legal Studies and other responsible commissions, boards (El). Actually, the principles of equality and non-discrimination have been running as a red thread throughout all human rights treaties and declarations (Mak07), (Uni08), (Uni10).

In order to comprehend the hidden nature, motivation, and trends of discrimination as well as support decision-making process analysis, law

and policies establishment, the problem of discrimination has been also studied for a long time by economists, sociologists, politicians, journalists, public servants, and legislators. There are numbers of researches on discrimination in both market and non-market fields. For example, Holzer and Ludwig (HL03) conducted natural experiments in the labor market domain to test the effectiveness of anti-discrimination laws and how policy changes affect minority groups. Goldin and Rouse (CR00) focused on a setting to measure discrimination in hiring. Chay and Greenstone (CG00) examined trends in black-white infant health outcomes between 1955 and 1975. The US Committee on National Statistics considered the definition of race and racial discrimination, reviewed the existing techniques used to measure racial discrimination, and identified new tools and areas for future research (oNS04), etc.

In general, those researches can be divided into the following directions (oNS04), (H.A03a), (H.A03b):

- *Regression studies*: theirs motivation is to discover the effect of race or gender coefficients by using OLS (Ordinary Least Square) model, for instance, wage model of Neal and Johnson 1996 (DW96).
- *Audit studies and kindred field experiments*: randomize race parameter to reveal possibly disparate treatments to minor groups in sensitive areas such as job application, mortgage, health care, insurance, etc. (BM02), (CU01).
- *Lab experiments*: conduct surveys or experiments to find evidence of discrimination based on race or gender attributes, e.g., (Ell05).
- *Quasi-experiments*: applied in cases when race or gender information is either revealed or concealed. There are few researches for this direction such as (CR00).
- *Learning models*: search structural models for correlations between productivity and belief of workers over the time. Those models are built primarily from information gathered by econometricians through a number of sources (test scores, surveys), not by employers. (AP01) is a typical example.

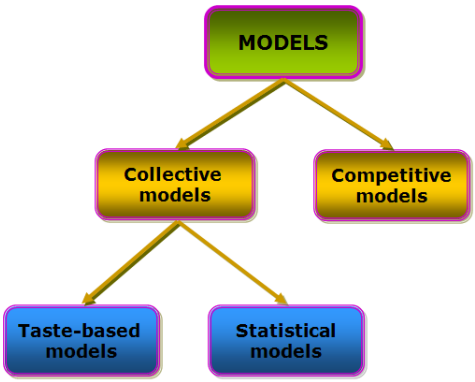


Figure 14: Models of discrimination.

For conducting experiments in alternative approaches, two groups of economic models of discrimination are often applied:

- *Competitive models*: focus on individuals maximizing behaviors which may contain disparity.
- *Collective models*: study operations of groups in which groups may act collectively against each other. This category attracts more attention from economists in economic analysis. It may split further:
  - *Taste-based models*: were mentioned the first time in (Bec57). Discrimination happens in the form that a certain group of objects is treated less favorably by prejudice. For example, in human resource market, employers hold a "taste for discrimination" and they underestimate ability of minority workers. Consequently, minority workers have to accept a lower wage for a given productivity or higher productivity for the average wage.
  - *Statistical models*: were coined by researches of Phelps (Phe72) and Arrows (Arr73). Discriminatory conclusion is drawn by analyzing statistics calculation on observable characteristics. For instance, a company has limited information about skills

of job applicants; the employer may use data of race or gender to infer expected productivity of applicants. The group of these models has been received more attention compared to the above one since they can provide more general or systematic information about how discrimination occurs other than individual or exceptional cases.

The common principle of these models is revealing tracks of discrimination and one of the most popularly used methods is *situation testing* which has been showing its potential via numerous cases. The role of evidence as the burden of proof and methodology of collecting evidence by situation testing will be discussed in the next section.

3.1.2 The weight of the burden of proof and Situation testing

Proving discrimination *per se* is a challenging task due to various reasons. First, it is not easy to discover straightforward proofs of discriminatory treatments since discrimination is often hidden or indirect. Second, there is the lack of transparency practices in decision making process as well as powerful means (theoretically and practically) to uncover discrimination. Traditional tools such as documentary records, witness statements or expert views are not sufficiently compelling for the analysis of unfair treatment. Third, having the right to access the source of information is not always feasible. Hence, victims of discrimination need special tools in order to achieving evidences to convince that they have been suffered from discrimination. To tackle this problem, legislators, policy-makers, and scientists mainly focus on two aspects:

- Building approaches and directives for the purpose of collecting evidences in a way that these methodologies must be feasible, legal, and effective, e.g., not breach the privacy of the considered people. The content of this section will mention about one of the most popularly used methods, *situation testing*, explaining how data mining theory, especially association analysis can be based on this tool in discrimination discovery.

- Developing statistical measures on observable attributes of obtained data to build more comprehensive, persuasive and satisfactory explanation or evidences of disparate treatments. This is actually the work on *Statistical models*.

Among the available methods for collecting proofs of discrimination, *situation testing* is a favorite tool of researchers. It shows its potential especially in revealing covert discrimination under different pretexts.

It is often stated that illegal discrimination occurs when: i) there are different treatments between people belonging to groups of different genders or ethnics or any prohibited ground; ii) one is treated worse or "less favorably" than the other; iii) whereas their situation are almost similar; iv) different treatments are distorted on the base of prohibited grounds defined in law, e.g., nationality, gender, age, disability, belief, sexual orientation, ...(Eur10), (U.K10). Based on this regulation, situation testing is formed for the purpose of exploring discrimination "on the spot" (Ror09) by conducting controlled experiments analyzing discrimination. This is a systematic approach for scrutiny of possibly discriminatory cases. It can clarify whether a person (or group of persons) possessing a specific characteristic is treated worse (less favorably) than others who do not possess that characteristic in a comparable situation.

Basically, in situation testing, pairs of candidates are established in a way that their general characteristics are almost the same (experiences, eligibility or clothes for entrance requirements of job, etc.) by selecting, training, except a single characteristic reflecting the discriminatory ground such as sex, race, age, and so on. They then will experience the same process of selection, for instance, applicants for job applications, loan, accommodations, or clients of restaurants, nightclubs, etc. If one of the two is behaved worse than the other, it can be stated that there is discrimination in treatment of the observed entity/person. The aim of this method is to establish a circumstance in which a person (or an entity) is observed without his/her recognition while making discriminatory treatments.

The first usage of situation testing was in the 1970s (Ror09) by the USA and UK to measure and evaluate the level of discrimination and

build policy to forbid discrimination. Nowadays, it is widely used by social scientists as a fundamental tool in discrimination discovery and policy-oriented studies.

The idea of situation testing can be well fitted with the methodologies of data mining and specifically association analysis. Obviously, given records of decisions taken in a context, for each member of the protected-by-law group with a negative decisions (an individual who may claim to be a victim of discrimination), individuals with similar, legally plausible characteristics, apart from being or not in the protected group are under scrutiny. If significantly different decisions between individuals of the protected group and the unprotected group can be recorded, it is possible to ascribe the negative decision to a bias against the protected group. When this situation frequently happens, the relationship between the discriminatory decisions and the two groups can be represented in the form of *association rules* among these related objects. These association rules can be discovered by statistics metrics and algorithms of association mining. Additionally, it is possible to analyse these discriminatory associations at different levels of abstraction by the *generalized association rules*. Also, rules might be generated from a neutral background whose similarity is generally modelled via a distance function. Hence, grouping rules sharing a given distance function may provide more valuable information to discrimination discovery. Details of implementing this idea in data mining will be introduced in the next chapter.

In the next sections, how the problem of discrimination discovery is solved in data mining will be presented. Despite the fact that it is a novel approach and requires to combine multiple disciplines, it has harvested considerable achievements.

3.2 Discrimination-aware data mining

In data mining, classification models can be used as an explanatory tool to distinguish between objects of different classes. They are built from historical data for two purposes: on the one hand to provide features defining class membership and on the other hand to predict class la-

bel of unknown patterns. Therefore, classification models are usually used in supporting decision making system in areas related to social analysis such as wages discrimination (AP01), criminal justice system (M.J91), (WA00), health care (AAK$^+$00) and racial profiling and redlining (SO01), etc. ...The risk arises while applying classification models to decision making system; it is possible for them to rapidly multiply effect of (not so often) discriminatory decisions on the bases of gender, age or other protected-by-law grounds to a vast number of data. The situation was directly coined in research of (Cli03) of classifiers that might cause discrimination. It is then followed by a variety of researches mentioned about discrimination from the statistics view by (HH97), (Tho00), (Han01) in the field of credit scoring, (PRT08), (PRT09b) experimented discrimination-aware data mining in the area of loan application, (KC09), (KCP09), (KCP10), (CV10), (KC10) for discrimination prevention. And as (PRT07) found, it becomes more serious when traditional prejudices may be amplified by learning from historical data. This situation is really dangerous since it leads to strictly outlawed malpractice as presented in the previous section.

Recent studies focus on discrimination discovery and (if possible) prevention based on data mining methodologies. They follow alternative but non-mutually exclusive approaches to solve the matter of discrimination. The first one is to adapt the direction of data preprocessing so that influence of sensitive information on the final decision is weakened, reduced or provided at different levels of generalization. These works are based on data sanitization (HISK05), knowledge hiding (VEB$^+$04), and hierarchy-based generalization (Swe01), (Swe02), (WFY05), or controlled distorted training set (KC09), etc. The second one is to post-process the produced classification models, in which (PRT09a), (PRT07), (PRT08) propose a confidence-altering approach for classification rules based on the CPAR algorithm (YH03). The third one is to modify the mining process by integrating specialized steps and measures for the discrimination analysis, which is the motivation of my research. In order to provide a clear and objective view of contributions of my work, the state-of-the-art works of the first two approaches are provided in this

section.

3.2.1 Data pre-processing approach

The main idea of this approach is removing the negative effect of protected-by-law attributes on the outcome by either hiding the part of this sensitive information in the original dataset before data publishing or modifying the learning criterion without causing critical information loss. It is *so-called* data sanitization, or formally stated as in (VEB$^+$04):

"Given a database D, a set R of relevant rules that are mined from D and a subset R_H of R, how can we transform D into a database D' in such a way that the rules in R can still be mined, except for the rules in R_H?"

The strategies of hiding sensitive information generally relies on decreasing the number of records supporting a given sensitive data, i.e. the original data will be distorted or cleaned. This idea can be implemented in either one of the two following directions or by combination of both of them:

- *data suppression* transforms the original data in a way such that the sensitive information is turned to implicit, normally replaced by unknown values. This direction is inspired by (VSC01) and then developed by (OZ03), (VEB$^+$04), (HISK05), etc.
- *generalized data* in which sensitive information is replaced by a less specific but semantically consistent value, which are works of (Swe01), (Swe02), (WFY05), (KC09), etc.

Practically, the former direction is based on the following principles:

- Only rules supported by disjoint large itemsets are disclosured.
- Association rules are removed (hidden) by decreasing either their support or confidence based on the side effects on the information that is not sensitive.

Various algorithms have been developed to implement the two strategies:

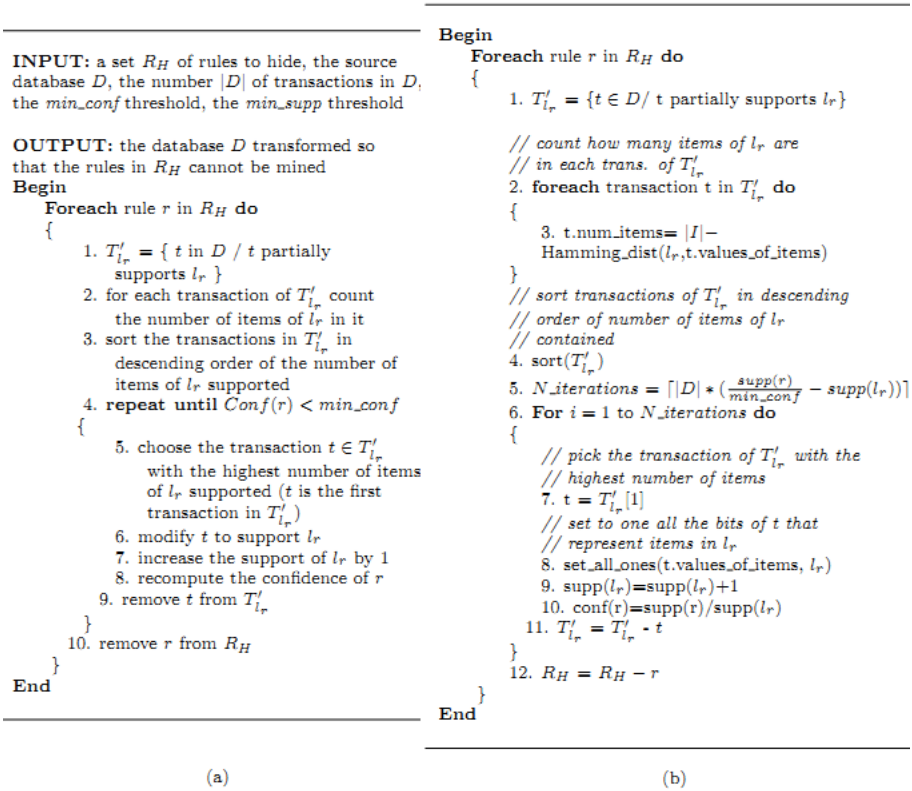


Figure 15: An example of rule-oriented algorithms.

- *rule-oriented* in which support or confidence of given sensitive rules is decreased until the rules are hidden. An example of this approach is shown in Fig. 15, and

- *itemset-oriented* which decreases the support of a given large itemset until it is below a user-specified threshold, thus, no rules can be derived from the selected itemset. Fig. 16 provides a typical illustration for this strategy.

For the approach of *data generalization*, (Swe01), (Swe02) have built

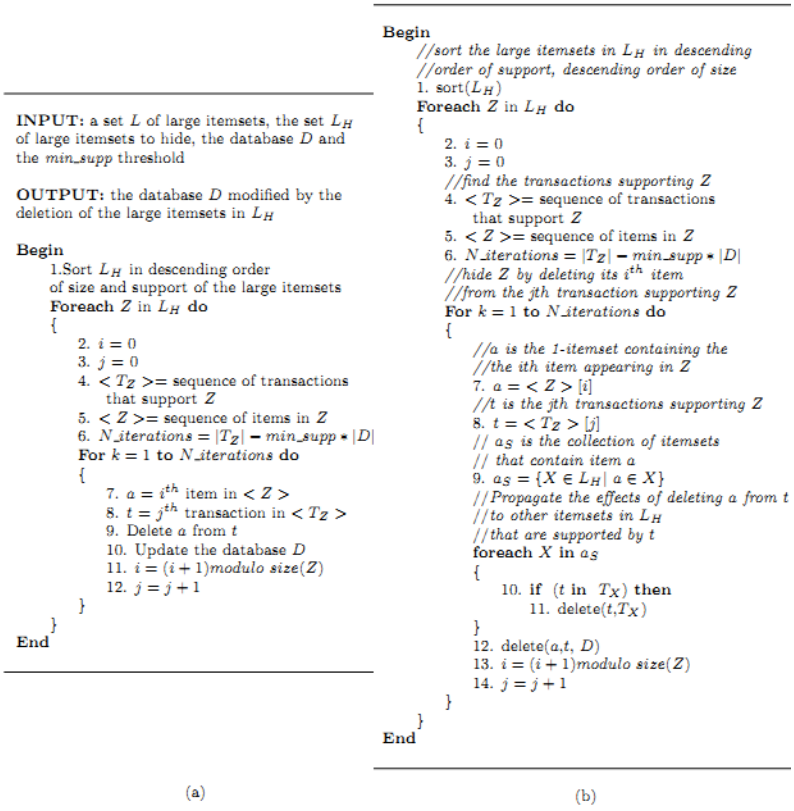


Figure 16: An example of itemset-oriented algorithms.

formal frameworks for sensitive data disclosure control in order to achieve a given level of anonymity by applying metrics of quality as well as anonymity on the distorted data in combination with data suppression. Originated from a simple idea: a piece of sensitive information is replaced by a less specific, more general but accurate to the original value. For example, the two values 02138, 02139 were generalized to 0213∗ as shown in Fig. 17. The new ZIP code semantically indicates a larger geographical area, hence the risk of revealing the actual place through the ZIP code is weakened.

Given an attribute A, a *generalization for an attribute* is a function f on A such that $f : A \rightarrow B$ and B is a generalization. A *generalization sequence* or a *functional generalization sequence* is defined as:

$$A_0 \xrightarrow{f_0} A_1 \xrightarrow{f_1} \ldots \xrightarrow{f_{n-1}} A_n$$

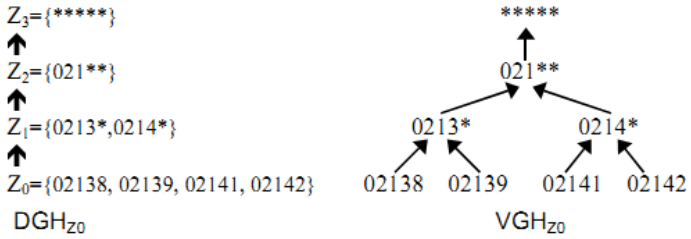


Figure 17: ZIP domain and value generalization hierarchies including suppression.

Given an attribute A of a private table PT, a *domain generalization hierarchy* DGH_A for A is defined as a set of functions $f_h : h = 0, \ldots, n-1$ such that:

$$A_0 \xrightarrow{f_0} A_1 \xrightarrow{f_1} \ldots \xrightarrow{f_{n-1}} A_n$$
$$A = A_0 \text{ and } |A_n| = 1. \; DHG_A \text{ is over:}$$
$$\bigcup_{h=0}^{n} A_h$$

Given a domain generalization hierarchy DGH_A for an attribute A, if $v_i \in A_i$ and $v_j \in A_j$ then $v_i = v_j$ if and only if $i = j$ and: $f_(i-1)(...f_i(v_i)...) = v_j$.

This relationship implies the existence of a *value generalization hierarchy* VGH_A for the attribute A. Indeed, given table PT, generalization can be effective in producing a table RT based on PT that adheres to k-anonymity because original values in RT are substituted with their generalized replacements. There are a number of algorithms to perform the generalization, for instance, the Preferred Minimal Generalization (MinGen) Algorithm shown in Fig. 3.18(a), and Core DataFly algorithm shown in Fig. 3.18(b), etc.

Input: Private Table **PT**; quasi-identifier QI = $(A_1, \ldots, A_n)$, k constraint; domain generalization hierarchies DGH_{Ai}, where i=1,…,n, and *preferred()* specifications.
Output: MGT, a minimal distortion of PT[QI] with respect to k chosen according to the preference specifications
Assumes: $|PT| \geq k$
Method:

1. **if PT**[QI] satisfies k-anonymity requirement with respect to k **then do**
   1.1. MGT ← { **PT** } // **PT** is the solution
2. **else do**
   2.1. *allgen* ← {T_i : T_i is a generalization of **PT** over QI}
   2.2. *protected* ← {T_i : $T_i \in$ *allgen* ∧ T_i satisfies k-anonymity of k}
   2.3. MGT ← {T_i : $T_i \in$ *protected* ∧ there does not exist $T_z \in$ *protected* such that *Prec*(T_z) > *Prec*(T_i) }
   2.4. MGT ← **preferred**(MGT) // select the preferred solution
3. **return** MGT

(a) Preferred Minimal Generalization (MinGen) Algorithm

Input: Private Table **PT**; quasi-identifier QI = $(A_1, \ldots, A_n)$, k constraint; hierarchies DGH_{Ai}, where i=1,…,n.
Output: MGT, a generalization of PT[QI] with respect to k
Assumes: $|PT| \geq k$
Method:

1. freq ← a frequency list contains distinct sequences of values of PT[QI], along with the number of occurrences of each sequence.
2. **while there exists** sequences in freq occurring less than k times that account for more than k tuples **do**
   2.1. **let** A_j be attribute in freq having the most number of distinct values
   2.2. freq ← generalize the values of A_j in freq
3. freq ← suppress sequences in freq occurring less than k times.
4. freq ← enforce k requirement on suppressed tuples in freq.
5. **Return** MGT ← construct table from freq

(b) Core DataFly algorithm

Figure 18: Algorithms of generalization.

Algorithm 1 *Classification with No Discrimination (CND)*

Input $(D, s, SA, +)$

Output Classifier CND learnt on D without discrimination

1: $(pr, dem) := Rank(D, SA, s, +)$
2: $existDisc := Disc(D, SA, s, +)$
3: Calculate M, the number of necessary modifications based on $existDisc$
4: **for** M times **do**
5: Select the data object from the top of pr
6: Change the class label of the selected object in D
7: Select the data object from the top of dem
8: Change the class label of the selected object in D
9: Remove the top element both of pr and dem
10: **end for**
11: Train a classifier CND on the modified D
12: **return** CND

Figure 19: CND algorithm.

Another methodology of data pre-processing is adapting the classification scheme for learning unbiased models on biased training data based on controlled distorted dataset (KC09), (KCP09), (CV10), (KCP10). They propose numerous approaches for generating a discrimination-free results after the classification. It can be either removing the sensitive attributes which are considered to cause discrimination on the class label of the related tuples, or adapting the construction of the classifier like the spliting criterion. The first approach proposes removing existing discrimination by a ranking function learned on the biased data. The sanitized data are then used in building a non-discriminatory model. Intuitively, removing a "disfavored" sensitive attribute can be a feasible solution (KC09). The algorithm of *Classification with No Discrimination* (CND) depicted in Fig. 19 illustrates this idea, it is able to degrade discriminatory or prejudicial behaviors in the future. In addition, this work provides a formal definition of the non-discriminatory classification problem which also involves a measure for assessing the discrimination in a dataset.

Algorithm 1 *Preferential Sampling*

Input (D, SA, c) **Output** Classifier CND learnt on D without discrimination

1: For all the data objects with $SA = s$: add in DP if $c = +$ else add in DN
2: For all the data objects with $SA = \overline{s}$: add in PP if $c = +$ else add in PN
3: Learn a ranker which arranges the data objects w.r.t. their probability of being in the desired class
4: Arrange the elements of DP and PP in ascending order and elements of DN and PN in descending order
5: Calculate the expected size for each combination of $v \in SA$ and $c \in C$ by $\frac{|v| \times |c|}{|D|}$
6: Change the sample size for each group by either re-substitution or skipping of top order elements iteratively
7: Move the top order elements with their duplicates (if exists) to the bottom of ranking after each iteration
8: Train a classifier CND on the re-sampled D
9: Return CND

Figure 20: Preferential Sampling algorithm.

(KCP09) provides other two strategies for the removal of the sensitive attribute from the input dataset. It can be implemented by either *Massaging* or *Reweighing* strategy. The former exchanges the class labels of objects which are opposite in a selected sensitive attribute to balance the training set. The selection is decided by a ranker. The undesired dependency between the selected attribute and the class in the classifier is, thus, removed. Actually, *Massaging* can be considered as an adaptation of CND. To obviate the modification of class label, the latter selects a biased sample to neutralize the impact of discrimination by assigning different weights to different data tuples. The training set is then generated by sampling the original data due to the assigned weights.

As CND is considered "intrusive", (KC10) developed another solution of using *Preferential Sampling* (PS) for creating discrimination free training set, in which the class relabeling is not needed. Alternatively,

Algorithm 1 Modifying Naive Bayes

Require: a probabilistic classifier M that uses distribution $P(C|S)$ and a data-set D
Ensure: M is modified such that it is (almost) non-discriminating, and the number of positive labels assigned by M to items from D is (almost) equal to the number of positive items in D

Calculate the discrimination *disc* in the labels assigned by M to D
while *disc* > 0.0 **do**
  numpos is the number of positive labels assigned by M to D
  if *numpos* $<$ the number of positive labels in D **then**
    $N(C_+, S_-) = N(C_+, S_-) + 0.01 \times N(C_-, S_+)$
    $N(C_-, S_-) = N(C_+, S_-) - 0.01 \times N(C_-, S_+)$
  else
    $N(C_-, S_+) = N(C_-, S_+) + 0.01 \times N(C_+, S_-)$
    $N(C_+, S_+) = N(C_-, S_+) - 0.01 \times N(C_+, S_-)$
  end if
  Update M using the modified occurrence counts N for C and S
  Calculate *disc*
end while

Figure 21: Modifying the Naive Bayes classifier.

Preferential Sampling (PS) modifies the distribution of diverse data objects. It is motivated from the observation that the data objects close to the decision boundaries are likely to suffer a "less positive" outcome. PS uses a ranking function to measure this closeness; the "least ranking" data (closest to the boundary) are then removed when sampling the dataset. Hence, training set is expected to be non-discriminatory, which makes the classifier non-discriminatory as well. The possible discrimination in classification in the future, therefore, is reduced. The approach is realized by the algorithm Preferential Sampling as in Fig. 20.

The above solutions focus on *altering the training dataset* to build a non-discrimination classifier. There is another approach of modifying the construction of the classifier. Particularly, (CV10) introduces three Naive Bayes classifiers without causing discrimination: i) modifying the probability of the sensitive attribute values given the class values to obtain a balance of the positive treatments imposed on favored and disfavored sensitive values. This method is performed by the algorithm

Algorithm 1: Decision Tree Induction

Parameters: Split evaluator $gain$, purity condition $pure$
Input Dataset D over $\{A_1, \ldots, A_n, Class\}$, att_list
Output Decision Tree $DT(D, att\_list)$

Create a node N
if $pure(D)$ or att_list is empty **then**
  Return N as a leaf labeled with majority class of D
end if
Select $test\_att$ from att_list and $test$ s.t. $gain(test\_att, test)$ is maximized
Label node N with $test\_att$
for Each outcome S of $test$ **do**
  Grow a branch from node N for the condition $test(test\_att) = S$
  Let D_S be the set of examples x in D for which $test(x.test\_att) = S$ holds
  if D_S is empty **then**
    Attach a leaf labeled with the majority class of D
  else
    Attach the node returned by $Decision\_Tree(D_S, att\_list - \{test\_att\})$
  end if
end for

Figure 22: Decision Tree Induction algorithm.

in Fig. 21. ii) training a separate model for each sensitive attribute value. When applying on the real dataset, the chosen classifier is the one having the same sensitive attribute to the real dataset. iii) adding a "latent variable" in the Bayesian model: Attributes which make the decision non-discriminatory will be represented by using a latent variable. The model parameters are optimized for the likelihood using expectation maximization. Especially, (KCP10) presents a *Discrimination Aware Decision Tree Learning*. Along this line, the splitting criterion of a tree node is considered to have an influence on discrimination, which can be measured by special information gain metrics with regard to a sensitive attribute. Therefore, a modification in this splitting criterion can be helpful in constructing discrimination-free decision tree as shown in Fig. 22. Also, the decision tree can be sanitized after the construction by the two approaches: discrimination-aware pruning and relabeling. In order to avoid

being discrimination-propagandized from its biased similarities, the label of a leaf is decided by the most frequent class of the set of tuples it belongs to. Label of a given leaf is adapted in a way that the discriminatory association is weaken with a minimal loss in accuracy. The algorithm is proposed due to the correspondence between the optimal leaf relabeling and the combinatorial optimization problem KNAPSACK.

These works mostly serve the purpose of assisting decision system in refraining discriminatory treatments, which is considerably useful in discrimination prevention and prediction. However, they do not directly address the problem of discrimination disclosure, especially in exploring the possible hidden causes of unequal treatments. The next section will mention studies in data mining that explicitly concentrate on revealing discrimination.

3.2.2 Data post-processing approach

Typical studies in this area (PRT08), (PRT07), (PRT09a), (PRT09b) provide an approach against discrimination in data mining by settings based on concept of *discriminatory classification rules*. Based on the judgment that disparity might result from the combination of a number of characteristics which are not discriminatory in isolation, simply excluding all sensitive-to-discriminatory attributes is not sufficient. They offer a new classification of terminologies for the convenience of discrimination discovery. Items or attributes which are sensitive to discrimination and normally under protection by law such as minority ethnicity in disadvantageous neighborhoods, divorced women, senior people in weak economic conditions are considered *potentially discriminatory* (PD). Their complementary attributes are called *Potentially Non-Discrimination* (PND). Based on this division of attributes, two kinds of *discrimination* might occur:

- *direct discrimination* is modeled by PD rules which are classification rules of the form $A, B \rightarrow C$ containing *potentially discriminatory* itemset A in their premises. For instance:

  personal status = female divorced/separated, savings status = no known savings $\rightarrow$ class = bad

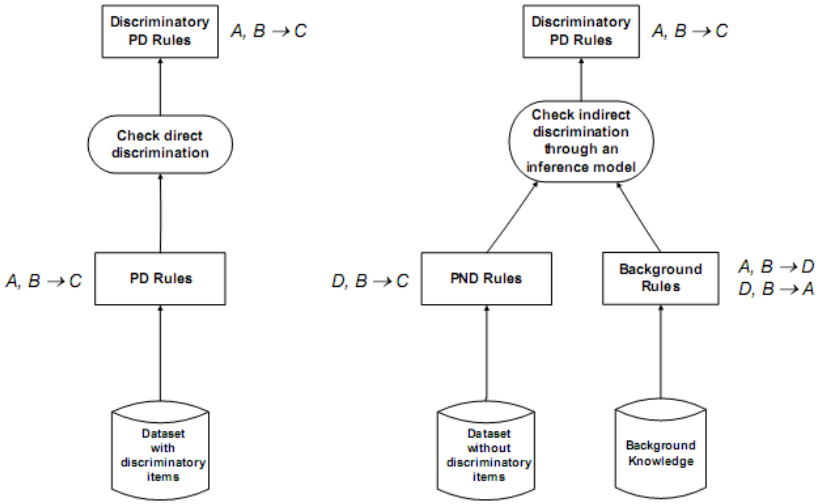


Figure 23: Modeling the process of direct (left) and indirect (right) discrimination control.

This is a PD rule as its premise contains the sensitive item "personal status = female divorced/separated" belonging to the group of PD attributes.

- *indirect discrimination* is modeled by *potentially non-discriminatory* (PND) rules $D, B \rightarrow C$, in which no PD itemsets is involved. For example:

  savings status = no known savings $\rightarrow$ class = bad

  This rule is not PD rule but a PND rule even though its antecedent is contained in the body of the previous rule.

Based on these concepts, an overall process of direct and indirect discrimination control was provided as depicted in Fig. 23. By defining the concept of PND rules, effects of background knowledge (in the form of

association rules relating a non-discriminatory itemset D to a discriminatory itemset A within the context B) on discriminatory decisions are also mentioned. Such an observation is critically important since it provides the context where treatments are weakly immune. Indeed, discrimination in one domain may diminish opportunities in other domains as noted in (oNS04). For example, disparity in house renting may affect residential options, which can also affect schooling and employment options.

In order to measure the level of disproportion a rule might impose, the notion of α-protection and a number of measures are provided.

Let R be a non-empty relation over attributes $a_1, \ldots, a_n$, namely $\emptyset \subset R \subseteq dom(a_1) \times ... \times dom(a_n)$, $dom(a_i)$ is the domain of values of a_i, assumed to be finite for every attribute a. An a-item is an expression $a = v$, where a is an attribute and $v \in dom(a)$. Also, an attribute c is fixed and called the class attribute. A c-item is called a class item. Let I be the set of all items, 2^I denotes the set of all itemsets. A transaction is a subset of I, with only one a-item for each attribute a. A database D of transactions is a set of transactions.An itemset $X \subseteq I$, for a transaction T of D, it is said that T verifies X if $T \models X$, namely for every $a = v$ in X, $T(a) = v$. The absolute support of an itemset X is the number of tuples in R verifying X: $asupp(X) = |T \in R|\sigma \models X|$, where $|\ |$ is the cardinality operator. The (relative) support of X is the ratio of transactions verifying X over the cardinality of D:

$$supp_D(X) = |T \in D|X \subseteq T|/|D| \tag{3.1}$$

Let $A, B \rightarrow C$ be an association rule such that A is a PD itemset, B is a PND itemset, and:

$$\gamma = conf(A, B \rightarrow C), \sigma = conf(B \rightarrow C) > 0$$

The relative gain in confidence due to the addition of the extra PD itemset A in the premise of the base rule $B \rightarrow C$ specified by non-discriminatory itemsets is defined as *extended_lift*:

$$elift = \gamma/\sigma \tag{3.2}$$

This parameter may be used as a suggestion for the desired level of protection against discrimination.

For a given threshold $\alpha \geq 1$, a rule is strongly $\alpha - protective$ if *glift* of that rule and its base smaller than α, where:

$$glift(\gamma, \delta) = \begin{cases} \gamma/\delta, & \text{if } \gamma \geq \delta \text{ ;} \\ (1-\gamma)/(1-\delta), & \text{otherwise} \end{cases} \tag{3.3}$$

If $glift(\gamma, \sigma) \geq \alpha$, the rule is said to be *strongly α-discriminatory*. Generally, α is a threshold of an acceptable level of discrimination that a rule might be. This parameter is set conforming to articles in laws, regulations, and jurisprudence. In the UK, for instance, a difference of 5% in confidence between female and male treatment is assumed by courts to be significant of discrimination against women (U.K10).

In order to extract discriminatory classification rules given an α threshold, a number of algorithms are proposed in these studies as example in Fig. 24.

Starting from a set of frequent itemsets ordered by the itemset size k, the *ExtractCR()* illustrates the extraction of PD and PND classification rules. Each k-frequent itemset containing a class item will generate a single classification rule, its confidence is computed based on the included itemsets of length $k-1$. If the extracted rule contains PD items, then that rule is considered as a PD rule, otherwise a PND rule. The outcome rules are grouped by the size *group* of the PND part of the premise. Therefore, all extracted rules are labelled as PD or PND rules. From results of *ExtractCR()*, the algorithm $CheckAlphaPDCR()$ on the right hand side of Fig. 24 checks the level of discrimination by calculating the extended lift $elift$ of a classification rule $A, B \longrightarrow C \in PD_{group}$ from its confidence and the confidence of the base rule $B \longrightarrow C \in PND_{group}$. The approach was later implemented on top of an Oracle database (PRT10).

Achievements of the above mentioned studies work well on well-defined data but do not warrant that they can unveil discrimination in semi-structured data whose implicit knowledge is probably valuable to discrimination discovery. Besides, discrimination is solely represented in a *flat* form, composed of basic classification rules, which may lead to limi-

ExtractCR()

$\mathcal{C}$ = { class items }
$\mathcal{PD}_{group} = \mathcal{PND}_{group} = \emptyset$ for $group \geq 0$
ForEach k s.t. there exists k-frequent itemsets
  $\mathcal{F}_k$ = { k-frequent itemsets }
  ForEach $\mathbf{Y} \in \mathcal{F}_k$ with $\mathbf{Y} \cap \mathcal{C} \neq \emptyset$
    $\mathbf{C} = \mathbf{Y} \cap \mathcal{C}$
    $\mathbf{X} = \mathbf{Y} \setminus \mathbf{C}$
    $s = supp(\mathbf{Y})$
    $s' = supp(\mathbf{X})$ // found in $\mathcal{F}_{k-1}$
    $conf = s/s'$
    $group = |\mathbf{X} \setminus \mathcal{I}_d|$
    If $group = |\mathbf{X}|$
      add $\mathbf{X} \rightarrow \mathbf{C}$ to $\mathcal{PND}_{group}$ with confidence $conf$
    Else
      add $\mathbf{X} \rightarrow \mathbf{C}$ to $\mathcal{PD}_{group}$ with confidence $conf$
    EndIf
  EndForEach
EndForEach

CheckAlphaPDCR(α)

ForEach $group$ s.t. $\mathcal{PD}_{group} \neq \emptyset$
  ForEach $\mathbf{X} \rightarrow \mathbf{C} \in \mathcal{PD}_{group}$
    $\mathbf{A} = \mathbf{X} \cap \mathcal{I}_d$
    $\mathbf{B} = \mathbf{X} \setminus \mathbf{A}$
    $\gamma = conf(\mathbf{X} \rightarrow \mathbf{C})$
    $\delta = conf(\mathbf{B} \rightarrow \mathbf{C})$ // found in $\mathcal{PND}_{group}$
    If $elift(\gamma, \delta) \geq \alpha$
      output $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$
    EndIf
  EndForEach
EndForEach

Figure 24: Classification rules extraction (left) and checking α-protection (right) algorithms.

tations in expressivity. Whereas the proliferation of data and social needs necessitate discrimination discovery in various kinds of data as well as the comprehension of discrimination at multiple points of view, from specific level to high abstract level. This part of discrimination discovery is the target of my thesis whose scope and solutions will be provided in the next chapter.

Chapter 4

Generalized discrimination discovery on semi-structured data supported by ontology

4.1 Introduction

The state-of-the-art of discrimination discovery introduced in previous chapter shows an interest in implementing knowledge discovery models, particularly association analysis, to the disclosure of discrimination . It is obvious that if data mining succeeds in finding associations proving the fact that the unequal treatments are strongly related to sensitive attributes, possibly illegal discrimination in treatment is irrefutable (with a few exceptions). For instance, in a tax database applied on garments, there is a difference in tariffs imposed on certain goods which is separately produced for men and women. If this difference can be exposed as association rules of the disparate tax and the gender attribute, then it is possible to judge that the tariff is negatively affected by the gender attribute.

However, data are not always ready for association analysis. Often

data and especially textual data are not well-structured or clean due to noises, incompleteness or inconsistency. This status results from many causes: by a variety of information sources, ambiguity by synonyms or negative meanings, etc, or mistakes of users. For example, a database of suppliers may assign several names for one item of goods for different local regions. It may also use varied descriptions for the same item that differ in producers or in materials. Direct mining on such data is impossible since the provided information is understandable for human beings but it is not in the form accepted by mining algorithms. It is required to implement data pre-processing in advance in a way that from an original input, sub-elements are extracted in the form of attribute/value pairs. In this case, text mining techniques such as text pre-processing, text transformation, pattern discovery may be helpful in parsing elementary components. Subsequently, association analysis can be applied on this well-structured data to reveal knowledge of hidden discrimination.

In this research, we propose solutions for discrimination discovery in the context of semi-structured business data, and represent this knowledge at multiple levels of abstraction. The scope of "semi-structured business data" used herein covers all business data that are not completely well-structured as pairs of attribute/value such as tables in relational databases. They still contain textual parts which can be split into small sub-components, which may semantically connect to each other. The approach is successful in representing the exposed discrimination in the form of association rules. Supported by a hierarchical common sense knowledge base, it can extract the semantic relations among data to enrich the obtained rules. Moreover, when the hierarchical structure is formalized as an ontology, generalized association rules can be dug, thus discrimination is explored at various conceptual levels.

In summary, the approach will address the three following aspects of the problem of discrimination discovery:

- *Semi-structured data*: need to be pre-processed so that they are eligible as pairs of attribute/value for the discrimination mining process. Also, possible semantic relationships among the sub-components should be reserved as they might provide valuable information to

the hidden discrimination.

- *The decision is not binary*: it is assigned different values (instead of normally two values "*Yes*" and "*No*") for different situations. It is then hard to specify a common context in order to unmask discrimination.

- *Adaptation of the traditional mining process* to discrimination exploration, especially at multiple levels of abstraction.

This chapter is organized as follows: first, the problem statement is provided, then typical examples of disparate treatments in business are provided in the second section. Finally, details of the framework is proposed.

4.2 Problem statement

According to the previous preliminaries, I is a database of itemsets, referring to a set of attributes:

$$A = a_1, \ldots, a_n,$$

where $dom(a_i)$ is the domain of the ith attribute.

Among these attributes there is often a special attribute a_{tg} used for classifying itemsets which is one of the purposes of the application. This attribute may be the cost of service, quality of service, e.g. granting or not a loan, level of health care service and so on. Its value depends on other attributes but this dependence cannot be represented as a simple function. Its domain is:

$$dom(a_{tg}) = c_1, \ldots, c_m$$

And des is a description attribute, containing a number of hidden elementary attributes and is expressed in a relatively free form. For instance, the following item of clothing merchandise database:

"*Men's or boys' suits, of synthetic fibers, not knitted or crocheted, containing* 36 *percent or more by weight of wool or fine animal hair.*"

This kind of data is significantly important since it contains covert properties of the product. However, directly digging knowledge from this item is impossible since the information is not well-organized for machine's operations. The available data mining techniques are able to work only on well-structured data such as attribute/value pairs. Even when being feasible with the above data, the achieved result will often lead to a too specific case, and consequently it will not provide useful or novel information. The matter turns simple if the elements contained in the description can be extracted as attribute/value pairs, which are items. What is needed then is to evaluate the effect of sensitive attributes like gender, age, etc on the target attribute a_{tg}.

Summing up, it is required to map these semi-structured items onto a set of well-defined transactions. For some of them, e.g. for the ID the mapping is one-to-one, but for others, and especially for the description, application of text mining might be performed on the description attribute before categorizing it into simpler items. The item given before as an example can be mapped into simpler sub-items:

- *Gender* = male
- *Name of goods* = suits
- *Form of production* = not knitted, not crocheted
- *Materials* = synthetic fibers, wool, fine animal hair
- *Quantity* = 36%

For the convenience of presentation, this description item *des* is said to refer to a set of sub-items:

$$a_{i1}, \ldots, a_{ip}, dom(des) = d_1, \ldots, d_m$$

These sub-items can be divided into a sensitive group such as the Gender item and a non-sensitive group such as Name of goods, Materials items. We define $S \subset A$ as the *Sensitive Attribute Set*, a set of protected-by-law attributes $\{v_i\}$, e.g., gender, race, age, sexual orientation which may cause discriminatory treatments, so-called *Potentially Discriminatory*-PD

attributes. Its complementary *Non-Sensitive Attribute Set* S' is composed of *non-sensitive* ordinary attributes $\{u_i\}$, and also *potentially non-discriminatory* (PND) attributes. The terminology PD-PND is borrowed from (PRT08), (RPT10). This set will form the background information or the context where discrimination happens. The transactions of the database are then expected to have the form of a three-tuple (u_i, v_i, c_i) where:

- u_i is the set of non-sensitive items, and
- v_i is the set of sensitive items, and
- c_i is the target item.

The objective of this research is to verify the influence of these PD attributes S on the target attribute a_{tg} using data mining techniques. While inspecting the database, if the variances in the target attributes a_{tg} considerably depend on sensitive attributes S, it can be said that this correlation is discriminatory, or represented as association rules with high confidence:

$$\{u_k\}, \{v\} \rightarrow c, \{u_k\}, \{v'\} \rightarrow c' \tag{4.1}$$

where:

- u_k is a subset of S' defining the background context where discrimination occurs.
- v_k, v'_k refer to a subset of the same sensitive attributes but differ in values.
- c, c' are target items with different values.

These association rules can then be used as proof of discrimination. Yet, the results of the above two general forms of association rules are hard to attain. Besides, they probably cannot provide much in details whether there is discrimination in the system or which item causes discrimination due to the following possible reasons:

- The target item is not binary. For example, the *accepted loan* item may have several levels for the amount of accepted money such

as over $50,000$\$, $10,000$\$ to $50,000$\$, and 0 instead of saying "yes" or "no" for the loan clients desire. It then requires a much more complex calculation to clarify: i) in which case(s) discrimination really happens for the target attribute, and ii) which attribute(s) has the negative effect causing the unequal values of the target attribute a_{tg}.

- It is difficult to construct all possible contexts to ascertain the influence of sensitive attributes on the outcome. These contexts must be flexible in order to cover all cases when discrimination might occur. The situation is even worse when the database is very large and/or the target attribute is not binary.

In addition, the extracted information of 4.1, solely considers items at a single concept level, hence, it may lead to limited or poor semantic results. If these findings can be represented at different levels of granularity, then more implicit, precise and meaningful knowledge can be revealed. For instance, garments importers want to know how much the tariff system differentiates between products for men and those ones for women. It is not sufficient to provide only diversified taxes of each individual apparel? Or is there any discrimination in mortgage loan system when women can receive an amount up to e.g. 80%, 60% or 50% of the loan whereas men are not often accepted even though they both are eligible for loan application? An expected solution is finding out all varied contexts (from general to detailed levels) when discrimination happens. A potential way is using generalized association rules that have been proved to convey a richer knowledge when compared to flat association rules due to their hierarchical structure of information. In the meantime, ontologies appear to be an effective means of formally descriptive categories of beings and their relations via a hierarchical organization and semantic dependence through object properties. An advantage of ontology engineering is ontology reasoning which provides answers to question not solved by normal SQL queries about classes and instances on ontology. Besides, ontologies are also easy for using, sharing, and maintaining of data among multiple parties. Therefore, it is very likely that

ontology can be used in association mining at different levels of granularity in a flexible and effective way. In addition, its intuition, utility and flexibility can be advantageous against hierarchical databases.

Given a dataset of historical decisions $\mathcal{D}$ and an ontology O in the form of a taxonomy, namely a hierarchical structure built from the domain of $\mathcal{D}$, the *problem of discrimination discovery* from $\mathcal{D}$ and O consists of unveiling varied contexts where decisions are discriminatory. This is implemented by mining generalized association rules, which provide evidence of discrimination at different levels of granularity of the taxonomy.

More precisely, two types of comparisons will be done, depending on whether:

- there is a different treatment for people with specific sensitive attribute values from the generality of people in the same context. For instance, we contrast the following two rules:

  Shoes, leather $= 30\%$, knitted = YES, gender = male $\rightarrow$ tariff = 4.5%

  Shoes, leather $= 30\%$, knitted = YES $\rightarrow$ tariff = 4%

  concerning shoes made from 30 percent of leather and knitted. In such a context, the overall tariff is 12.5 percent (0.5 percent in absolute value) lower than the one particularly devised for men. This implies that a final customer may pay the product differently on the basis of his/her gender. In other words, the tariff is discriminatory with regard to the gender of the customer. In general, this discrimination can be modeled by pairs of rules:

  $$\{C_i\}, \{R_j\}, \{R'_k\} \rightarrow d_1, \{C_i\}, \{R_j\} \rightarrow d_2 \tag{4.2}$$

  - $\{C_i\}$ is a subset of named individuals of the ontology;
  - $\{R_j\}$ is set of non-sensitive properties of those individuals, such as materials of garment products;
  - $\{R'_k\}$ is a set of sensitive properties of those individuals, such as the gender the garment is produced for;

– d_1, d_2 are different treatments, e.g., different tariffs.

- there are different treatments for people possessing different sensitive attribute values. For instance, we contrast the following two rules:

  Shoes, leather $= 30\%$, knitted = YES, gender = male $\rightarrow$ tariff = 4.5%
  Shoes, leather $= 30\%$, knitted = YES, gender = female $\rightarrow$ tariff = 3.5%

  concerning shoes made from 30 percent of leather and knitted. In such a context, the tariff for men is 28.57 percent (1 percent in absolute value) higher than that the one for women. Again, this means that the tariff may produce different treatments (namely, cost for buying the product) on the basis of the gender of the final customer. In general, this discrimination can be modeled by pairs of rules:

  $$\{C_i\}, \{R_j\}, \{R'_k\} \rightarrow d_1, \{C_i\}, \{R_j\}, \{R'_h\} \rightarrow d_2 \tag{4.3}$$

  where

  – $\{C_i\}$ is a subset of named individuals of the ontology;
  – $\{R_j\}$ is a set of non-sensitive properties of those individuals;
  – $\{R'_k\}$ and $\{R'_h\}$ are subsets of sensitive properties of those individuals, with different values (i.e., $\{R'_k\} \neq \{R'_h\}$) in the two rules;
  – d_1, d_2 are different treatments.

These two rules 4.2 and 4.3 provide associations between the unfair decision and the protected-by-law attributes, hence the existence of discrimination is proved. In order to achieve these rules, the traditional rule mining process should be adapted when applied to discrimination analysis by integrating discrimination calculation and a hierarchical structure of the given domain within it.

In the next section, real life cases of discrimination will be mentioned; they show the need of the work as well as its potential applicable fields.

4.3 Examples

4.3.1 HTS dataset

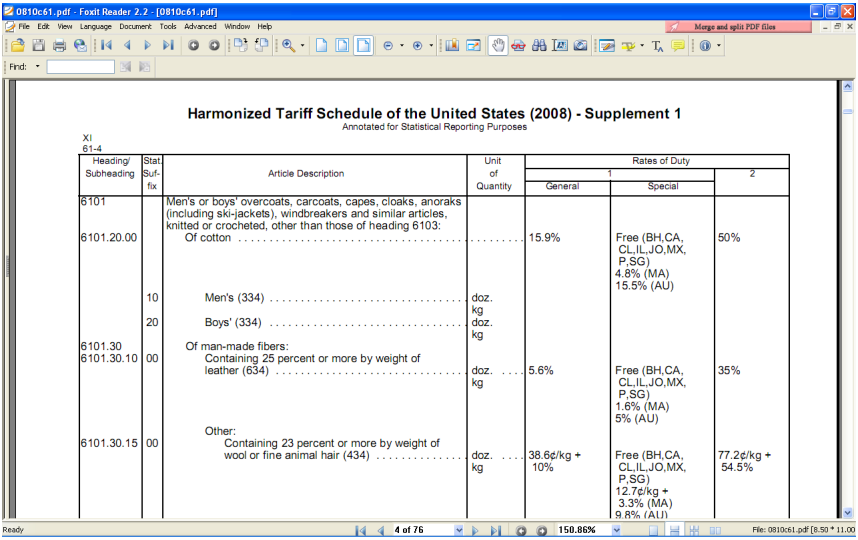
The Harmonized Tariff Schedules (HTS) (U.S78b) of the United States is the main source in determining tariff classifications for goods imported into the US. It was promulgated in the Omnibus Trade and Competition Act in 1988 and effective since January 1st 1989. This schedule is built based on the international Harmonized Commodity Coding and Classification System - Harmonized System (HS) of the International Custom Bureau whose head quarter is located in Brussels, Belgium. The HS is an internationally standardized system of names and numbers for classifying traded products developed and maintained by the World Customs Organization. At the present, the US International Trade Commission (ITC) is responsible for building, managing, and updating the HTS.

The existing HTS is a hierarchical structure with 99 chapters and a variety of additional sections such as appendices and indexes. It provides a detailed tariff classification system for merchandise imported to US, including nomenclatures (name), descriptions for goods, formulas for calculating tariff rates, ad valorem, specific and estimated ad valorem equivalent (AVE) tariffs as illustrated in Fig. 25.

The HTS also indicates products that are eligible for tariff preferences under free trade agreements such as with Canada, Mexico and Israel. There are products eligible for any preferential programs such as the Generalized System of Preferences (GSP), the Caribbean Basin Initiative (CBI) and the African Growth and Opportunity Act (AGOA).

All kinds of merchandise imported into the United States are subject to duty or duty-free entry in accordance with their classification under the applicable items in the HTS. When merchandise is dutiable, ad valorem, specific, or compound rates may be assessed:

- An *ad valorem rate*: the most often applied type of rate which is a percentage of the value of the merchandise, such as 5 percent ad valorem.
- A *specific rate*: is a specified amount per unit of weight or other



Harmonized Tariff Schedule of the United States (2008) - Supplement 1
Annotated for Statistical Reporting Purposes

XI
61-4

Heading/ Subheading	Stat. Suf-fix	Article Description	Unit of Quantity	Rates of Duty 1 General	Rates of Duty 1 Special	Rates of Duty 2
6101		Men's or boys' overcoats, carcoats, capes, cloaks, anoraks (including ski-jackets), windbreakers and similar articles, knitted or crocheted, other than those of heading 6103:				
6101.20.00		Of cotton		15.9%	Free (BH,CA, CL,IL,JO,MX, P,SG) 4.8% (MA) 15.5% (AU)	50%
	10	Men's (334)	doz. kg			
	20	Boys' (334)	doz. kg			
6101.30		Of man-made fibers:				
6101.30.10	00	Containing 25 percent or more by weight of leather (634)	doz. kg	5.6%	Free (BH,CA, CL,IL,JO,MX, P,SG) 1.6% (MA) 5% (AU)	35%
		Other:				
6101.30.15	00	Containing 23 percent or more by weight of wool or fine animal hair (434)	doz. kg	38.6¢/kg + 10%	Free (BH,CA, CL,IL,JO,MX, P,SG) 12.7¢/kg + 3.3% (MA) 9.8% (AU)	77.2¢/kg + 54.5%

Figure 25: An example of HTS data.

quantity, such as 5.9 cents per dozen.

- A *compound rate*: is a combination of both an ad valorem rate and a specific rate, such as 0.7 cents per kilo plus 10 percent ad valorem.

Rates of duty for imported merchandise may vary depending upon the country of origin. Most merchandise is dutiable under the Most-Favored-Nation (MFN), now referred to as normal trade relations rates in the General information. Merchandise from countries to which these rates have not been extended is dutiable at the full or "statutory" rates.

Free rates are provided for many subheadings of the tariff schedule. Duty-free status is also available under various conditional exemptions which are reflected in the Special column of the tariff schedule. It is the importers burden to show eligibility for a conditional exemption from duty (refer to the Fig. 25).

Even though this system is carefully built and updated conforming to the constitution and government's policies, Michael Barbaro has uncovered an oddity of the tariff system: duties on men's and women's

Heading/ Subheading	Stat. Suf-fix	Article Description	Unit of Quantity	Rates of Duty 1 General	Rates of Duty 1 Special	Rates of Duty 2
6112 (con.)		Track suits, ski-suits and swimwear, knitted or crocheted (con.):				
		Men's or boys' swimwear:				
6112.31.00		Of synthetic fibers		25.9%	Free (BH,CA, CL,IL,MX, P,SG) 5.3% (JO) 7.8% (MA) 15.5% (AU)	90%
	10	Men's (659)	doz. kg			
	20	Boys' (659)	doz. kg			
6112.39.00		Of other textile materials		13.2%	Free (BH,CA, CL,E*,IL,JO, MX,P,SG) 3.9% (MA) 11.8% (AU)	90%
	10	Of cotton (359)	doz. kg			
		Other:				
	15	Containing 70 percent or more by weight of silk or silk waste (759)	doz. kg			
	90	Other (859)	doz. kg			
		Women's or girls' swimwear:				
6112.41.00		Of synthetic fibers		24.9%	Free (BH,CA, CL,IL,MX, P,SG) 5.1% (JO) 7.5% (MA)	90%

Different taxes for same apparels for men and women	 ***Cotton pajamas***	 ***Fur felt hats***	 ***Coats***
Men and boys	**8.9%**	**0**	**38.6¢/kg + 10%**
Women and girls	**8.5%**	**96¢/doz + 1.4%**	**64.4¢/kg + 18.8%**

Figure 26: Different taxes for same apparels for men and women.

garments are different, for no apparent reason (Mic07b), (Mic07a). For instance, cotton pajamas for men and boys are imposed a tariff of 8.9 percent whereas those ones for women and girls are imposed a tariff of 8.5 percent. Or coats which contain 23 percent or more by weight of wool or fine animal hair are imposed a tax of 38.6 cents per kilo plus 10 percent ad valorem for men whereas up to 64.4 cents per kilo plus 18.8 percent ad valorem for women as shown in Fig. 26.

This problem is really without rhyme or reason to the tariff system but it happened. Barbaro calculated that the government imposes a 14 percent tariff on women's, but only 9 percent on men's on overall (Mic07b). According to data and evaluation of (MG08) and (Mic07b), US importers have overpaid more than 1.3 billion dollars for discriminatory duties.

Consequently, a number of clothing makers filed a lawsuit alleging discrimination in tariff on the ground of gender like Steve Madden, Asics and Columbia Sportswear, etc (MG08), (Mic07a), (Mic07b). They hoped this case would force the tariffs on similar items to be equalized. These enterprises even expected that they might win damages for the higher tariffs they paid on certain items - an idea that seems to presume that the lower rates are somehow more legitimate in all cases. Lawyers worked hard to find the covert motivation for this one of greatest-considered-mysteries in retailing. Its first appearance was dated back to the mid 1800s to protect domestic textile manufacture. However, it is no longer appropriate to the present situation when US has committed to free trade when they joined WTO and when human beings are unceasingly trying to build a better life equal to every individual. Yet, this complaint was dismissed by the U.S. Court of International Trade (USCIT) since the Court found that the argument and evidence Totes alleged were inadequate and not persuasive to claim that the US government offended against constitution: "In order to state such a claim for violation of the equal protection clause based on gender, Totes must allege that the government has engaged in gender-based discrimination without an exceedingly persuasive justification, or in other words, that the government has used discriminatory means that are not substantially related to important government objectives, ... In so doing, Totes' complaint must include a factual allegation that demonstrates a governmental purpose to discriminate." (USCIT. 2008: p14)

From the point of view of knowledge discovery, there is a possibility for applying data mining techniques to detect possible relations between attribute(s) of apparel (gender, kind of apparel, materials, etc) and the discrimination in tariff. And if it exists, then potentially discriminatory bases of the problem may be revealed. Moreover, if the domain of apparels can be organized in a hierarchical schema, new information about unequal taxes imposed on each kind of garment will be even unmasked at multiple levels of abstraction.

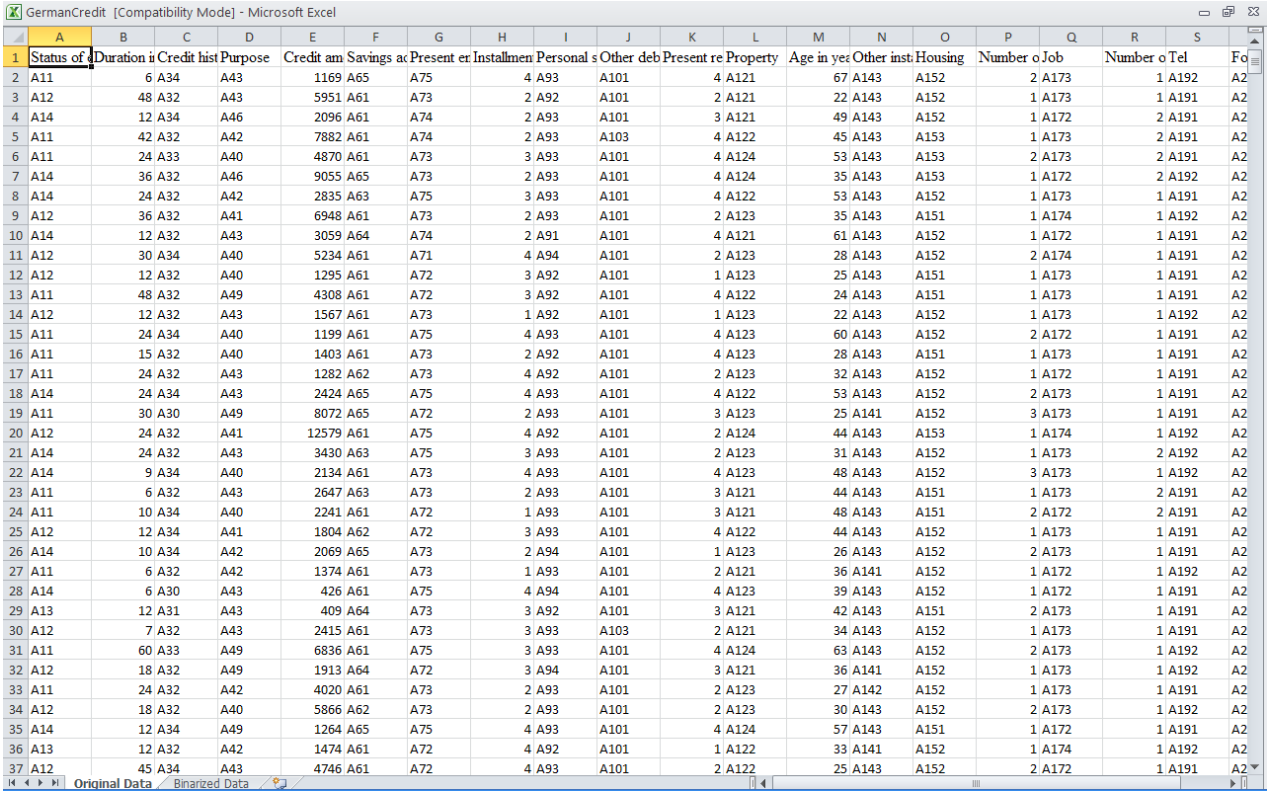
4.3.2 German credit data

The German credit data provided by the UC Irvine (UCI) Machine Learning Repository (NHBM98) provides ranking of clients of a bank applying for a loan. These clients are described by a set of 20 attributes from which each applicant was rated as "*good*" (700 cases fulfilled terms of credit agreement) or "*bad*" (300 cases defaulted on loan payments). Example of this dataset is illustrated in Fig. 27. Generally, attributes of applicants can be categorized as:

- attributes on private life such as *age, marital status, sex, foreign worker, residence since*.
- attributes on credit information like *credit history, existing credits, other installment plans*.
- attributes on details of the application, for example, *the purpose of the loan, co-applicant or guarantor* of the application.

Due to first analysis results of (PRT08), (PRT07), it appears that the assessment of applicants is negatively impacted by protected-by-law attributes like age, gender, marital status. The explored results, however, are just represented as *flat* association rules, which may lead to limited expressivity. It would provide richer information when representing this discrimination at different levels of abstraction.

The next section proposes a 3-step framework as presented in Fig. 28 which is able to analyze semi-structured business data to detect possible discriminatory associations. First, a knowledge base is built in the form of an ontology in which the hierarchical structure of the data domain is represented by the TBox. After the input data is pre-processed with support of semantic rules, clean data are transferred to ABox as individuals. Second, matching analysis is implemented to "*binarize*" the non-binary target attribute, which makes data mining process feasible. Subsequently, *Potentially-Discriminatory* (PD) Patterns are semi-automatically defined by components of the ontology. They are used in the reasoning service over the Abox to discover instances satisfying these patterns.



GermanCredit [Compatibility Mode] - Microsoft Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	Status of e	Duration i	Credit hist	Purpose	Credit am	Savings ac	Present er	Installmen	Personal s	Other deb	Present re	Property	Age in yea	Other inst	Housing	Number o	Job	Number o	Tel	Fo
2	A11	6	A34	A43	1169	A65	A75	4	A93	A101	4	A121	67	A143	A152	2	A173	1	A192	A2
3	A12	48	A32	A43	5951	A61	A73	2	A92	A101	2	A121	22	A143	A152	1	A173	1	A191	A2
4	A14	12	A34	A46	2096	A61	A74	2	A93	A101	3	A121	49	A143	A152	1	A172	2	A191	A2
5	A11	42	A32	A42	7882	A61	A74	2	A93	A103	4	A122	45	A143	A153	1	A173	2	A191	A2
6	A11	24	A33	A40	4870	A61	A73	3	A93	A101	4	A124	53	A143	A153	2	A173	2	A191	A2
7	A14	36	A32	A46	9055	A65	A73	2	A93	A101	4	A124	35	A143	A153	1	A172	2	A192	A2
8	A14	24	A32	A42	2835	A63	A75	3	A93	A101	4	A122	53	A143	A152	1	A173	1	A191	A2
9	A12	36	A32	A41	6948	A61	A73	2	A93	A101	2	A123	35	A143	A151	1	A174	1	A192	A2
10	A14	12	A32	A43	3059	A64	A74	2	A91	A101	4	A121	61	A143	A152	1	A172	1	A191	A2
11	A12	30	A34	A40	5234	A61	A71	4	A94	A101	2	A123	28	A143	A152	2	A174	1	A191	A2
12	A12	12	A32	A40	1295	A61	A72	3	A92	A101	1	A123	25	A143	A151	1	A173	1	A191	A2
13	A11	48	A32	A49	4308	A61	A72	3	A92	A101	4	A122	24	A143	A151	1	A173	1	A191	A2
14	A12	12	A32	A43	1567	A61	A73	1	A92	A101	1	A123	22	A143	A152	1	A173	1	A192	A2
15	A11	24	A34	A40	1199	A61	A75	4	A93	A101	4	A123	60	A143	A152	2	A172	1	A191	A2
16	A11	15	A32	A40	1403	A61	A73	2	A92	A101	4	A123	28	A143	A151	1	A173	1	A191	A2
17	A11	24	A32	A43	1282	A62	A73	4	A92	A101	2	A123	32	A143	A152	1	A172	1	A191	A2
18	A14	24	A34	A43	2424	A65	A75	4	A93	A101	4	A122	53	A143	A152	2	A173	1	A191	A2
19	A11	30	A30	A49	8072	A65	A72	2	A93	A101	3	A123	25	A141	A152	3	A173	1	A191	A2
20	A12	24	A32	A41	12579	A61	A75	4	A92	A101	2	A124	44	A143	A153	1	A174	1	A192	A2
21	A14	24	A32	A43	3430	A63	A75	3	A93	A101	2	A123	31	A143	A152	1	A173	2	A192	A2
22	A14	9	A34	A40	2134	A61	A73	4	A93	A101	4	A123	48	A143	A152	3	A173	1	A192	A2
23	A11	6	A32	A43	2647	A63	A73	2	A93	A101	3	A121	44	A143	A151	1	A173	2	A191	A2
24	A11	10	A34	A40	2241	A61	A72	1	A93	A101	3	A121	48	A143	A151	2	A172	2	A191	A2
25	A12	12	A34	A41	1804	A62	A72	3	A93	A101	4	A122	44	A143	A152	1	A173	1	A191	A2
26	A14	10	A34	A42	2069	A65	A73	2	A94	A101	1	A123	26	A143	A152	2	A173	1	A191	A2
27	A11	6	A32	A42	1374	A61	A73	1	A93	A101	2	A121	36	A141	A152	1	A172	1	A192	A2
28	A14	6	A30	A43	426	A61	A75	4	A94	A101	4	A123	39	A143	A152	1	A172	1	A191	A2
29	A13	12	A31	A43	409	A64	A73	3	A92	A101	3	A121	42	A143	A151	2	A173	1	A191	A2
30	A12	7	A32	A43	2415	A61	A73	3	A93	A103	2	A121	34	A143	A152	1	A173	1	A191	A2
31	A11	60	A33	A49	6836	A61	A75	3	A93	A101	4	A124	63	A143	A152	2	A173	1	A192	A2
32	A12	18	A32	A49	1913	A64	A72	3	A94	A101	3	A121	36	A141	A152	1	A173	1	A192	A2
33	A11	24	A32	A42	4020	A61	A73	2	A93	A101	2	A123	27	A142	A152	1	A173	1	A191	A2
34	A12	18	A32	A40	5866	A62	A73	2	A93	A101	2	A123	30	A143	A152	2	A173	1	A192	A2
35	A14	12	A34	A49	1264	A65	A75	4	A93	A101	4	A124	57	A143	A151	1	A172	1	A191	A2
36	A13	12	A32	A42	1474	A61	A72	4	A92	A101	1	A122	33	A141	A152	1	A174	1	A192	A2
37	A12	45	A34	A43	4746	A61	A72	4	A93	A101	2	A122	25	A143	A152	2	A172	1	A191	A2

Original Data | Binarized Data

Figure 27: Example of German credit dataset.

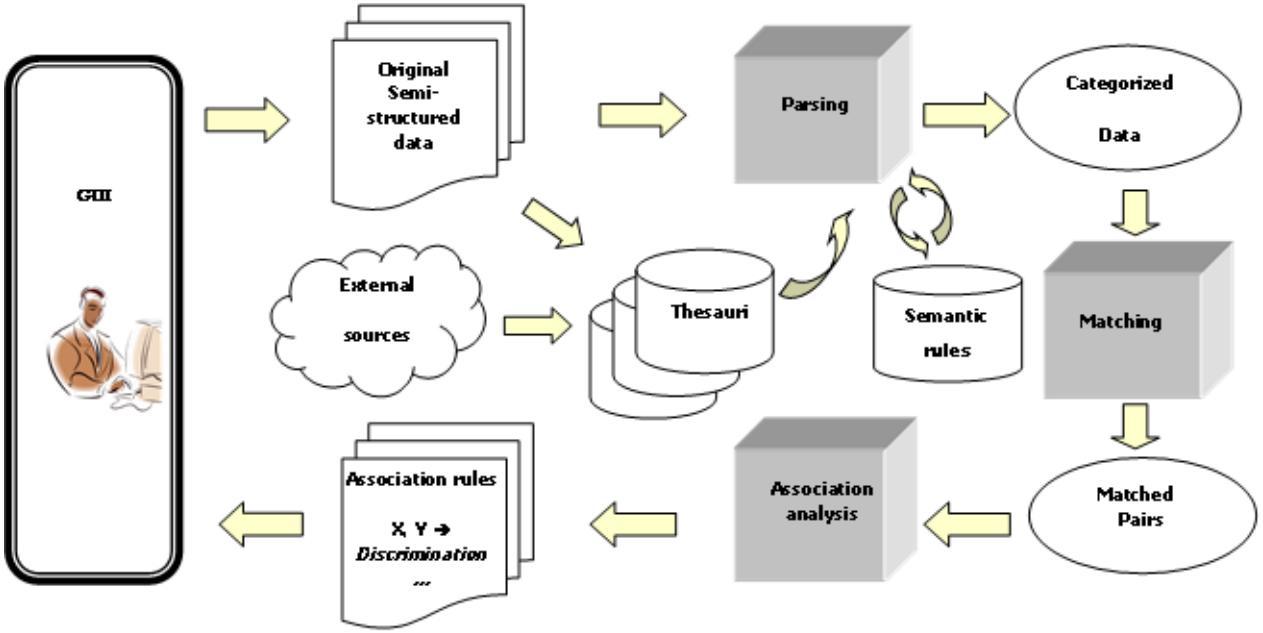


Figure 28: Framework of discriminatory association analysis on semi-structured data.

The data mining engine then searches over obtained instances for generalized discriminatory association rules. Let us expound components and steps of the framework.

4.4 Generalized discrimination discovery on semi-structured data supported by ontology

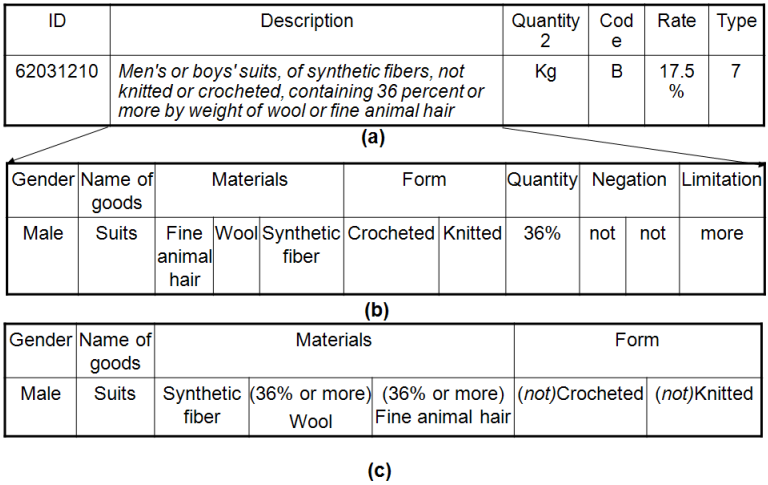
In order to provide a coherent content, the part of building the knowledge base in the form of an ontology will be presented after the first two steps. Therefore, the framework is introduced in the following sequence:

- Parsing: pre-processes data so that they are ready for the mining process.
- Matching analysis: defines an intermediate binary parameter as an alternative to the non-binary target attribute by pairs of matching items.
- Association analysis: formalizes the ontology for the purpose of discrimination discovery, and implements association mining on this ontology.

4.4.1 Pre-processing semi-structured data for discrimination discovery

In order to be adaptive to associations mining for discrimination, semi-structured data need to be transformed to the well-complete format of pairs of attribute/value. We suggest using semantic analysis borrowed from the idea of *Post of Speech - PoS* technique in syntax analysis of text mining (Vou95b), (JH00) supported by a common sense knowledge base.

Fig. 29a represents a typical case of not-well-structured raw data extracted from the HTS data introduced in Section. 4.3.1. It is obvious that the *description* field des can be still divided into more detailed subfields (sub-items) as shown in Fig. 29c with categories such as *gender*, *name of goods*. Each description of the input data can be formalized as:



ID	Description	Quantity 2	Cod e	Rate	Type
62031210	*Men's or boys' suits, of synthetic fibers, not knitted or crocheted, containing 36 percent or more by weight of wool or fine animal hair*	Kg	B	17.5 %	7

(a)

Gender	Name of goods	Materials			Form		Quantity	Negation		Limitation
Male	Suits	Fine animal hair	Wool	Synthetic fiber	Crocheted	Knitted	36%	not	not	more

(b)

Gender	Name of goods	Materials			Form	
Male	Suits	Synthetic fiber	(36% or more) Wool	(36% or more) Fine animal hair	*(not)*Crocheted	*(not)*Knitted

(c)

Figure 29: Example of transformation from semi-structured data to well-structured data.

$$des_i = \{(w_i + \varepsilon_i), (t_i + \zeta_i)\}$$

where:

- w_i, t_i are meaningful terms.
- ε_i, ζ_i are considered noise or redundant words, which should be cleaned.

In our framework, the original data will be firstly combined with external sources of information such as WordNet's thesaurus (grob) or Roget's thesaurus (groa) to build the common sense knowledge base. It is structured as a hierarchy of components (if any) and attributes of the sub-items extracted from des. From the structure of the common sense knowledge base, thesauri are built to support the extraction of sub-items from *des*. Particularly, each description will be parsed with the help of these thesauri, for which their elements are normalized and categorized as sub-items in the form of attribute/value. There are two groups of thesauri:

- *Form thesauri* FT: serve the purpose of normalizing all different words so that the description contains only "standard word" of the knowledge domain, e.g. thesaurus of synonyms, abbreviations. Each *Form thesaurus* consists of a set of pairs of terms that are similar in semantics: $FT_k = \{(t_{ki}, t_{kj})\}$, $k = 1, ..., p$ where p is the number of Form thesauri. We call t_{ki} the *relevant term*, and t_{kj} the *normalized term*. For instance, ("*not*", "*excluding*") is an entry of the *Negative* thesaurus. If any word extracted from the description matches a relevant term of an entry of a certain thesaurus, e.g. "*excluding*", it will be replaced by its corresponding term - the *normalized term*, and in this case, that term is "*not*".

- *Category thesauri* CT: clarify categories (attributes) of normalized terms of Form thesauri: $CT_i = \{t_{ij}\}$, $i = 1, ..., q$ where q is the number of Form thesauri. For example, the thesaurus of *Materials* of goods: silk, fur, ...

Therefore, when parsing the description, redundant information such as stop words, for example, "to", "the", "of", "and", ...is removed; ambiguous words are transferred into unique categories. Thus, after the parsing, all meaningful terms will be categorized into a particular thesaurus of the *Category* group (a particular attribute). In other words, the original description item is transformed into pairs of attribute/value as well-defined sub-items (called clothing sub-itemset):

$$des_i = \{t_{ij}\}, i = 1, ..., n, \text{and } j = i_1, ..., i_m,$$

where:

- t_{ij} is a categorized term, it is an entry of a specific Category thesaurus: $t_{ij} \in CT_i$.

- n is the number of descriptions des, and

- i_m is the number of terms of a des_i.

Fig. 29b provides result of parsing the example given in Fig. 29a.

Algorithm Semantic_Rule_Forming (dataset of category termed descriptions D)

BEGIN

$R = \varnothing$ // R is the set of semantic rules

R' = simple initial rules added by expert users built from category thesauri

ForEach description d_i in dataset D

begin

seq_{ij} = sequence of category term t_{ij} in d_i

$r_i = \{(seq_{ij}, t_{ij})\}$ // build order of categories of d

$R_i = r_i \cup R'$ //if r_i satisfies initial rule(s) of R',

//add both to the combinatorial semantic R_i

If $R_i \not\subset R$ Then $R \leftarrow R_i$ // if R_i is new to R then add R_i to R.

end

END.

Figure 30: Semantic_Rule_Forming algorithm.

However, these data do not convey the semantic relationships between sub-items. For instance, in a garment database: a number followed by a material often specifies the quantity of that material in a clothing item, e.g. "less than 30% of cotton". We propose to represent these semantic relationships by means of semantic rules that are formalized by sub-items through syntactic analysis step added to the parsing. This idea is borrowed from Part-of-Speech (PoS) technique, in which extracted categorized terms of each description are sequentially checked through rules. Typical semantic rules have the form of sequences of items of category thesauri, specifying their relationships: $r_i = \{(seq_i, t_{ij})\}$. For example:

Gender-name-quantity-materials-forms of production
Gender-name-materials 1-parts-materials 2 -forms of production

The first rule above tells the system that the *quantity* supports the *material* element (not others) of the extracted clothing sub-itemsets if the checked sub-itemset satisfies the order of the rule. Also, it might happen

the case when there are sub-items of the same category but they semantically support sub-items of different categories. The "*level*" information is then used to classify these relationships. The second rule states that the first *materials* belonging to the object level (*name*) have level 1; whereas the second *materials* belonging to the component of object (*parts* item) have level 2 if the checked sub-itemset satisfies the order of the rule.

In order to implement semantic parsing, we introduce the *Semantic_Rule_Forming* algorithm in Fig. 30. The algorithm scans terms of each description, finds and stores order of category terms then combines this order with support from (expert) users to extract semantic rules. First rules are cultivated by users, which are their experience in a domain. They are relatively simple rules, for instance, an initial rule for a clothing database:

(number "%" + materials) + forms

It means a percentage number followed by a material item(s) and form item(s) will reveal the amount of that *material(s)* item but not the *form* item(s) in a particular garment product. Subsequently, complex rules are generated by combining these initial rules in tandem or with other items of category thesauri when scanning the database. Particularly, when the sequence of all terms of a given description is specified, it is checked through the store of rules. If that sequence is not found in any rule, it is considered as a new rule, and will be added to the store. This process might need the supervision of experts. At the end of the semantic parsing, all descriptions are attached to semantic rule.

Finally, the description is a well-defined structure with semantics constraints as shown in Fig. 29c:

$$des_i = \{(seq_i, t_{ij})\}$$

4.4.2 Matching analysis

It is said that a decision system is discriminatory if there are bias treatments on the same context for varied protected-by-law attributes. However, it is not easy to expose this context, especially when the decision is

not binary, i.e. simply *yes* or *no* since there is no inherent base to judge discrimination in a vast number of data. Thus, it is needed to define an "intermediate" binary parameter as a substitute for the non-binary target attribute. This parameter can be activated when ascertaining itemsets or more specifically *individuals* having the same background information but different (or opposite) sensitive attributes. If their target items are different, it can be set as "*Yes*", and "*No*" otherwise. Actually, its objective is the same as the idea of *situation testing* as presented in Chapter 3. The discrimination mining is then feasible on the whole dataset since contexts at different levels of generalization can be built by loosening the highly specific background information of the matching items. In particular, for each individual (itemset):

- Find individuals (itemsets) that are different in the sensitive attributes S (PD attributes)
- For each individual (itemset) found in the previous step, comparing their PND attribute. If there is any individual (itemset) which is completely the same in all non-sensitive attributes, it is said that they match each other.

For example, the data in Fig. 29a will match the following original data:

"*Women's or girls' suits, not knitted or crocheted, of synthetic fibers, containing 36 percent or more of wool or fine animal hair.*"
which has nearly the same sub-items as the itemset presented in Fig. 29c except the *gender* sub-attribute ("*female*" vs "*male*") as illustrated in Fig. 31.

Nevertheless, it is often difficult to discover matching pairs even when descriptions are analyzed as in Fig. 29c. Theoretically, it is required to exhaustively scan the database, which is not always feasible, especially when the dataset is huge. We propose an approach as in Fig. 32 for seeking matching itemsets using semantic rules. Given an itemset, this algorithm checks only itemsets satisfying the semantic rule of that itemset. Hence, it obviates the need to scan the whole dataset and all the attributes as well, which helps to increase both performance and precision.

Definition 1 *Let* (u_i, v_i, c_i), (u_i, v_i', c_i') *be two triples, where:*

ID = 62031210

Gender	Name of goods	Materials			Forms	
male	suits	synthetic fibers	(*36% or more*) wool	(*36% or more*) fine animal hair	(*not*) crocheted	(*not*) knitted

ID = 62041310

Gender	Name of goods	Materials			Forms	
female	suits	synthetic fibers	(*36% or more*) wool	(*36% or more*) fine animal hair	(*not*) crocheted	(*not*) knitted

Figure 31: An example of matching itemset.

- u_i *is the set of PND items, and*
- v_i, v_i' *are items referring to alternative PD items, such as the female gender versus male gender,*
- c_i, c_i' *are the target items.*

the discriminatory indicator θ *is defined as follows:*

$$\theta = \begin{cases} Yes & \text{if } c_i \neq c_i'; \\ No & \text{if } c_i = c_i' \end{cases} \tag{4.4}$$

The above definition can be explained as follows: For any itemset, the discriminatory indicator is activated when another itemset which is different in values of PD attribute(s) and target attribute on the same context is found. It means that the multiple-valued target item is replaced by a binary item, the discriminatory indicator (θ is Yes or No). This result is extremely important since it is now feasible to implement the mining of discriminatory relations between the PD attribute(s) and the target attribute by the means of discriminatory indicator. Indeed, when applying the mining process to the original data, the possible achieved results

```
Algorithm MatchFindingByRules(itemset t_i)
   BEGIN
     isMatch = true
     R_i = the semantic rule of t_i
     v_ij = sensitive attribute j of t_i
     C_i = {t_ij| t_ij |= R_i}//the set of itemsets satisfy semantic R_i
     ForEach itemset t_ik in C_i
     begin
            v_kj = sensitive attribute j of t_ij
            If v_kj ≠ v_ij Then
                   ForEach non-sensitive attribute u_i in AttributeSet
                          If u_ii ≠ u_iik Then
                          begin
                                 isMatch = false
                                 break
                          end

            If (isMatch)
            Then
                   Add itemset t_ij to MatchedList;
            Else
                   isMatch = true
     end
END.
```

Figure 32: Matching algorithm based on semantic rules.

might be limited as the volume of candidates of frequent sets is small because of the multiple-valued target attribute. If this non-binary target attribute is substituted by the binary discriminatory indicator, these candidates will include the discriminatory indicator and other normal attributes, which makes it easier to collate tuples of data. Thus, the volume of candidates of frequent sets is improved, and more hidden rules are exposed in comparison to outcome extracted from the original data. If these relations are discovered, then discriminatory treatments can be confirmed.

If only one discriminatory indicator is set, it is hard to precisely identify the effect of each PD attribute (if there are more than one PD attributes) on the target attribute. Hence, each PD attribute has its own discriminatory indicator. The individual effect of each PD attribute on the target attribute is then calculated by the support of that discriminatory indicator on the whole database. In particular, the following situa-

tions may happen:

- there are a number of itemsets satisfying the triple:

$$(\{u_k\}, v_i, \theta = Yes)$$

  In this case, the following association can be generated:

$$\{u_k\}, v_i \rightarrow Discrimination \tag{4.5}$$

  where $\{u_k\}$ is a set of PND items forming the context, whereas v_i is a single PD item. When $\{u_k\}$ is empty, it means that the discrimination caused by the PD attribute does not depend on a specific case, which is the simplest case.

- there are a number of itemsets satisfying the triple:

$$(\{u_k\}, \{v_i\}, \{\theta_i = Yes\})$$

  In this case, the following association is generated, one for each PD item in the set $\{vi\}$:

$$\{u_k\}, \{v_i\} \rightarrow Discrimination \tag{4.6}$$

  where $\{v_i\}$ is a set of more than one PD items.

4.5 and 4.6 are called *discriminatory association rules*. These rules are interpreted that the presence of a PD attribute v_i (and correspondingly v_i) will lead to discrimination shown in the target attribute in a context built by given PND attributes $\{u_k\}$. Particularly, changes in the target attribute are strongly related to the PD attribute v_i given the context $\{u_k\}$. Thus, it can be said that the sensitive attribute v_i negatively affects the target attribute.

4.4.3 Association analysis

If only flat discriminatory associations are required, direct mining can be then implemented to compute the confidence of each of the possible

discriminatory associations mentioned in the previous step. Examples of flat discriminatory association rules will be presented in the chapter of Experiments.

However, flat association rules are proved to have weak descriptive power in the sense that the retrieved knowledge is insufficiently either general or detailed. Hence, generalization of these association rules is a reasonable solution for rich semantic association mining. In this section, we introduce the work of mining discrimination at different levels of abstraction on ontology.

The generalized discriminatory association analysis step is centered around a knowledge base in the form of an ontology. Data is populated from relational databases or external resources. The TBox of the ontology represents the hierarchical structure of the data domain, whereas individuals of the ABox are well-structured data transferred from the relational databases. Background knowledge is modelled by a domain expert, while patterns of possible discrimination are specified by an anti-discrimination analyst by defining a few concepts of interest, as described in Section 4.4.4. These patterns are used in the reasoning service over the Abox for discovering instances satisfying these patterns. The system for discrimination discovery, called discrimination reasoner and described in Section 4.4.5, extracts from the ontology individuals that satisfy patterns of possible discrimination and that, according to the legal methodology of auditing, can be considered as discriminated. Starting from such a set, generalized classification rules summarizing contexts of discrimination are calculated, and ranked on the basis of the number of discriminated individuals (rule support), and on the precision of the summarization (rule confidence). Details of this step are displayed in Fig. 33.

4.4.4 Ontology Structure for Discrimination Discovery

Typically, an ontology $\mathcal{O} = \langle \mathcal{T}, \mathcal{A} \rangle$ is designed by domain experts (as per the TBox), and populated (as per the ABox) from external data sources by means of schema-mapping or ETL processes. For the example of garment products, categories of clothing items are gathered in the *Apparels*

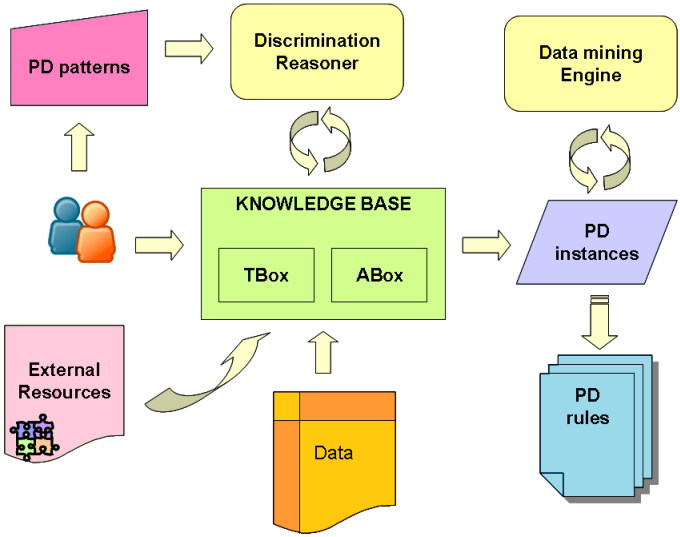


Figure 33: Generalized discriminatory association analysis.

concept, while other relational attributes are modelled as object properties, e.g., *hasGender*, *hasMaterial*, *hasPrice*. We assume that the following concepts are at top level in an ontology designed for supporting discrimination discovery. They are set up by the anti-discrimination analyst as a preliminary step of the analysis.

Relevant concepts. While ontologies often contain a large set of concepts, which intertwine to describe a knowledge domain, we assume a main concept that is the subject of the discrimination discovery problem. In our running example, this is the *Apparels* concept, because we are interested in unveiling disparate taxation practices in the HTS system. We call such a main concept and its sub-classes the *relevant concepts*.

PD and PND attributes. Object properties of relevant concepts link to *Potentially Discriminatory* (PD) or *Potentially Non-Discriminatory* (PND) groups of concepts, on the basis of whether they refer or not to sensitive or non-sensitive personal attributes respectively. According to anti-discrimination laws, PD attributes include gender, age, marital status,

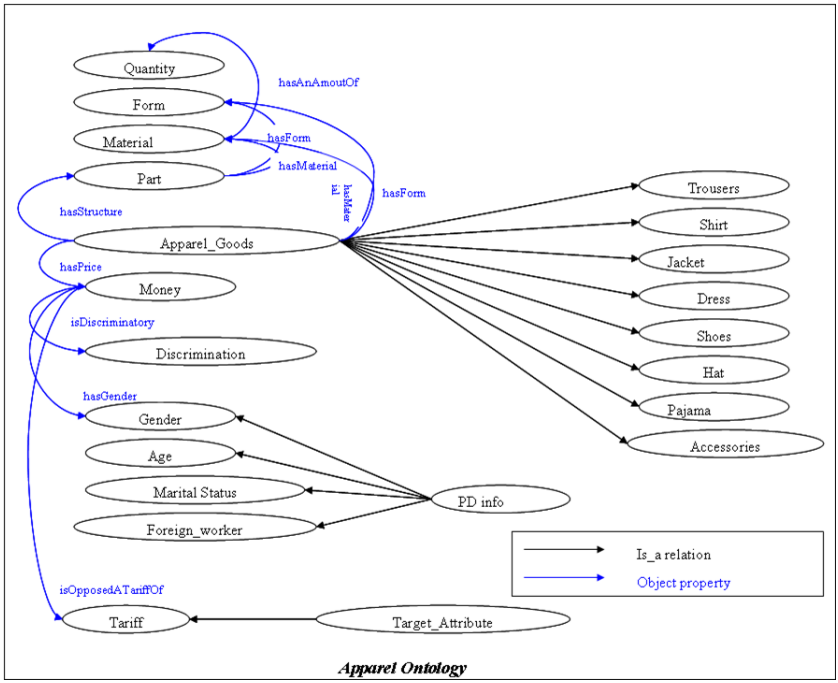


Figure 34: HTS ontology

nationality, ethnicity, and so on. In our running example, the only sensitive attribute is the gender the garment is produced for (specifically, we will consider the female gender, which is protected by the anti-discrimination laws), and the object property *hasGender* connects *Apparels* to the PD concept *Gender*. As another example, the property *hasPrice* connects *Apparels* to the PND, or non-sensitive, concept *Money*.

Target attribute. A specific concept, called the *target attribute*, is assumed to model the decision that may have discriminatory effects. In our example, it is the taxation applied to a garment, expressed as a percentage value. In Fig. 52, the *Tariff* concept models the target attribute, and the *isAppliedATariffOf* object property specifies the amount of tax applied to a garment. We use the meta-variable τ to represent the decision value for an individual. The domain of τ can be discrete, e.g., a yes/no

decision to a loan application, or continuous, as in our running example. In the HTS example, a taxation tariff τ_1 is worse than τ_2 if $\tau_1 > \tau_2$.

Finally, we assume a *discriminatory attribute* concept, taking only yes/no values, which is linked to relevant concepts via the $isDiscriminatory$ object property. Intuitively, such a property is intended to label as discriminated or not each individual belonging to relevant concepts on the basis of its PD, PND, and target attributes. In our example, it labels as discriminated or not each apparel on the basis of the gender it is produced for, of its characteristics (material, form, quantity, etc.), and on the basis of the tariff applied to it. The formal definition of the $isDiscriminatory$ property is presented in the next subsection.

Importing data

After defining the structural hierarchy of the knowledge domain as the TBox of the ontology, instances and assertions are inserted in the Assertion axiom (A-Box) of the knowledge base as the input of the mining process. Those assertions are for particular instances that can relate an instance to its appropriate class or two instances to each other. In fact, these instances are specific items of transactions in the databases. They can be copied from relational database tables to the corresponding class in the ontology via a mapping table by establishing the match of tables and their attributes in database with classes and properties in ontology. However, this method works just in case the data are already well-defined. When data are complex and contain overlapping relationship such as the HTS data with information about materials of the whole garment and several parts of the garment, that direct mapping does not work well. In addition, this mapping does not allow users to maintain logic relationships, for instance one kind of shoe is made from "*silk*" whereas its outer is made from "*rubber*". Thus, it is really time-consuming and cumbersome to fill in thousands of those axioms even in a small database like the HTS. Hence, tagging format like XML file can be a good way to maintain their relationships. Besides, using XML format allows a more flexible mechanism of mapping. Moreover, it is more convenient to access the data when storing them in an XML file than in a database. In

our approach, data achieved after the analysis can be directly converted to an XML file without transferring them to the database.

4.4.5 Modelling Discrimination

In social and legal sense, discrimination occurs in contexts when a person is treated unequally or less favorably than the another in the same conditions. We model contexts by expressions Δ of the form:

$$\Delta = C \wedge R_1 \wedge ... \wedge R_t$$

where C is a relevant concept, at any level of the hierarchy, and $R_1, \ldots, R_t$ are object properties that link C to PND properties. As an example, the context:

$$Outerwear \wedge hasMaterial \wedge hasForm \tag{4.7}$$

mentions a kind of garments, $Outerwear$, that are made by some (not further specified) material and that have some (not further specified) form.

The legal methodology of *auditing* (Ben07a; Ror09), also known as field experiments or situation testing, follows a quasi-experimental approach to investigate for the presence of discrimination by controlling the factors that may influence decision outcomes. It consists of using pairs of *testers* (also called *auditors*), who have been matched to be similar on all characteristics that may influence the outcome except race, gender, or other grounds of possible discrimination. The tester pairs are then sent into one or more situations in which discrimination is suspected, e.g., to rent an apartment or to apply for a job, and the decision outcome is recorded. A different outcome between the paired testers is then considered a *prima-facie* evidence of discrimination.

We rephrase the auditing approach as follows. The concept C in Δ is used to select individuals as candidate testers, while the object properties $R_1, \ldots, R_t$ in Δ are used to compare individuals to select pairs of similar (w.r.t. those properties) characteristics. More formally, we define the *realization set* Φ of Δ over the ontology the set of pairs $(x, \{y_1, \ldots, y_t\})$, called *realizations* $\Phi(x, \{y_1, \ldots, y_t\})$, such that:

- $\mathcal{O} \models C(x)$
- for $i = 1 \ldots t$, $\mathcal{O} \models R_i(x, y_i)$

Therefore, the individual x is a candidate tester, while $y_1, \ldots, y_t$ are its properties in the context Δ. We look for another realization Φ' that can be paired with it by introducing a notion of similarity over ontologies. The path-distance of two concepts C_1 and C_2 in a hierarchy is defined as the number of edges in the shortest path between C_1 and C_2 (and in particular, it is 0 when $\mathcal{O} \models C_1 \equiv C_2$). The path-distance of two individuals x and y, denoted as $p(x, y)$, is the path-distance of the concepts C_x and C_y, whose x and y are direct instances of. The basic similarity measure $sim()$ between individuals x, y is defined as a monotonically decreasing function of their path distance. In experiments, we set:

$$sim(x, y) = 2^{-p(x,y)}$$

It shows that if two individuals belong to the same concept, their similarity is 1, otherwise it exponentially decreases with their path-distance. We extend similarity to a pair of realizations $\zeta_1 = \Phi_1(x_1, \{y_i^1\})$, $\zeta_2 = \Phi_2(x_2, \{y_i^2\})$ as follows:

$$sim(\zeta_1, \zeta_2) = \frac{sim(x, y) + \sum_{i=1}^{t} ssim(y_i^1, y_i^2)}{t + 1}$$

where $ssim()$ is a similarity function between scalar values. For nominal domains, it boils down to an equality indicator: $ssim(a, b) = 1$ if $a = b$, and $ssim(a, b) = 0$ otherwise. For continuous values, normalized in the interval $[0, 1]$, we assume:

$$ssim(a, b) = 1 - |a - b|.$$

It can be easily seen that sim ranges over $[0, 1]$.

The similarity measure sim is used to search for pairs of testers. Given a candidate ζ_1, we look for ζ_2 such that[1] $sim(\zeta_1, \zeta_2) \approx 1$, namely a realization ζ_2 with a similar individual and similar object properties as ζ_1.

[1] In our experiments on the HTS dataset, we have observed no significant difference between $sim(\zeta_1, \zeta_2) = 1$ and $sim(\zeta_1, \zeta_2) \geq 0.95$, and thus we simply require $sim(\zeta_1, \zeta_2) = 1$, namely we look for garments of the same type and with exactly the same characteristics. However, the sensitivity of the results to the approximation "≈ 1" may change from a domain to another.

With reference to the description (4.7), we look for individuals of the *Outerwear* concept that are similar with respect to the material they are made of (e.g., synthetic fiber) and form (e.g., knitted).

The legal notion of "different treatment", which is the basis for claiming discrimination, is interpreted as follows. Consider a realization $\zeta_1 = \Phi_1(x_1, \{y_i^1\})$. Let s_{ζ_1} be the PD attribute value(s), and τ_1 the target attribute value for the individual x_1. In our example, s_{ζ_1} is the gender the garment x_1 is produced for, and τ_1 is the tariff applied. We label ζ_1 as discriminated if there exists a realization ζ_2 close to ζ_1 (namely, $sim(\zeta_1, \zeta_2) \approx 1$), with $\zeta_2 = \Phi_2(x_2, \{y_i^2\})$, for which the PD attribute value s_{ζ_2} of x_2 is different from s_1 (in our example, it refers to another gender), and the target attribute value τ_{ζ_2} for x_2 is better than τ_{ζ_1} (in our example, taxation τ_{ζ_2} is lower than τ_{ζ_1}, i.e., $\tau_{\zeta_2} < \tau_{\zeta_1}$). We formalize such a auditing reasoning by introducing the following *discriminatory indicator*:

$$\theta_\Delta(\zeta_1) = \begin{cases} \text{yes} & \text{if there exists } \zeta_2 \text{ such that} \\ & \quad sim(\zeta_1, \zeta_2) \approx 1 \text{ and} \\ & \quad s_{\zeta_2} \neq s_{\zeta_1} \text{ and } \tau_{\zeta_2} < \tau_{\zeta_1} \\ \text{no} & \text{otherwise} \end{cases}$$

We are now in the position to label an individual x_1 as discriminated or not by setting the object property $isDiscriminatory$ of x_1 to the value of $\theta_\Delta(\zeta_1)$, where ζ_1 is the realization whose first element is x_1 and the second element is the set of its object properties values w.r.t. Δ. With reference to (4.7), an outerwear garment produced for women is labeled as discriminated if the amount of taxation applied is higher than the one applied to the same, or similar as per material and form, garment produced for men. Summarizing, for a fixed context Δ, the $isDiscriminatory$ property is populated for individuals in its realization set according to the discriminatory indicator θ_Δ. However, we are still faced with the problem of extracting a high-level, intelligible, characterization of the individuals labeled as discriminated. We resort to (generalized) classification rule mining by extracting rules of the form:

$$C(?x) \wedge R_{i_1}(?x, v_1) \wedge ... \wedge R_{i_n}(?x, v_t) \rightarrow isDiscriminatory(?x, yes)$$

that we call *discriminatory classification rules*. They are extracted from the subset of realizations having PD property values (in our case, from garments produced for women) enriched with their discriminatory indicator value. Here, $v_1, \ldots, v_n$ are PND attribute values characterizing specific properties of an individual $?x$ belonging to the concept C. Object properties $R_{i_1}, \ldots, R_{i_n}$ are a subset[2] of $R_1, \ldots, R_t$, or, in symbols, $\{i_1, \ldots, i_n\} \subseteq \{1, \ldots, t\}$. Classical interestingness measures, such as support and confidence, are applicable. The support of a rule is the percentage of individuals that satisfy both sides of the rule, while the confidence is the percentage of individuals satisfying the left hand side that are labeled as discriminated. Notice that we adopt a human-readable SWRL (Semantic Web Rule Language) syntax (**?**). In our running example, the rule:

$$Outerwear(?x) \wedge hasMaterial(?x, \text{“}synthetic\ fiber\text{”}) \wedge$$
$$hasForm(?x, \text{”}\ knitted\text{”}) \rightarrow isDiscriminatory(?x, yes)$$
$$(supp = 5\%, conf = 100\%)$$

summarizes that 5% of outerwear apparels produced for women are made of synthetic fiber and knitted and with a tariff that is higher than those produced for men; and that 100% of the outerwear apparels produced for women that are made of synthetic fiber and knitted is applied such a disparate taxation.

Mining over the ontology

The pseudo-code of the overall rule generation algorithm is shown in Fig. 35. For each concept C that is specified by the user as relevant for the analysis, the following steps are executed. First, all object properties of the concept are retrieved from the ontology to build a pattern

[2]Since differences between individuals are controlled by means of the discriminatory indicator, a discriminatory rule involving only a subset of $R_1, \ldots, R_t$ can be safely read as the fact that the individuals in such a subset are discriminated (in proportion to the confidence of the rule). This conclusion cannot be made by the early approaches using classification rule mining (PRT09b; RPT10).

Δ. The set of realizations Γ of such a pattern is then calculated, consisting of all individuals belonging to concept C and of its PND attributes. For each realization with PD property values, the discriminatory indicator is calculated, and the object property $isDiscriminatory$ of the individual is set according. After the inner loop, each individual in the realization set is then labelled as discriminated or not, on the basis of the legal methodology of auditing. A summarization of the conditions for which individuals are discriminated can be extracted via discriminatory classification rules, as reported in Fig. 35. It is worth noting that such rules could be computed by standard (generalized) classification/association rule mining algorithms by exporting the dataset consisting of tuples $(x, y_1, \ldots, y_t, \theta_\Delta(\zeta))$ for each realization $\zeta = \Phi(x, \{y_i\})$. In our actual implementation, described next, we rather exploit the capability of ontology management systems of answering queries, specifically SWRL queries. We have adopted the SWRL syntax for classification rules exactly with the purpose of using such a query language for computing the basic support counting primitives of (Apriori-based) association rule mining. While this is not an immediate technical advancement over reusing existing classification rule extraction algorithms, it demonstrates that the whole process of analysis can be coded within an ontology management systems.

To avoid exhaustively scanning the ontology hierarchy, we applied two optimizations:

- calculate the number of assertions for each group of subclasses in advance. Therefore, only items satisfying rule pattern will be checked for each scanning.

- apply the Apriori principle to reduce the number of generated rules by filtering unfrequent patterns.

Given a user-specified threshold (for the minimum support or minimum confidence), if any discriminatory association rule is retrieved, it will: i) prove that there is unjust discrimination in the given system on the bases of PD attributes and ii) reveal which attribute(s) causes that

for all relevant concepts C **do**
  let $R_1, \ldots, R_t$ be all object properties of C
  let $\Delta = C \wedge R_1 \wedge \ldots, R_t$
  let Γ = realizations(Δ)
  retract all $isDiscriminatory$
  for all PD $\zeta = (x, \{y_i\}) \in \Gamma$ **do**
    assert $isDiscriminatory(x, \theta_\Delta(\zeta))$
  end for
  extract rules with $supp \geq minsupp$
    and $conf \geq minconf$ of the form
      $C(?x) \wedge R_{i_1}(?x, v_1) \wedge ... \wedge R_{i_n}(?x, v_t)$
        $\rightarrow isDiscriminatory(?x, yes)$
end for

Figure 35: Discrimination rule generation algorithm.

discriminatory treatment. Moreover, among PND attributes causing discrimination, it might be found that some of them are not reasonable. If no convincing argument is given for the negative effect of such attributes in the discrimination matter, they should be considered wrong/illegal and removed in the decision-making processes for business.

4.5 Discrimination measures

A general principle mentioned in (Kno86) is to consider group under representation as a quantitative measure of the qualitative requirement that people in a group are treated "less favourably" than others (European Union Legislation 2009 (Eur10)), UK Legislation 2009 (U.K10). Or such that "a higher proportion of people without the attribute comply or are able to comply" (Australian Legislation 2009 (Aus10)) to a qualifying criterion. For the purpose of quantifying the level of discrimination, we also propose a number of measures applied on the discriminatory association rules. Whereas (PRT08), (PRT09b) use discrimination measure as a way to attain the expected level of discrimination in the system having binary target attribute, our relatively similar measures are used to obtain a precise vision of how discriminatory a non-binary decision system will

be when a sensitive attribute(s) (not value) appears in certain contexts. Measures are also used to investigate which PD values are more favored (benefiting a more positive outcome treatment for the whole itemset) in comparison with others.

Definition 2 *Let $s', s \rightarrow \theta_i = 1$ (discrimination) and $s' \rightarrow \theta_i = 1$ be association rules with confidences correspondingly α and ϕ. We have:*

- Absolute difference*:*

$$abs\_lift = |\alpha - \phi| \tag{4.8}$$

- Relative difference*:*

$$rev\_lift = |\alpha - \phi|/\phi \tag{4.9}$$

- Ratio gain of difference*:*

$$rg\_lift = \alpha/\phi \tag{4.10}$$

where:

- *$s' \in S'$ are PND items*
- *$s \in S$ is a set of PD items*

Many legislative documents regulate a threshold which is the acceptable maximum difference. For example, in the UK, a difference of 5% in confidence between female and male treatment is assumed by courts to be significant of discrimination against women (U.K10). The absolute difference reveals whether the discrimination caused by s_i will be gained or decreased when a certain extra PD attribute is added to the context which means that the discrimination will be worse when these two PD attribute go in tandem.

For the same meanings, the relative difference computes the relative reduction of equality with the presence of another PD attribute besides the context. The US Legislation (U.S78a), for instance, states that "a selection rate for any race, sex, or ethnic group which is less than four-fifths

(80%) of the rate of the highest rate will be generally regarded as evidence of adverse impact."

The ratio gain of difference tells how much the degree of difference will be amplified if more sensitive part is added to the context.

When $|s| = 1$ and both the following two discriminatory association rules are achieved:

$$s = v_1, s' \rightarrow \theta_i = 1,$$
$$s = v_2, s' \rightarrow \theta_i = 1$$

where v_1 and v_2 are two opposite values of the single PD attribute s_i, and the corresponding confidences are α and $1 - \alpha$. And through the matching step, θ_i is specified as the possible benefit such as a favored action like accepting a loan. We can compare by then the disparity between different values of the sensitive items such as what is the difference of income between male and female:

- *Absolute opposite difference*:

$$abs\_lift = 1 - 2\alpha \tag{4.11}$$

- *Relative opposite difference*:

$$rev\_lift = (1 - 2\alpha)/(1 - \alpha) \tag{4.12}$$

- *Ratio gain opposite difference*:

$$rg\_lift = (1 - \alpha)/\alpha \tag{4.13}$$

θ can be a numeric value, calculated as the average value of the itemsets sharing the antecedent of the rule.

4.6 Another approach

These proposals and (PRT08), (PRT09a) for discrimination discovery and (KCP09) for discrimination prevention have followed the legal principle of under-representation. Unfortunately, they suffer both from legal

weaknesses, due to the use of aggregation measures over undifferentiated sets of people, and technical limitations, such as the restriction to nominal attributes and decisions and to local models of discrimination. In order to overcome both the legal and the technical drawbacks, we will propose another approach, which is still inspired by the legal experimental procedure of situation testing. Practically, it looks for pairs of people with similar characteristics apart from membership to a protected-by-law group. We approximate situation testing by a variant of the k-nearest neighbor (k-NN) classification, which labels each tuple in a dataset as discriminated or not. Discrimination discovery then reduces to build a classifier providing a global description of the conditions where discrimination occurs. Discrimination prevention is tackled by a pre-processing step, changing the historical decision for tuples labeled as discriminated, before training a classifier.

To make the content of the thesis coherent, that part of the thesis will be presented in the Chap. 7, after the parts of system design and experiments of the generalized discrimination discovery on semi-structured data.

Chapter 5

System design

The system is designed to perform the following three tasks:

- Prepare data, particularly sanitize semi-structured data so that they are ready for discrimination mining.
- Define common contexts where the occurrence of discrimination is strongly related to the protected-by-law attributes by finding matching itemsets.
- Mine and represent discrimination over the ontology in the form of generalized association rules.

The architecture of the system is organized as in Fig. 36.

5.1 Pre-processing data

The input of this phase is an Excel file (as illustrated in Fig. 25, which is a database of garments, containing semi-structured data (the *description* field). The output is expected to be well-structured data: all fields are represented in the form of attribute/value pairs, and they are subsequently loaded onto the ontology. Thus, this phase is composed of two sub-steps:

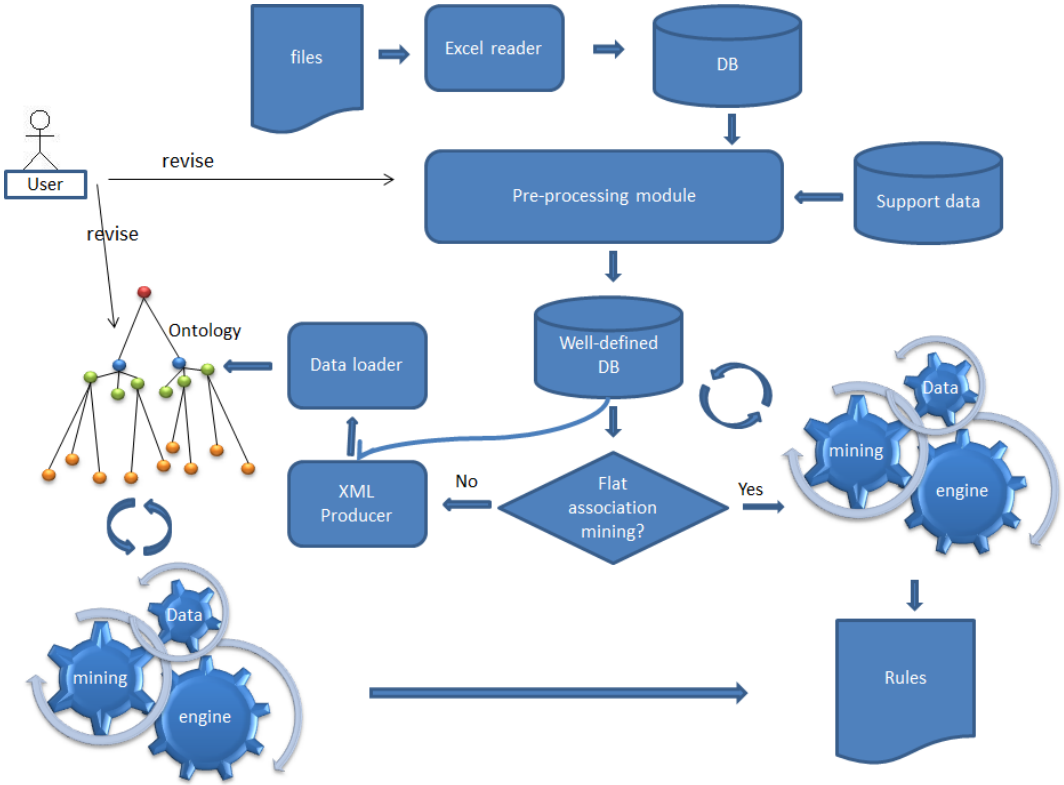


Figure 36: System architecture.

- Read the Excel file, extract and store data into the relational database. The *description* field is parsed in a way that its constituents are classified into pairs of attribute/value based on thesauri of the common knowledge base.

- Load the well-structured data onto the corresponding ontology as instances.

5.1.1 Pre-processing semi-structured data

In order to load data from the Excel input file, an ExcelImporter module is written to retrieve information saved in columns of the input Excel file and store these data into corresponding fields of a relational database. Generally, these data are well-structured (containing single component) except the *description* (containing multiple sub-components), which must be parsed to be well-defined form. We realize this task by the two modules *Normalize Description* and *Analyze Structure*:

- There is no warranty that the semi-structured data are clean, they thus need sanitizing before the classification. Hence, the module *Normalize Description* is designed to implement the task of data normalization by support of *Form thesauri*. Particularly, while parsing the *description*, extracted terms are checked with terms of thesauris. If a term matches with an entry of the thesauri, it is categorized due to that entry, e.g. abbreviations are replaced by full words of Abbrevation thesauri, synonyms are normalized by a the normalized term. The output of this module is almost clean descriptions, whose terms are sanitized.

- Subsequently, *Analyze Structure* extracts meaningful terms from these normalized data, and assigns them to the corresponding categories based on *Category thesauri*. The thesauri are built in advance for the given domain knowledge with support from users. They might be enriched while parsing descriptions by collecting new terms, which are constituents of descriptions. Review of users is certainly needed.

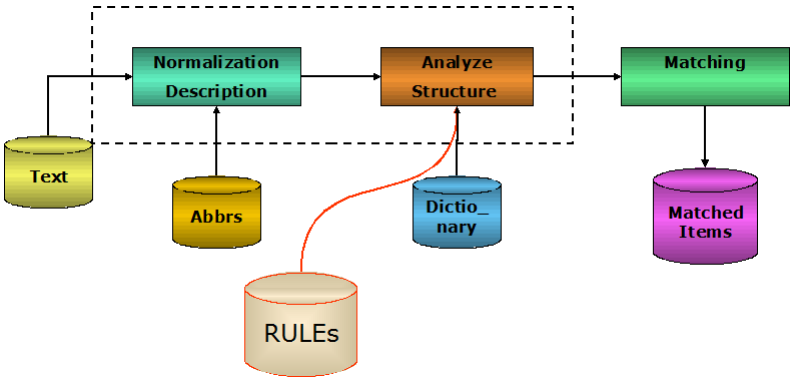


Figure 37: Pre-processing semi-structured data.

Also, the overall operations of the system should be controlled by (expert) users to increase its precision. The architecture of the whole unit is organized as in Fig. 37.

In a trivial parsing phase, sub-components of the descriptions are completely extracted. That is the descriptions are analyzed into pairs of attribute/value, and assigned to a specific category. It is then possible to conclude that the semi-structured descriptions are well-structured.

Yet these analyzed data do not convey the semantic relationships between constituents of the description. Since this information carries precious knowledge about the data object, it should be maintained in the processed data. Syntactic analysis borrowed from PoS technique of natural language processing can be a good idea to represent this dependence.

The general mechanism of syntactic analysis is to scan constituents (terms) of each categorized description, find and store order of categorized terms then compare with existing semantic rules as shown in Fig. 38. The initial semantic rules are generated by users, according to their experiences in a specific domain. These rules are relatively simple, they are *patterns* of sequences of categorized items. For instance, an initial rule for a clothing database *(number % + materials) + forms* means a percentage *number* followed by a *material* item(s) and *form* item(s) will

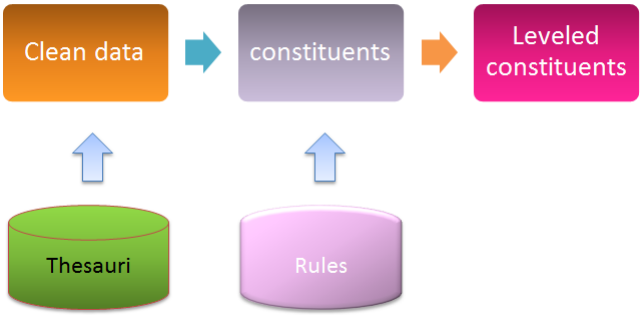


Figure 38: Semantic analysis module in data pre-processing.

reveal the amount of that *material* item(s) but not the *form* item(s) in a particular garment product. Subsequently, complex rules are generated by combining initial rules in tandem or with other items of *Category thesauri* if new sequences of category terms are found when scanning descriptions. This mechanism is realized in the *Semantic_Rule_Forming* algorithm (see Fig. 30). When it is applied to the experiment with the HTS dataset, we achieved a number of semantic rules as depicted in Fig. 39.

5.1.2 Load data onto ontology

For the purpose of resettling well-structured data available in relational databases in the form of rows in tables onto ontology, data need to be transformed into instances and assertions in the ABox of the knowledge base. Actually, there is an approach of mapping relational databases to ontology by establishing the correspondence between tables and their attributes in database with classes and properties in ontology. When data are complex and contain overlapping (semantic) relationship such as the HTS data with information about materials of the whole garment and several parts of the garment, that direct mapping does not work well. In this case, XML format is usually recommended as it allows a more flexible mechanism of mapping. Moreover, it is more convenient to access

Number	Structure
	Genders - names - main materials (except materials) – {"containing" + number % + specific materials} – {(not) forms}
	Genders - names – {(not) forms } – materials – (forms)
	Names – form — material – (not) gender
	Names – {parts + materials} – {part + (not) materials} – {(not) forms} – {(other than) gender } –{ "valued" + "not over" $}
	Names – {parts + materials} – forms – genders
	Names – {parts + materials} + ("nesoi") – {parts + (not) materials}
	Parts – name – materials – genders
	Genders - names - main materials (except materials) – {(not) forms} – {("containing" + number + specific materials}- (parts))
	Genders - names –{ (not) materials} –{ (not) forms} – (not) parts
	Genders - names – materials – {parts (forms)} – {(not) forms }
	Genders - names - main materials (not/except materials) – {(other than) parts + forms + materials} – (not) forms
	Genders - names – (not) forms – main materials (except materials) – {(not) "containing" + number % + specific materials} – (forms)
	Genders - names – (not) forms – {"containing" + number % + materials}
	Genders - names – {"containing" + number % + materials} – (not) forms
	Genders – forms – names – materials
	Genders – names – {parts+names }– materials (except materials) – {"containing" + number % + materials} – forms

Figure 39: Semantic rules for the HTS problem.

and manipulate data when they are stored in an XML file rather than in a database. Hence, we choose to convert obtained data to the XML format. It can even deployed without transferring the clean data to the relational database.

First, data (retrieved after the analysis) are transferred to an XML file that maintains their (semantic) relationships via tags, which are corresponding to classes and properties on the ontology. Example of structure of the XML file is shown in Fig. 40.

Second, information of the XML file is read and passed to the ABox

```
- <htsitems>
  - <Sleepwear id="0" negation="false" name="bathrobes">
    - <Quantity name="70%" negation="false">
        <Material name="silk" negation="false" />
      </Quantity>
      <Material name="textile" negation="false" />
      <Material name="material" negation="false" />
      <Material name="wool" negation="true" />
      <Form name="knitted" negation="false" />
      <Form name="crocheted" negation="false" />
      <Limitation name="more" negation="false" />
    - <HTS_7>
        <Amount value="0.008" />
      </HTS_7>
    </Sleepwear>
  - <Underwear id="0" negation="false" name="dressing gowns">
    - <Quantity name="70%" negation="false">
        <Material name="silk" negation="false" />
      </Quantity>
      <Material name="textile" negation="false" />
      <Material name="material" negation="false" />
      <Material name="wool" negation="true" />
      <Form name="knitted" negation="false" />
      <Form name="crocheted" negation="false" />
      <Limitation name="more" negation="false" />
    - <HTS_7>
        <Amount value="0.008" />
      </HTS_7>
    </Underwear>
  - <Others id="0" negation="false" name="similar articles">
    - <Quantity name="70%" negation="false">
        <Material name="silk" negation="false" />
      </Quantity>
      <Material name="textile" negation="false" />
      <Material name="material" negation="false" />
      <Material name="wool" negation="true" />
      <Form name="knitted" negation="false" />
      <Form name="crocheted" negation="false" />
      <Limitation name="more" negation="false" />
    - <HTS_7>
        <Amount value="0.008" />
      </HTS_7>
    </Others>
```

Figure 40: XML file serves mapping data from relational database to ontology.

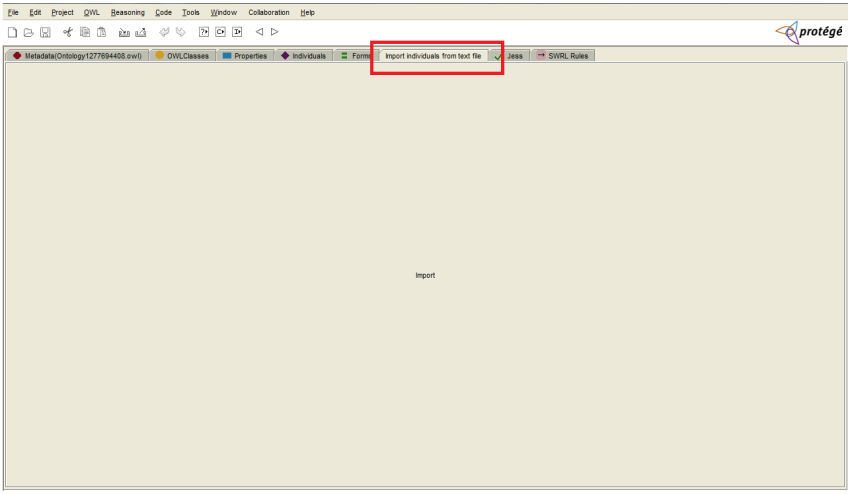


Figure 41: Importing data onto ontology Tab.

by an Importer plug-in. Each data object contained in tags then becomes an instance of the class corresponding to its outermost tags; its axioms are built based on all possible object properties between the outermost tag (*domain class*) as subordinate tags (*range classes*). For instance, the item having $id = 0$, name = *bathrobes* is loaded as an instance of the class *Sleepwear* since its outermost tag is "*Sleepwear*". One of its object properties, i.e., *hasMaterial*(*wool*) is caused by one of its subordinate tags, "$Materialname =$ "*wool*"". The Importer plug-in is added to Protégé, an open-source ontology editor and knowledge acquisition framework, via Protege OWL APIs. In the interface of the Protégé tool, this plug-in is shown as a separated Tab mounted to the main display (see Fig. 41). Data loaded onto ontology is as shown in Fig. 42. It shows information of the original descriptions; in which each description becomes an individual, extracted sub-items become properties and their values.

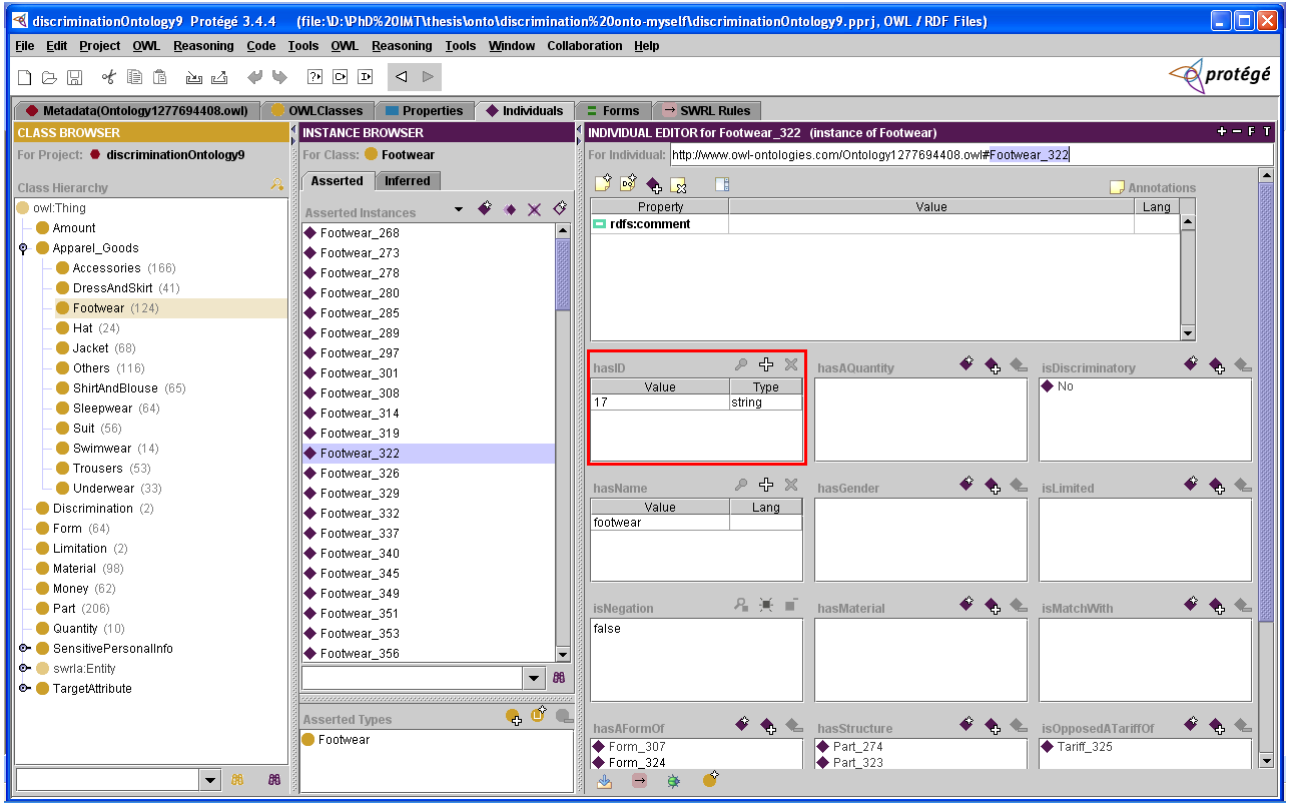


Figure 42: Loading data to ontology for the HTS problem.

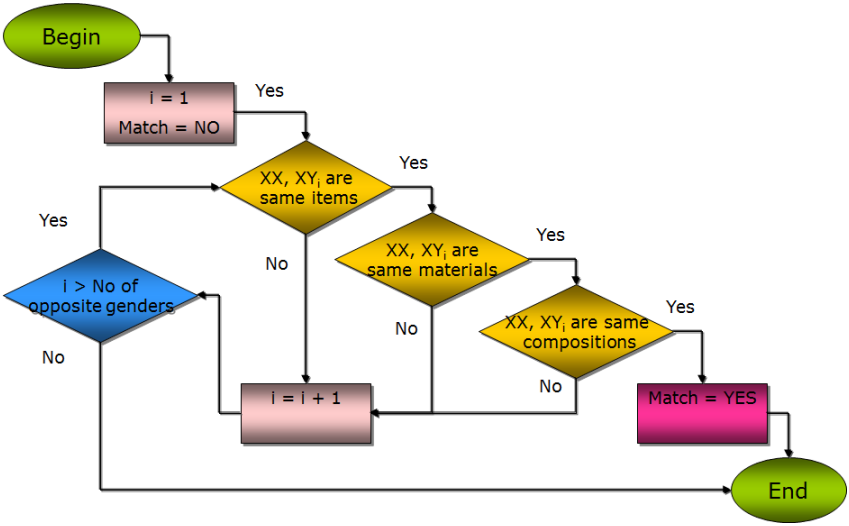


Figure 43: Schema of matching itemsets.

5.2 Matching analysis

The common way to reveal discrimination is establishing a circumstance in which significantly different outcomes between individuals of the protected group and the unprotected group are discovered, e.g., the method of situation testing. Particularly, this process is composed of two steps:

- Seek pairs of items owning a common background - *matching items*
- Check their corresponding decisions. If they are unequal, then a conclusion of discriminatory treatments can be drawn.

The former step - *matching analysis* is obviously the core task. A typical flow for this process is represented as in Fig. 43. It is illustrated by the problem of HTS whose input is the set of well-structured itemsets in relational databases (or instances on ontology), and the output is the set of their matching itemsets.

This phase can be realized on relational data by the *Matching* algorithm based on semantic rules. Then information of matching items are

directly loaded onto ontology as in Fig. 32. Otherwise, matching items can be sought by investigating all possible combinations of each subclass of the Interesting item and its properties. That is the combinations of *Apparel_Goods*, and its axioms such as *hasMaterial*, *hasPart*, its sensitive information like *Gender* and the target attribute such as *isOpposedATariffOf* in advance. Subsequently, these combinations are converted to SWRL rules by using the conjunction operator ($\wedge$) between axioms. This function is represented by a sub-tab in the SWRL Rules Tab of Protégé. The Tab is supported from SWRL API and Jess Rule Engine as shown in Fig. 42. Given an itemset, if itemsets having the same PND properties apart from its PD properties are discovered, then they are considered matching partner of the investigated itemset.

Technically, SWRL rules are stored as OWL individuals with their associated knowledge base when the considered OWL ontology is saved. Classes that describe these individuals are described by the SWRL Ontology. These SWRL rules can be edited by the SWRL Editor, an extension to the Protégé OWL Plugin. Additionally, developers can directly manipulate SWRL rules in an OWL knowledge base by a Java API called the SWRL Factory. This API supports a rule engine plugin mechanism that permits API-level interoperation with existing rule engines via SWRLRuleEngineBridge and SWRLRuleEngine. To connect with Jess engine, the chosen reasoning engine for our system, the RuleEngineFactory is used. Besides, an element of the SWRL Editor, SWRLGUIAdapter, is extended to create our own sub-tab in the SWRL Tab. All rules are instances of swrl:Imp built-in. Instances, classes, and super-classes satisfy any of rules that are marked as items possibly causing discrimination and then used in the discriminatory association mining in the last step.

5.3 Mining over ontology

After the above phase, we have collected matching itemsets (instances) receiving unequal treatments on a common background with different sensitive attributes. The next step is to mine (*discriminatory*) association rules on ontology with support of these matching items. Running min-

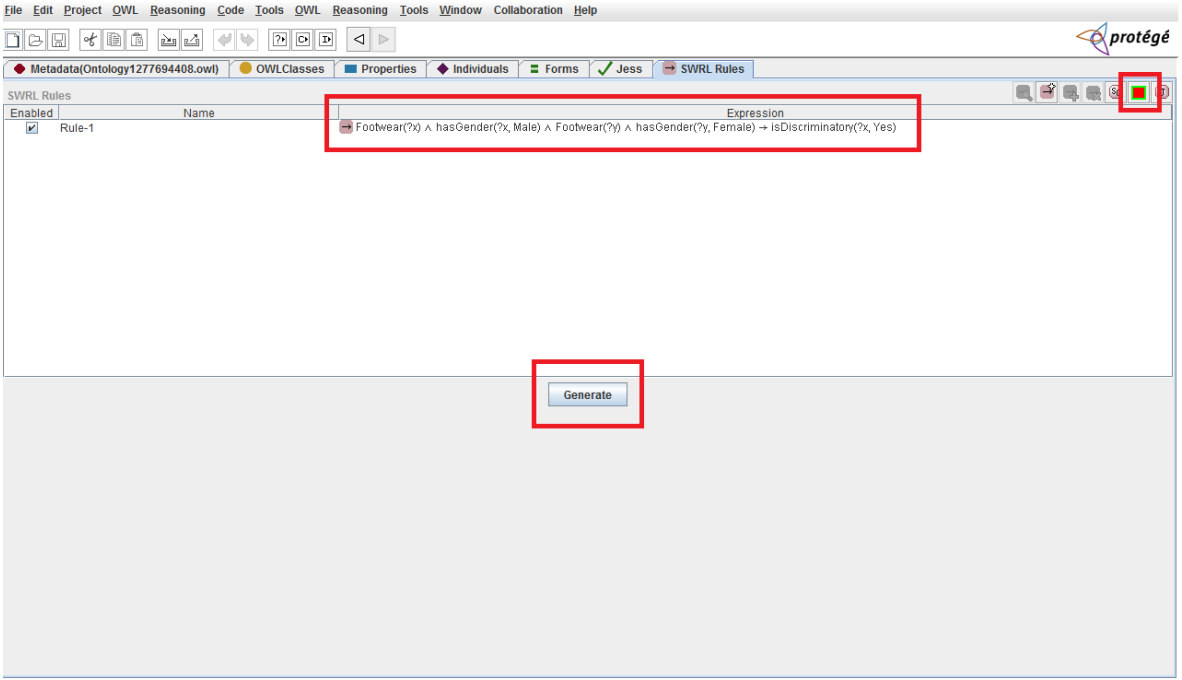


Figure 44: SWRLTab.

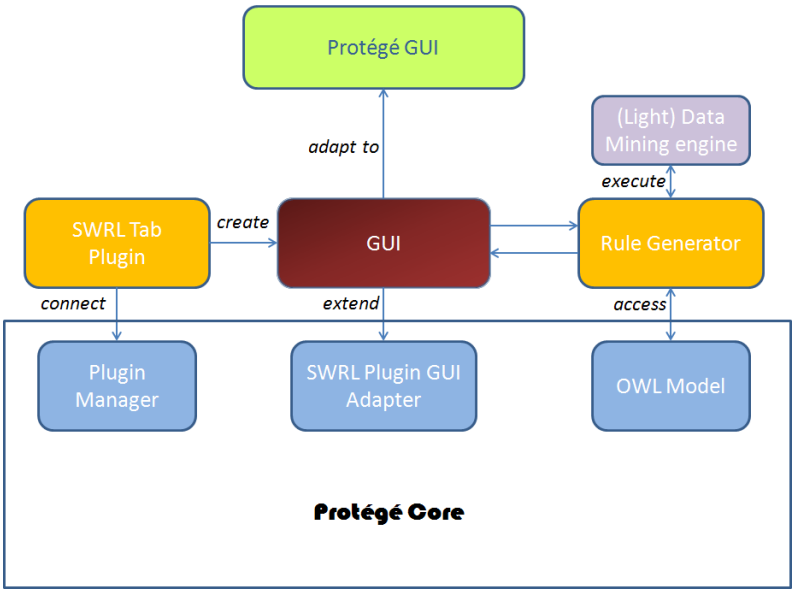


Figure 45: Connecting with Protégé.

ing algorithm on ontology requires support from an ontology editor, in this case, we have still used Protégé. This Java-based framework is extensible, and provides a plug-and-play environment that makes it a flexible base for rapid prototyping and application development.

To be adaptive with Protégé framework, our system must be injected as one of its plugins, working with the Protégé core via the PluginManager as shown in Fig. 45.

The system of mining over ontology is composed of main modules:

- *SWRLTab Plugin*: registers this add-in system with the PluginManager of the Protégé. It is considered as an entry point for our add-in system. Its purpose is to manage all tasks of the system, e.g., the SWRL tab can activate and deactivate it from the Protégé core.

- *RuleGenerator Plugin*: accesses OWLModel for the information of ontology's structure, its instances, axioms to build rules. It trans-

fers this information to a (light) data mining engine after generating rules conforming the PD patterns.

- *Data Mining engine* checks if the generated rules are sufficiently strong, which means to specify if they satisfy the conditions of $minsup, minconf$.

The interface of the system is inherited from the JPanel implementing the SWRLPluginGUIAdapter of the ProtégéGUI. It shows the contexts where discrimination happens in the form of *non-sensitive* rules and which sensitive attributes possibly cause that discrimination in the form of *sensitive* rules as in Fig. 46.

Results of the experiments are presented in the next chapter.

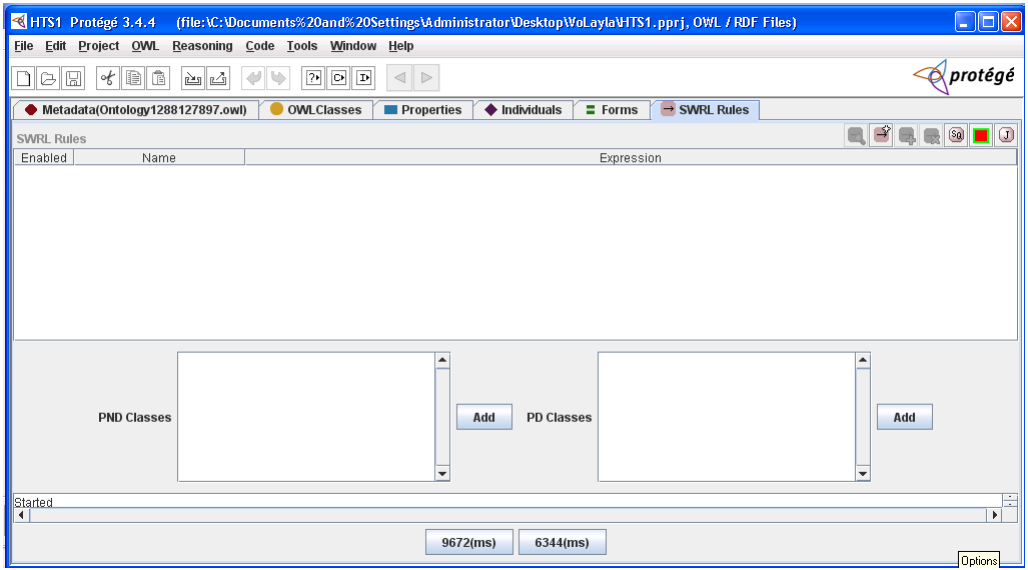


Figure 46: Rule generation on ontology.

Chapter 6

Experiments

To prove its potential and flexibility, the framework has been applied on two different datasets of the above examples:

- HTS: a database of tariff imposed on imported garments.
- German credit data: a database of classifying loan applicants.

6.1 HTS problem

The HTS data provide a typical case of semi-structured data that requires to be pre-processed before applying association mining model. We applied the framework to the mentioned HTS problem for the purposes of:

- Finding possible relationships between attributes of apparel and discriminatory tariff in the form of discriminatory association rules.
- Then, verifying the charge that the discrimination in tariff for certain apparels is based on the gender attribute.

To realize this component, we have built a common sense knowledge base as illustrated in Fig. 47 from the original data as presented in Fig. 29a. Thesauri of synonyms, abbreviations, negative meanings are then built based on the structure of this common sense knowledge.

Figure 47: Common sense knowledge base for the HTS problem.

Through the common sense knowledge base, data are cleaned and standardized as in Fig. 29b. It is then hierarchically structured as in Fig. 29c: from the original description. For each kind of apparel, information is categorized in the name of the apparel, its gender (which group of sex it is produced for), material, form of production, quantities of each material, and quantities for its components (structure) as built in the common sense knowledge base and their corresponding levels. To be able to precisely extract data, syntax analysis by means of rules is deployed. The result of this step is a store of refined and categorized data corresponding to a set of elementary items harvested from the common sense knowledge base. It is easily seen that the tariff attribute is the target attribute, the gender attribute is the PD attribute and all the others form the background (the context). Fig. 48 provides an illustration for the analysis of the HTS data.

Subsequently, the matching step is implemented to seek for matching items by comparing their name, material, form of production, structure, quantity, limitation, and gender information to identify matching items. The correlative tariff of matching items is checked later to set "*activated*" or "*not_activated*" for the discriminatory indicator according to the comparison of taxes imposed on the two matching items. Based on semantic

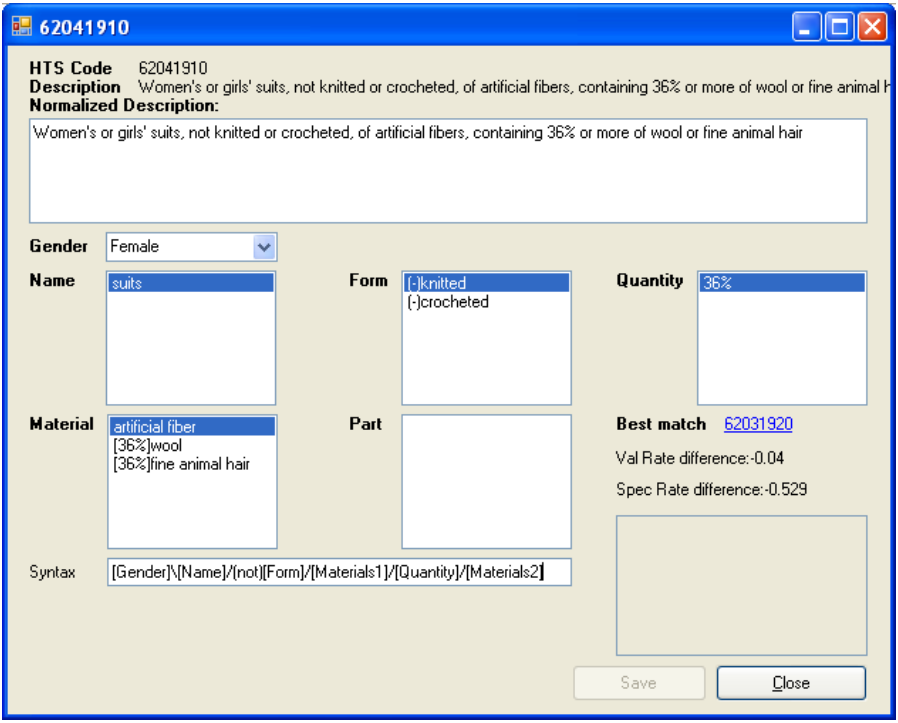


Figure 48: Analyzing HTS data.

rules, matching pairs are found in an easier and more precise way than with the basic approach without support of semantic rules as shown in Table. 2. At the end, a step of association analysis is performed to dig out possibly discriminatory relations among attributes. The Apriori algorithm is applied at varied min_supps over the 427-item database of data which is categorized and has the information of gender.

By applying the association rule mining, and setting different minimum supports, discriminatory rules are found. For instance, when the minimum support is equal to 5 percent, rules are discovered such as:

Table 2: Comparison between semantic and without semantics parsing

Parameters	No semantics analysis	With semantics analysis
Extracted matching pairs	345	290
Exactly extracted pairs	275	289
Accuracy	79.71%	99.66%

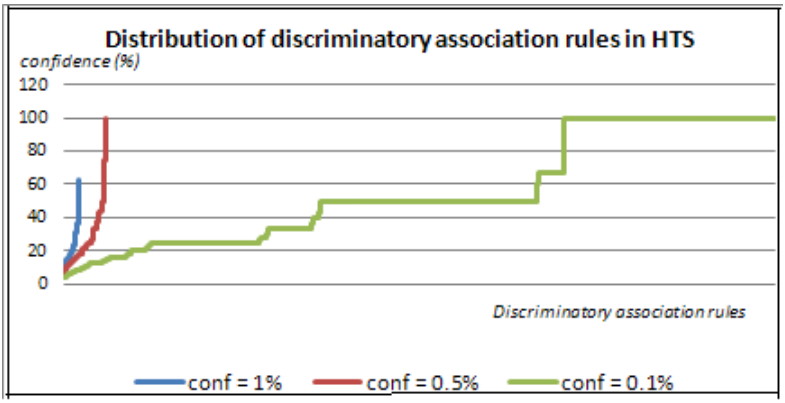


Figure 49: Discriminatory rules for the HTS.

form = *"knitted"*, *"crocheted"*, *material* = *"fine animal hair"*, *"wool"* $\rightarrow$ *discrimination* ($conf = 73.53\%$)

form = *"not_knitted"*, *"not_crocheted"*, *material* = *"fine animal hair"*, *"wool"* $\rightarrow$ *discrimination* ($conf = 50\%$)

The first rule states that for *knitted*, *crocheted* garment products made from *fine animal hair* and *wool*, it will lead to discrimination when the *gender* information is added with the probability of 73.53%.

When reducing the minimum support to 1 percent, the achieved result is more specific, revealing more surprising information about attributes affecting discriminatory tariff. For example:

form = "not_knitted", "not_crocheted", material = "silk", "textile", "not man-made fiber", "not wool", "not cotton", quantities = "70%", MFN_rate_type = 7 → discrimination ($conf = 66.67\%$)

That MFN is the abbreviation for Most Favored Nation, and it implies the normal status in international trade. For a given merchandise, a MFN rate type equal to 7 means that the duty for that merchandise is an ad valorem rate which is a percentage of the value of the merchandise, such as 5 percent ad valorem.

To avoid the misunderstanding that the union of the two item groups "*knitted*", "*crocheted*" and "*not knitted*", "*not crocheted*" are the whole HTS dataset, it is necessary to note that not all items in the dataset are specified as "*knitted*", "*crocheted*" or "*not knitted*", "*not crocheted*", e.g the garment item 65010030:

"*Hat forms, hat bodies and hoods, not blocked to shape or with made brims; plateaux & manchons; all of fur felt, for men or boys.*"

It is obvious that if the minimum support is set relatively high, i.e, greater than 5 percent, it will then affect the number of selected candidates. Indeed, the number of rules decreases as well as the number of items in the antecedent of rules. For instance, when the minimum support is set to 10 percent, the most specific achieved rule is:

MFN_rate_type = 7, form = "not_knitted", "not_crocheted" → discrimination ($conf = 29.58\%$)

This status is shown in Fig. 49 where the number of extracted rules increases when the minimum support decreases. Besides, when the minimum support is sufficiently low, specific rules are exposed, they are actual cases, hence, their confidence is absolute.

In order to examine which gender, male or female, negatively affects the tariff, we also check the difference in tariff shown in measures of abs_lift, ref_lift and rg_lift. In fact, it shows that the ratio of apparels for male which are imposed a lower tariff than for female is often much higher than that one for female as illustrated in Fig. 50 and Fig. 51. In-

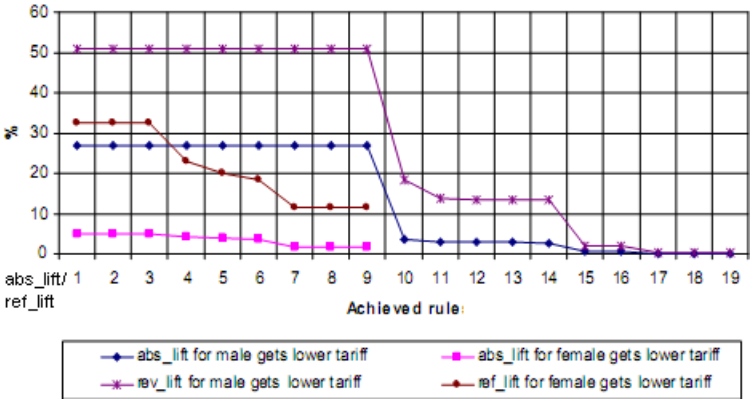


Figure 50: Discrimination in tariff between male and female in HTS shown via measures of abs_lift and rev_lift.

tuitively, for a common kind of apparels, male often receives lower tariff than female, which is shown via the longer lines of abs_lift and rev_lift of male products receiving lower tariff in comparison to the corresponding ones of female products. Moreover, the difference of male/female products rate is not high if female is imposed lower tariff which is opposite to the situation when male products retrieves lower tariff. It is illustrated from the figure that for abs_lift, tax imposed on male products can obtain up 30% lower than that tax imposed on female products, whereas the maximum lower level that female products can reach is only around 5%. The same situation is also observed in rev_lift measure. In general, it can be said that apparels for male are favored in the tariff. This result comparatively fits researches of (MG08), (Mic07a), (Mic07b).

For the purpose of achieving discriminatory rules at multiple levels of abstraction, from the original data, a HTS ontology has been built for Apparel_Goods as shown in Fig. 52. PND elements are classes of attributes of garment items, e.g., *Material, FormOf, Parts*, etc. PD elements are grouped in subclasses in the *SensitivePersonalInfo* class, namely the female gender a garment is produced for. The target attribute is the taxation tariff applied, shown in the property *isAppliedATariffOf*. Measures of support and confidence of a discriminatory classification rule can be re-

Figure 51: Discrimination in tariff between male and female in HTS shown via measure of rg_lift.

spectively interpreted as the proportion of discriminatorily taxed apparels produced for women that are recalled by the rule; and the precision of the discrimination conclusion of the rule given that the antecedent holds.

The rule extraction algorithm is applied to this HTS ontology, adopting a minimum support threshold of 1% and a minimum confidence threshold of 10%. Outcome rules are represented as Horn clauses, using the "conjunction" ($\wedge$) operators by automatically grouping individuals of each subclass of the *Apparel_Goods*, its axioms such as *hasMaterial*, *hasPart*, sensitive information like Gender and the target attribute *isImposedATariffOf*. The immediate advantage of relying on an ontology is that discriminatory rules at different levels of abstraction can be considered. As an example, the extracted rule:

$$Shorts(?x) \wedge hasMaterial(?x, \text{“}fine\ animal\ hair\text{”}) \\ \rightarrow isDiscriminatory(?x, yes)$$

with a confidence $conf = 66.67\%$ can be directly compared with its ancestor rule at the grand-parent level (the concept $Shorts$ is a sub-class of $Outerwear$):

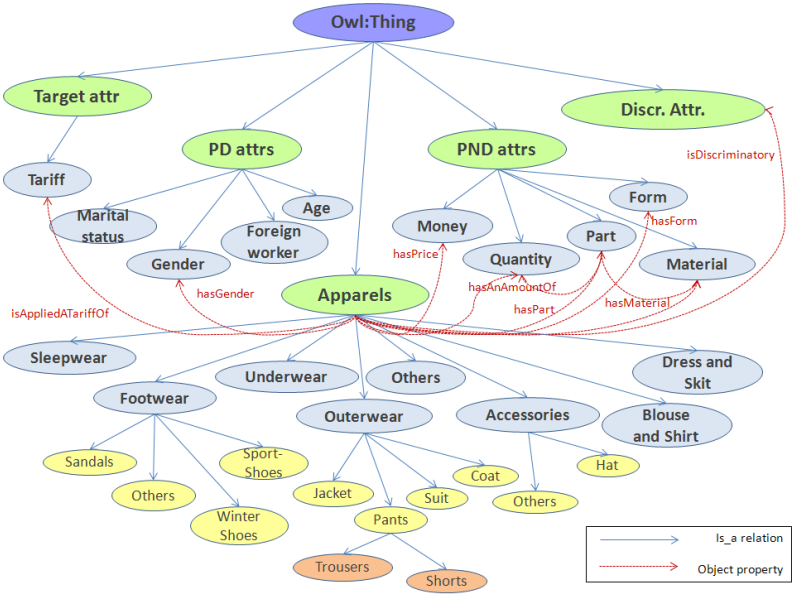


Figure 52: HTS ontology

$$Outerwear(?x) \wedge hasMaterial(?x, \text{“}fine\ animal\ hair\text{”}) \\ \rightarrow isDiscriminatory(?x, yes)$$

which has a lower confidence of $conf = 57.78\%$. Intuitively, this can be read as the fact that tariffs for shorts are more discriminatory than the ones at the level of outerwar. It is stated before that differences between individuals are controlled by means of the discriminatory indicator, a discriminatory rule involving only a subset of $R_1, \ldots, R_t$ can be safely read as the fact that the individuals in such a subset are discriminated (in proportion to the confidence of the rule). Thus, the two rules above concern pairs of similar garments, i.e., shorts/outerwar with the same (PND) object properties, but with different gender property. Hence, they already control for characteristics other than the one explicitly mentioned in the rule, i.e., $hasMaterial$, which now assumes the expected role of *summarizing* under which conditions gender discriminatory tariffs apply.

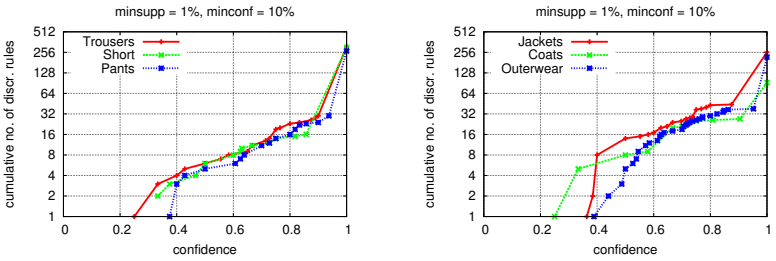


Figure 53: Cumulative distribution of discriminatory classification rules by confidence.

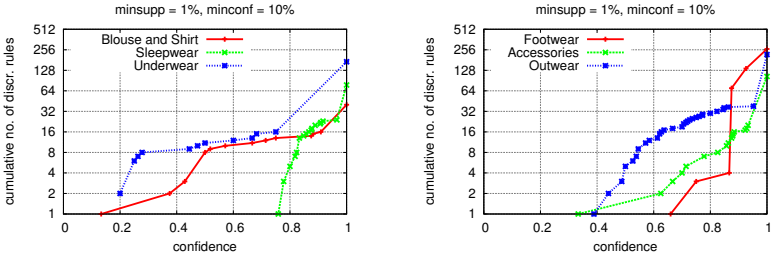


Figure 54: Cumulative distribution of discriminatory classification rules by confidence.

Fig. 53 and Fig. 54 show the cumulative number of discriminatory classification rules extracted for a minimum confidence of 10%. Fig. 53 (left) restricts to trousers and shorts and to their parent concept, namely pants (see the hierarchy in Fig. 52). Most of the rules have a confidence of 100%, namely they summarize contexts of garments with different male/female tariffs. Also, the three distributions are rather similar, which means that the conditions for discriminatory tariffs are not very specific of trousers or shorts or pants, but rather of their properties, such as the materials they are made of. Fig. 53 (right) shows a plot with characteristics similar to the previous one, but now comparing higher level concepts, namely jackets, coats, and their parent concept, i.e., outerwear. Fig. 54 instead reports the distribution of six concepts at the highest level. Some differences are now visible. Sleepwear and footwear appear to

have the steepest distributions, which means that the contexts of discrimination summarized by their classification rules are very precise. In addition, footwear has the highest number of rules, hence exhibiting the largest number of discriminatory contexts[1]. On the contrary, the blouse and shirt concept shows the lowest number of discriminatory contexts. Summarizing, the plots in Fig. 53 and Fig. 54 provide interesting hints to an anti-discrimination analyst on how to prioritize subsequent analyses.

Compared to the rules retrieved by data mining on relational database, in which elements neither join in relationship nor support each other, the new rules are more expressive when elements in the body of rule are semantically related to each other, i.e. a specific material (*fine animal hair*) for a specific group of garment products (*shorts, outerwear, ...*) instead of a apparel products. Besides, based on the subsumption relation (*is-a*) between concepts, rules at different levels of generality are achieved, e.g., from a specific kind of *Trousers* to a general garment product such as *Outwear* category. Hence, they provide a richer semantic for the knowledge of discrimination discovery. Especially, it shows that the discriminatory indicators help to expose much more hidden information of discrimination in the HTS set, which was not revealed before as the number of candidates of frequent itemsets was too small. Also, the number of candidates for the extraction of rules is reduced as well because of the constraints between products and its properties.

It is also worth to investigate discrimination occurring in very specific categories (not transactions) of garments that are opposed to a discriminatory tariff, because when rule mining is performed at high level, highly particular rules cannot be found, thus, we decreased the *minsup* to a very low value, i.e. 0.0001. The details of discrimination for each subcategory shown in Fig. 55 prove that the more specific categories are, the more likely (more often) they appear to be discriminatory, which is further supported by the high confidence (almost 100%) of most of rules. Fig. 55 shows the general trend of these rules, in which for each pair

[1]Since the grain of the dataset is a tariff applied to a garment, this does not imply that footwear exhibits the largest discrimination in terms of male/female tariff *differences*, or in terms of *market-value* of the discriminated garments.

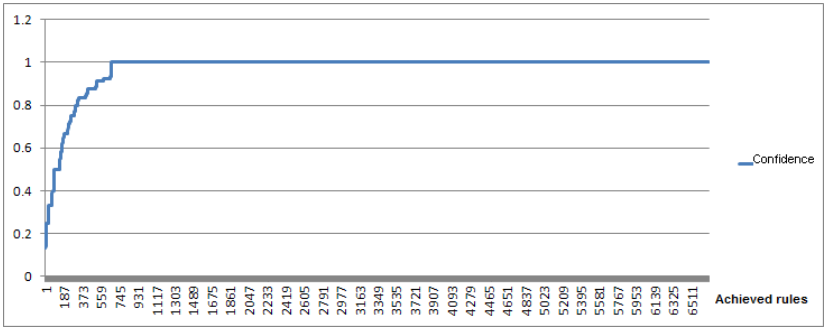


Figure 55: Generalized discriminatory rules in HTS dataset at lowsup.

(X, Y), X is the number of rules with a confidence $\leq Y$.

Finally, we mention that, due to the small size of the HTS dataset, running time and memory occupation of the rule extraction algorithm are negligible on current PC desktop machines. For example, the execution completes in $9891ms$ for $minsup = 1\%$, and $37532ms$ for $minsup = 0.01\%$. In general, the bulk of the resources needed by the algorithm in Fig. 35 is due to the classification rule extraction phase. While it is non-optimal in our SWRL-based implementation, we have already observed that it can performed by optimized external modules (the "Data mining engine" in Fig. 45).

Looking back at the initial targets of our experiment, the following results are achieved:

- Finding relationships between attributes of apparels and discriminatory tariff in the form of discriminatory association rules with a high confidence. Through the achieved results, verify the charge from Totes-Isotoner and others that the discrimination in tariff for certain apparels is considerably affected on the basis of gender.

- Other likely hidden causes of this matter are the MFN rate type equal to 7, "knitted", "crocheted", "not knitted", "not crocheted" for the form of production, "silk", "textile", "wool" with a relatively high confidence and high frequency for different values of

minimum support. This result strongly supports the findings of (PRT08), (PRT09a), and (PRT07) when PND attributes may have certain effect on discrimination. If there is no reasonable explanation, this kind of discrimination should be excluded in business otherwise it may lead to loss in benefit or troubles in law such as anti-dumping duty. Although the experimental results well satisfy the initial requirements, a new question arises: why the above three attributes have a strong connection to the disparity status in the HTS? Is there any historical explanation for it?

- Verifying the potential of the framework. Through the experiment, the proposed framework shows the ability to transform the ambiguous original HTS data to the well categorized data suitable for analysis. By its result, improvement and advances for the business process might be attained.

6.2 German credit data

We also applied the framework on the German credit dataset to verify if the "bad" judgment has a strong relation with sensitive information, i.e. gender, age, foreign worker, or marital status. The possible effects are judged by the measure *e-lift* (PRT08) with $\alpha = 1$ which means if *e-lift*> 1 then the sensitive information c has negative effect on the rate. Thus, it is possible to conclude there is discrimination against the sensitive information c. An ontology is built is in Fig. 56 according to the template of an ontology serving the purpose of discrimination discovery introduced in Chapter 4.

Example of rules extracted from German credit data is as following:

LoanApplicant(?x) $(\wedge)$ *hasEmploymentInfo(?x, ?y)* $(\wedge)$ *Employed(?y)* $\rightarrow$ *getClassification(?x, C2)*, *conf* = 29.96%

LoanApplicant(?x) $(\wedge)$ *hasEmploymentInfo(?x, ?y)* $(\wedge)$ *Skilled(?y)* $\rightarrow$ *getClassification(?x, C2)*, *conf* = 30.46%

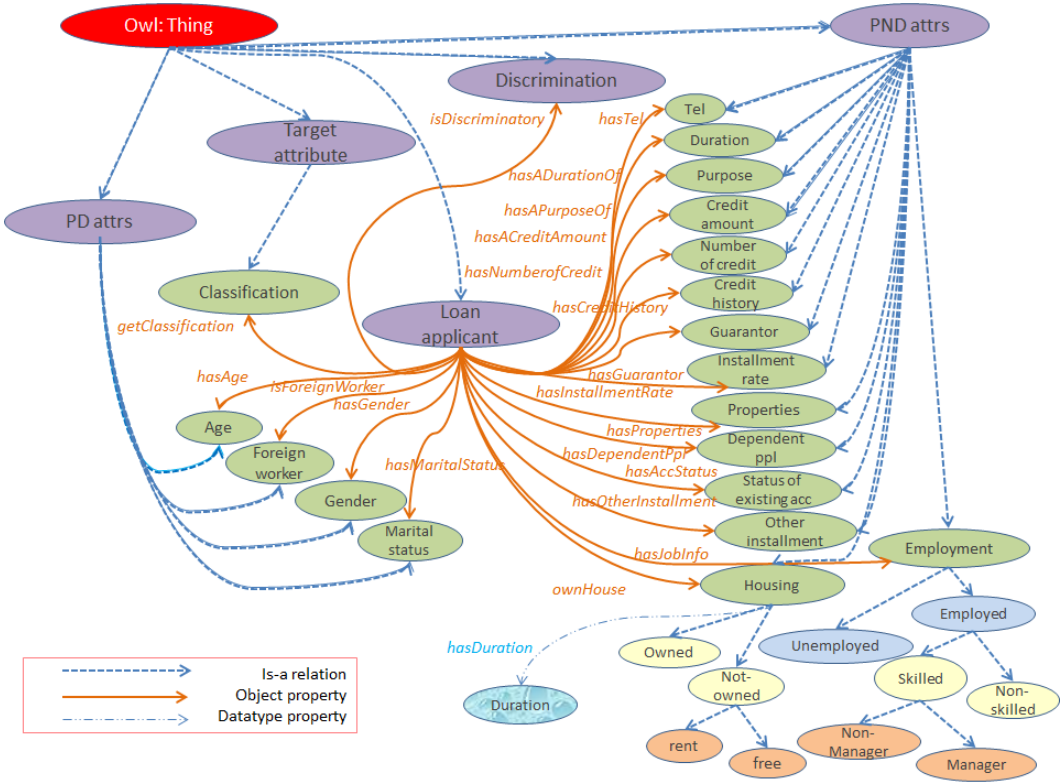


Figure 56: German credit ontology.

This pair of rules shows that varied levels of abstraction can also be discovered at the level of properties not only at the level of concepts as in the HTS problem. Particularly, from the general information (Employed or not) of the Employment status to a specific status of Employment (Skilled job). Through the two experiments, the approach proves its ability to discover discrimination at different levels of generality as well as its flexibility in representing different aspects of generality:
Minsup = 1%, minconf = 20% for base rules
Achieved base rules: 12836 rules

The achieved results show that the classification is significantly affected by this sensitive information, however, the effects are disparate when mining is deployed at varied levels. When randomly collecting rules at more general level, foreign worker is likely the least favorite customers, and followed by young people who do not often have a considerable amount of credit at bank as well as a long credit history. It is almost guaranteed that client is classified "bad" if that one is a foreigner. Young and female applicants are not favoured as well. Whereas rules are collected in a brute force way, results are diversified:

- Gender has negative influence on the rating. Particularly, male applicants are likely more favored than female applicants. It is shown by observing that most of rules with added "male" information have lower confidence than base rules whereas rules with added "female" information make the opposite effect as shown in Fig. 57, and Fig. 58.

- And single men (A93) are even more positively discriminated which can be seen by the observation that a larger number of rules have lower confidence than base rules compared to the case when only "male" information is added as illustrated in Fig. 59. While men who have marital problems (A91) are likely not favored anymore as shown in Fig. 60.

- Surprisingly, being a foreign worker does not affect negatively on the rating system when it is examined at more specific levels, in-

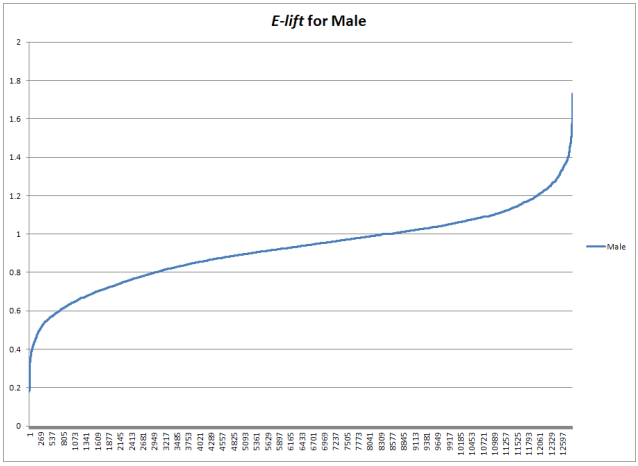


Figure 57: Generalized discriminatory rules w.r.t *Male* info in German credit dataset.

stead of only general information. In general, rules with this attribute have the same confidence as base rules as presented in Fig. 63. This might be explained that foreign workers often have the same living conditions, it is then not necessary to add this information to their background as it is very likely typical for this attribute only.

- Among applicants, young people seem to be disfavored by the rating system which is represented the large number of rules increased their confidence when young age is added as presented in Fig. 64, Fig. 65, Fig. 66.

- The rate of badly treated old persons is not so high, but the difference, but the difference of *bad* rate in comparison to other groups are the largest, up to 5 times, the highest *e-lift* in all experimented cases.

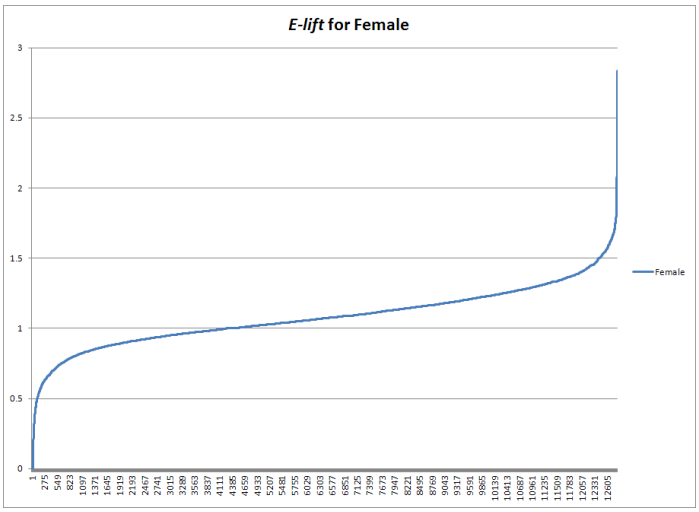


Figure 58: Generalized discriminatory rules w.r.t *Female* info in German credit dataset.

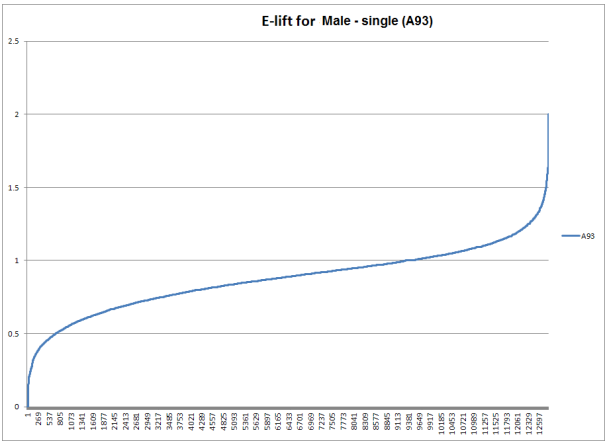


Figure 59: Generalized discriminatory rules w.r.t *Male-single* info in German credit dataset.

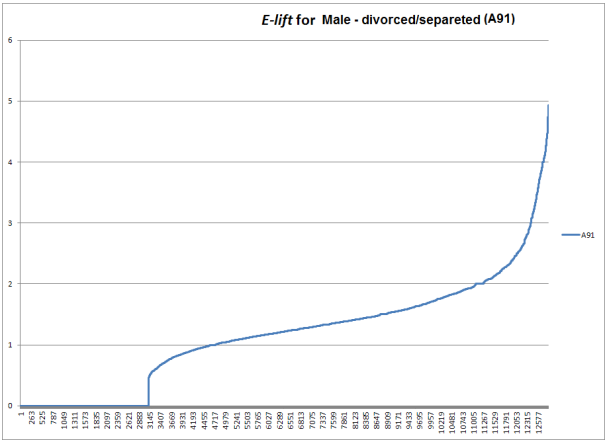


Figure 60: Generalized discriminatory rules w.r.t *Male-troubled in marriage* info in German credit dataset.

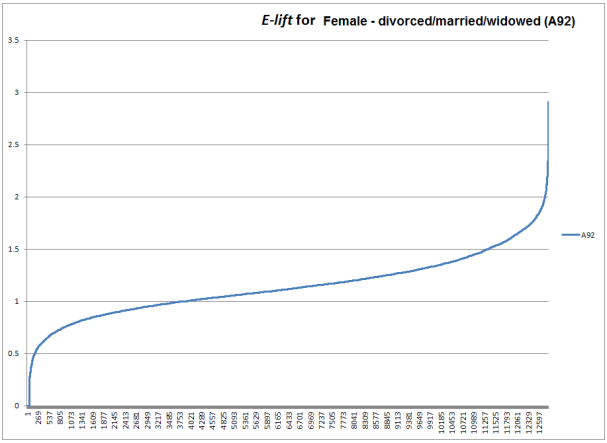


Figure 61: Generalized discriminatory rules w.r.t *Female-divorced/married/widowed* info in German credit dataset.

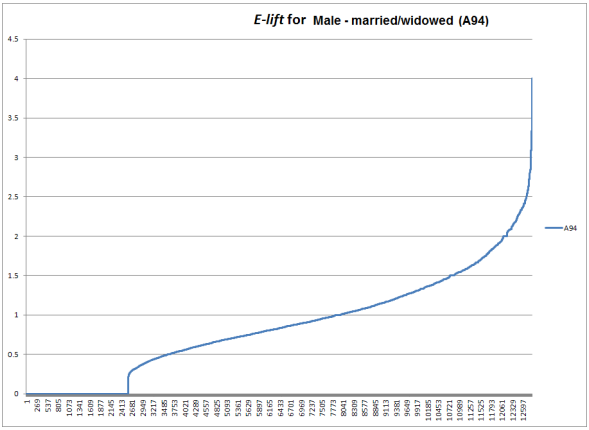


Figure 62: Generalized discriminatory rules w.r.t *Male-married/widowed* info in German credit dataset.

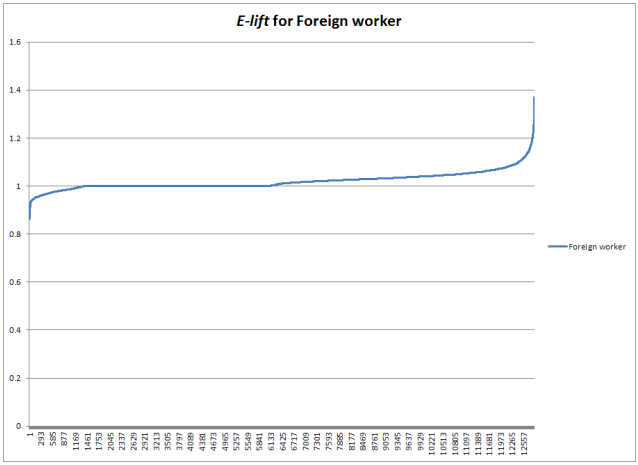


Figure 63: Generalized discriminatory rules w.r.t *Foreign worker* in German credit dataset.

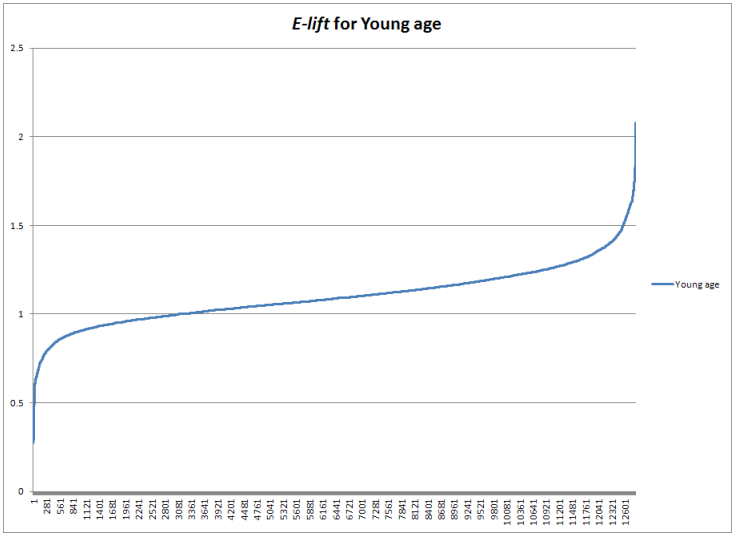


Figure 64: Generalized discriminatory rules w.r.t *Young person* in German credit dataset.

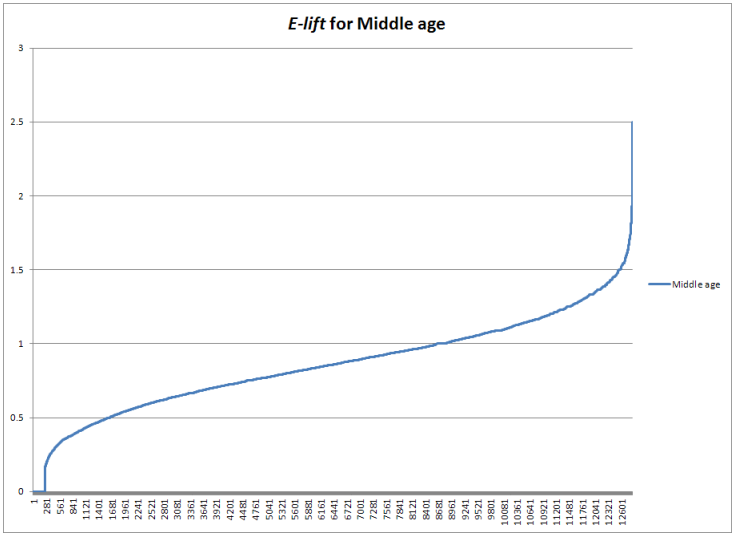


Figure 65: Generalized discriminatory rules w.r.t *Middle aged persons* in German credit dataset.

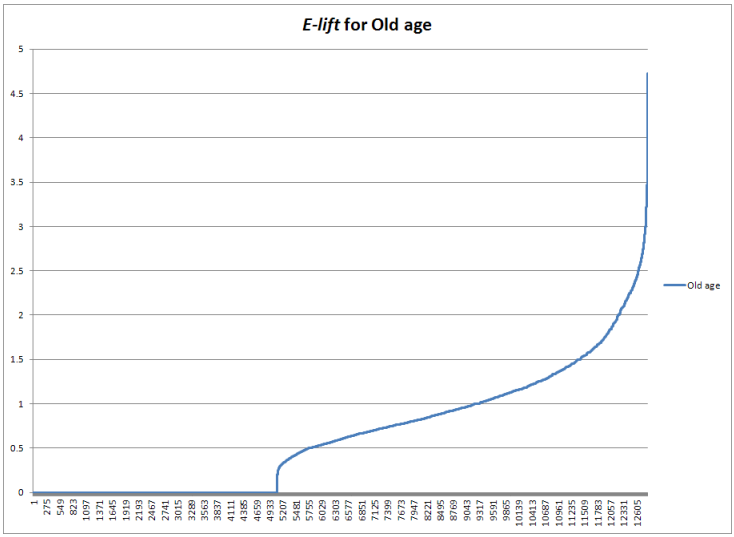


Figure 66: Generalized discriminatory rules w.r.t *Old persons* in German credit dataset.

Chapter 7

kNN as Situation Testing

7.1 Introduction

Generally, the *group under-representation* attaining a benefit is often considered a quantitative measure in studies of discrimination (Ell05), (Ler91). However, it has certain drawbacks in both legal and technical aspects. Consider I a dataset of historical decisions about granting or not a benefit (e.g., a loan, a job, a wage increase). Let p_1 (resp., p_2) be the proportion of people in the protected-by-law group (resp., not in the protected group) that were not granted the benefit. And let p be the proportion of all people (both protected and not) that were not granted the benefit. Group under-representation can be measured due to varied measures for the difference between p_1 and p_2, e.g., $p_1 - p_2$, adopted by law as shown in Fig. 69 (PRT08). The higher these measures are, the more under-represented the protected group suffers.

On the legal side, it can be contested the usage of aggregated measures over *undifferentiated* groups for exploring discrimination against protected-by-law groups. As an example, assume a high value for $p_1 - p_2$ in I, with women as the protected group and job hiring as the benefit. Since p_1 and p_2 are aggregated values, they mix decisions for people who possess different eligibilities for the application. A typical legal argument is that of *genuine occupational requirement*. For instance, if the

job requires a special driving licence, which most of the male applicants have whereas most of the female applicants do not have, then the non-hiring rates p_1 and p_2 cannot be adopted for comparing applicants with different attributes.

On the technical side, the outcome of the mining process is a (possibly large) set of classification rules, which provide local and (possibly) partially overlapping insights of possible discrimination.

Thus, the approach needs to be polished in order to address these limitations. We approximate the method of situation testing by a variant of the k-nearest neighbor (k-NN) classification, which labels each tuple in a dataset as discriminated or not. Direct discrimination discovery then reduces to build a classifier providing a global description of the conditions where discrimination occurs. Particularly, for each member of the protected-by-law group with a negative decision outcome (an individual who might claim to be a victim of discrimination) we seek testers with similar, legally admissible characteristics, apart from being or not in the protected group - k-nearest neighbors. If considerably varied treatments between the testers of the protected group and the ones of the unprotected group are recorded, it can be ascribed the negative treatment to a bias against the protected group. Similarity is modelled via a distance function as introduced in Section. 2.2.

7.2 Basic definitions

k-nearest neighbor (k-NN) is a lazy instance-based classification method. The classification model simply consists of storing the training set. Given a tuple r to be classified, k-NN: (1) seeks the k tuples in the training set closest to r w.r.t. a distance measure $d()$, i.e., its k-nearest neighbors; (2) then assigns as class value to r the most frequent class value among its k-nearest neighbors.

Let r^i be a tuple r whose id is i. A collection R is a set of tuples with id containing protected-by-law attributes. For a tuple r, we assign to every $r^i \in R$ a rank (as a neighbor of r) on the basis of its distance from r (or,

for equal distances, on the tuple id). Formally, it is defined:

$$rank_R(r, r^i) = |\{j | d(r, r^j) < d(r, r^i) \vee (d(r, r^j) = d(r, r^i) \wedge j \leq i)\}| \quad (7.1)$$

The k-NN set for a given tuple r is the set of top k tuples w.r.t. ranking.

Definition 3 *For a dataset R, we define:*

$$\begin{cases} kset_R(r, k) = \{r^i \in R | rank_R(r, r^i) \leq k\} \\ kset_R(r, k, d) = \{r^i \in R | rank_R(r, r^i) \leq k \wedge d(r, r^i) \leq d\} \end{cases} \quad (7.2)$$

7.2.1 Protected-by-law groups

It is set *protected*(r) iff r contains a vulnerable value, e.g., r[sex] = female, or r[race] = black. Using this *protected* relation, it is possible to separate tuples of people in the protected group from those of people not in the protected group - *unprotected group*.

Definition 4 *We define:*

$$P(R) = \{r^i \in R | protected(r^i)\}$$
$$U(R) = \{r^i \in R | \neg protected(r^i)\}$$

Notice that R = P(R) $\cup$ U(R) does not necessarily hold since tuples with unknown values to be tested by $protected$ are not included in P(R) nor in U(R).

7.2.2 Legally-grounded attributes for distance measurement

In k-NN classification, distance is used to seek neighbors of a given tuple. As in situation testing, such neighbors are searched with reference to attributes that are legally-grounded for being adopted in making the decision. Therefore, it can be assumed that $d()$ is defined on a subset of the attributes of the dataset, e.g., those that are legally admissible in a discrimination litigation. Additional attributes may be present in the dataset for other purposes, for extracting a description of cases where

discrimination happens. Let $G \subseteq \{1, ..., n\}$ be the set of attribute indices (or, equivalently, attribute names) that are legally-grounded. We write $\pi_G(r)$ to denote the projection of tuple r over attribute indices in G, e.g., $\pi_{\{1,3\}}(\langle 3, 5, 4\rangle) = \langle 3, 4\rangle$. Hereafter $d(r, s)$ are referred to by $d(\pi_G(r), \pi_G(s))$.
Distance is computed with reference to attributes indexes in G. In symbols, when writing $d(r, s)$ we actually restrict to consider $d(\pi_G(r), \pi_G(s))$.

In particular, when writing $kset_R(r, k)$ we now intend the k-NN set w.r.t. the distance over attribute indexes in G.

7.2.3 Target attribute

As previously stated in Sec.4.2, it is included in the dataset I a target attribute τ for differentiating tuples or a historical treatment. We write $dec(r)$ to denote the value of τ for r. We admit nominal, interval-scaled and ordinal target attributes. For nominal (typically binary, i.e., two-valued) target attributes, we denote by $\ominus$ the negative decision (deny of benefit), and by $\oplus$ the positive decision (grant of benefit). Examples of interval-scaled target attributes include wage in a dataset of personnel, and interest rate in a dataset of loans. Ordinal target attribute are basically treated as interval-scaled attributes once ordinal values are mapped into interval-scaled ones using the standard mapping recalled in Section. 2.2.

7.3 kNN as Situation Testing

The purpose of situation testing is to prove the existence of discrimination. Indeed, given records of decisions taken in a context, for each member of the protected-by-law group with a negative decision outcome (an individual who may claim to be a victim of discrimination); individuals with similar, legally plausible characteristics, apart from being or not in the protected group are under scrutiny. If significantly different decision outcomes between individuals of the protected group and the unprotected group can be recorded, it is possible to state the negative decision

to a bias against the protected group. Let us formalize this practice in kNN.

Let K_1 be the set of k-nearest neighbors of r in $P(R)\backslash\{r\}$ and K_2 be the set of k-nearest neighbors of r in $U(R)$. K_1 and K_2 represent persons whose attributes are close to the ones of r apart for being in the protected group $P(R)$ or in the unprotected group $U(R)$ respectively. The distance function is defined on attributes that are legally-grounded, i.e., legally admissible for being adopted in taking the decision of granting or not a benefit. For a nominal target attribute, the probability of the decision outcome for r can be estimated from K_1 (resp., K_2) as the proportion p_1 (resp., p_2) of tuples whose decision values are the same of r. The difference between the observed values p_1 and p_2 represents the bias of the decision against r due to membership to the protected group. Any discrimination measures proposed in (PRT08) can be used to estimate the discrimination.

Definition 5 *For $r \in P(R)$, $diff(r) = p_1 - p_2$, where:*

$$\begin{aligned} p_1 &= |\{r' \in kset_{P(R)\backslash\{r\}}(r,k) | dec(r') = dec(r)\}|/k \\ p_2 &= |\{r' \in kset_{U(R)}(r,k) | dec(r') = dec(r)\}|/k \end{aligned} \tag{7.3}$$

Assume that a negative decision is assigned to r, namely $dec(r) = \ominus$. A value $diff(r) = t \geq 0$ means that the decision is more frequent in the neighbors of the protected group P(R)with respect to the neighbors of the unprotected group U(R) by a percentage difference of t. This implies that the negative decision for r is not explainable on the basis of the legally-grounded attributes used for distance measurement, but rather it is biased by group membership. Whether the bias was intentional or not is irrelevant: laws also sanction *unintentional discrimination*. t is a measure of the strength of the discrimination bias. A value $t \leq 0$ indicates that the negative decision for r is not explainable by a worse treatment of the neighbors in the protected group. Hence, no discrimination conclusion can be drawn.

Conversely, assume a positive decision $dec(r) = \oplus$. By reasoning as above, $diff(r) \geq 0$ means a bias towards the positive decision due to group membership. This could be the result of *affirmative actions*, also

called *positive actions* or *reverse discrimination*, consisting in a range of policies or quotas to overcome and to compensate for past and present discrimination. Dually, $diff(r) < 0$ means that the positive decision is not explainable by a general better treatment of the neighbors in the protected group.

It is also valuable to study how the observed value $p_1 - p_2$ is affected by randomness in decisions of the dataset. A confidence interval provides us with a range for the true value over the entire dataset of decisions at a certain level of significance. Let us denote by π_1 and π_2 the true proportions over the protected and the unprotected neighbors. The Wald confidence interval for $\pi_1 - \pi_2$ at $100(1-\alpha)\%$ level of significance is $[(p_1 - p_2) - d, (p_1 - p_2) + d]$, where:

$$d_\alpha = Z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{k} + \frac{p_2(1-p_2)}{k}}. \tag{7.4}$$

The well explanation of this result can be found in (Agr02), (FLP03). Also,(New98)compares alternative methods for interval estimation. We mention here that when k is very low, the Wald interval becomes imprecise. Exact methods have been proposed in the statistics literature, where "exact" means that the actual discrete distribution of the statistical parameter is adopted in computing the confidence intervals, instead of approximating it with a normal distribution as done in Wald intervals. In our context, we extend the $diff()$ function as follows:

$$diff(r,\alpha) = \begin{cases} max\{0, p_1 - p_2 - d_\alpha\} & \text{if } p_1 - p_2 \geq 0; \\ min\{0, p_1 - p_2 + d_\alpha\} & \text{otherwise} \end{cases} \tag{7.5}$$

Intuitively, non-negative values of $diff()$ are corrected to the lower bound of the confidence interval, and negative values are corrected to its upper bound. $diff(r,\alpha) = 0$ when the null hypothesis $\pi_1 - \pi_2 = 0$ cannot be rejected.

The definition of $diff()$ extends to interval-scaled target attributes by setting:

$$\begin{aligned} p_1 &= |\{r' \in kset_{P(R)\setminus\{r\}}(r,k) | dec(r') = dec(r)\}|/k \\ p_2 &= |\{r' \in kset_{U(R)}(r,k) | dec(r') = dec(r)\}|/k \end{aligned} \tag{7.6}$$

Assume that higher decision values mean more negative outcomes, e.g., as when the target attribute is the interest rate to be paid for a loan or a mortgage. Intuitively, p_1(resp., p_2) measures the proportion of the neighbors in the protected (resp., unprotected) group that have a higher decision value. A difference $diff(r) = p_1 - p_2 \geq 0$ means a negative bias due to membership to the protected group. Finally, assume that lower decision values mean more negative outcomes, e.g., as when the target attribute is the percentage of salary increase. In order to have that positive values of $diff(r) = p_1 - p_2$, denoting bias against the protected group, the comparison predicates in Eq.7.6 must be changed from $\geq$ to $\leq$.

7.4 Discrimination discovery

Definition 6 *Let $t \in [0, 1]$ be a threshold value. It is said that r is t-discriminated, and write $disc(t, r)$, if $dec(r) = \ominus$ and $protected(r)$ and $diff(r) \geq t$.*

The first condition $dec(r) = \ominus$, warrants distinguish discrimination from favoritism, such as the one resulting from affirmative actions. In fact, if $dec(r) = \oplus$ and $diff(r) \geq t$, then r and its protected neighbors were granted a benefit in a higher proportion than its unprotected neighbors. Also, we require that protected(r) holds because we are interested in characterizing under which conditions a member of the protected group was discriminated or not. Obviously, also members of the unprotected group might be discriminated (again, or instance, because of affirmative actions), but labeling both protected and unprotected group members would mix different causes of discrimination. The value of t is selected due to law. For instance, the U.K. legislation (U.K10) for sex discrimination, sets $t = 0.05$, namely a 5% difference. The U.S. legislation (U.S10) for employment discrimination, sets a threshold (known s the "four-fifths rule") of 1.25 for the measure of selection lift *slift*(). After calculating the above parameters, a global description of the discriminated tuples is provided by solving the classification problem.

Definition 7 *The t-labeled version of a dataset I is the dataset obtained: (1) by*

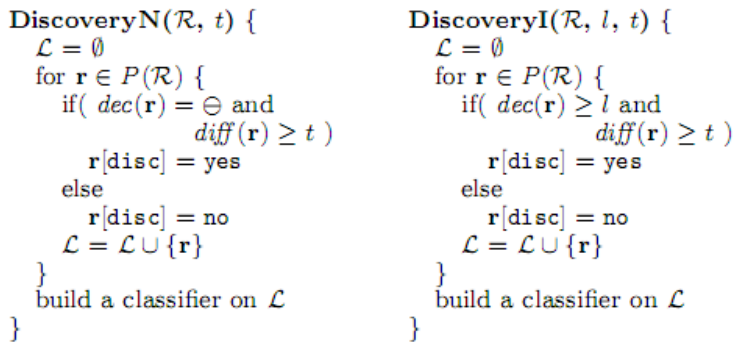


Figure 67: Procedures for discrimination discovery

adding a binary attribute, called disc, *assuming, for a tuple $r \in R$ the boolean value $disc(t, r)$; and (2) by restricting to tuples of the protected group only.*

Discrimination discovery reduces now to the extraction an accurate classifier over a labeled version of I with *disc* the class attribute. Accuracy will be evaluated with standard measures, e.g., precision and recall over the class value and $disc = yes$. The intended use of the classifier is descriptive, namely to provide the analyst with a description of the conditions under which discrimination occurred. Classification models that can be interpreted clearly and easily should be adopted, such as decision trees and classification rules. The overall procedure for discrimination discovery over nominal target attributes is shown in Fig. 67 as *DiscoveryN*() where the t-labeled version of I is computed in L.

Condition (2) in the above definition follows from the fact that for tuples not in the protected group $disc(t, r)$ is always false, hence, any classiffication model would derive assertions such as "if not $protected(r)$ then disc = no", assuming that the model is able to express the condition defining $protected$.

Finally, the overall procedure for discrimination discovery over interval-scaled target attributes is shown in Fig. 67 as DiscoveryI(). A further parameter is now required, namely the threshold l such that $dec(r) \geq l$ denotes the negative decision outcome. Notice that the class attribute

$disc$ in the labeled dataset L is still nominal, binary valued.

7.5 Discrimination prevention

Starting from a dataset of historical decision records, classification models are typically extracted with the intent to learn and predict the decision $dec(r)$ (the class attribute) starting from the other attributes of a tuple r. Preventing discrimination when training a classifier consists of balancing these two contrasting objectives: maximize accuracy of the extracted classification model; and minimize the number of predictions that are discriminatory. Within our framework, a prediction is discriminatory if the classified tuple is t-discriminatory, for a certain fixed threshold t. Why a classification model should be discriminatory? The main reason is due to statistical discrimination, namely the fact that the classification algorithm may recognize patterns in the data where, either directly or indirectly, the membership to a protected group is a proxy for lower performances (e.g., capacity to pay the loan installments). We now propose a simple pre-processing step, orthogonal to the classification algorithm, for tackling the discrimination prevention problem. The method consists of changing the decision value for tuples in the training set that are t-discriminated.

Definition 8 *The t-corrected version of a training set T is the dataset obtained by changing $dec(r)$ from $\ominus$ to $\oplus$ if $disc(t, r)$ holds.*

Given a protected group and a threshold t, in order to evaluate the effectiveness of the pre-processing method, we build two classifiers: one on the training set T , and the other on its t-corrected version T'. Both classifiers are evaluated on a test set V with respect to two measures: accuracy, and t-discrimination. Accuracy is measured as the percentage of correct predictions. t-discrimination is measured as follows. Consider the dataset V where the target attribute is set to the prediction of the classifier. t-discrimination is the percentage of tuples r with $diff(r) \geq t$ among the tuples n the protected group with negative decision. Graphically, we look at the value of the cumulative distribution of $diff()$ (such

```
PreventionN(T, V, t) {
  T' = ∅
  for r ∈ T {
    r' = r
    if( dec(r) = ⊖ and
        protected(r) and
          diff(r) ≥ t )
      r'[dec] = ⊕
    T' = T' ∪ {r'}
  }
  build classifiers on T and T'
  compare them on V
}
```

```
PreventionI(T, V, l, t) {
  T' = ∅
  for r ∈ T {
    r' = r
    if( dec(r) ≥ l and
        protected(r) and
          diff(r) ≥ t )
      r'[dec] = l − ε
    T' = T' ∪ {r'}
  }
  build classifiers on T and T'
  compare them on V
}
```

Figure 68: Procedures for discrimination prevention

as those in Fig. 70, Fig. 71) for the x-axis equals to t. The overall procedure is shown in Fig. 68 as PreventionN().

Finally, the overall procedure for discrimination prevention over interval-scaled target attributes is shown in Fig. 68 as PreventionI(). As for discrimination discovery, a threshold l, such that $dec(r) \geq l$ denotes the negative decision outcome, is required as a further input. Notice that the change of the decision for t-discriminated tuples is now implemented by setting the decision value to a minimum $l - \epsilon$ that keeps the tuple below the limit of negative decisions.

7.6 First experiments

7.6.1 Experiment of kNN as Situation testing

Again, we use the German credit dataset in the experiment. Attributes classified by their types are the following:

- *Interval-scaled*: duration, credit amount, installment_commitment, age, existing credits, num_dependents.
- *Nominal*: credit history, purpose, personal status, other parties, residence since, property magnitude, housing, job, other payment plans,

group	benefit: not granted	benefit: granted	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

$$p_1 = a/n_1 \qquad p_2 = c/n_2 \qquad p = m_1/n$$

$$diff(c) = p_1 - p_2 \qquad slift(c) = \frac{p_1}{p_2}$$

$$elift(c) = \frac{p_1}{p} \qquad olift(c) = \frac{p_1(1-p_2)}{p_2(1-p_1)} = \frac{ad}{cb}$$

Figure 69: Discrimination measures

own telephone, foreign worker, class.

- *Ordinal*: checking status, savings status, employment.

We consider two versions of the dataset: *credit*, the original dataset, with interval-scaled attributes; and *credit-d*, where interval-scaled attributes are discretized into 5 bins of equal width, and the resulting attributes are assigned the ordinal type. In addition to the German credit, the larger datasets shown in Fig. 69 will be considered. Strictly speaking, they are not concerned with decisions, but rather with census information related to people income (*adult, census-income*) or to communities and crimes (*crimes*). Therefore, the objective of discrimination analysis on these datasets is to discover or prevent forms of statistical discrimination. Notice that the target attribute is binary for all of them, apart for *crimes* where it is interval-scaled being the "total number of violent crimes per 100K population".

Consider the *credit* dataset, we fix the protected group to *non-single women* by setting $protected(r)$ to $r[personal\_status] = female, div/sep/mar$. Also, we set the target attribute to class, namely to the credit classification, with $\ominus = bad$ and $\oplus = good$. All the remaining attributes are included to use in distance metric calculation. Fig. 70(a) shows the cumulative distribution of $diff()$ (a measure in (PRT08)) for people in the protected group that have received the bad credit classification, namely

for tuples r such that $protected(r)$ and $r[class] = bad$. The plot shows the distributions for different values of k.

It is immediate to see that the distributions are shifted towards the positive half of the spectrum. More than 60% of the people have $diff(r) \geq 0.1$, which means that bad debtors are at least 10% more frequent among the k-most similar persons in the protected group than among the k-most similar persons in the unprotected group. Hence, the decision of classifying r as bad-debtor appears to be biased by her membership to the protected group.

The discrimination scenario becomes clearer if we consider the same distributions for people in the protected group having received the good-debtor decision, namely for tuples r such that $protected(r)$ and $r[class] = good$. They are shown n Fig. 70(b). Now the distributions are shifted towards the negative half of the spectrum. For $k = 16$, it turns out that $diff(r) \geq 0.1$ in less than 7% of cases, which means that only a small percentage of the good-debtor classification could be attributed to a bias in favor of non-single women.

Finally, observe that the distributions in Fig. 70(a,b) tend to shrink as k increases. This is intuitive, since for $k \to \infty$ the k-NN sets of protected and unprotected groups tend to include all the elements of the group, and then $diffr) \to \hat{p_1} - \hat{p_2}$ where $\hat{p_1}$ (resp., $\hat{p_1}$) is the proportion of decision $dec(r)$ in the whole protected (resp., unprotected) group.

It is also worth studying how the distributions vary for different parameters in the previous case. Fig. 70(c) shows the cumulative distribution of $diff()$ for protected groups defined on ranges of age. The plot clearly shows that youngsters suffer from a higher bias towards the bad-debtor classification than middle-age or older people.

Fig. 70(d) shows how the distribution varies with the set G of attributes used in distance measurement. The plot refers to three sets. G_1 includes all attributes except *sex*, used to define the protected group, and *class*, used as the target attribute. G_1 is the set used so far. G_2 includes attributes related to credit (history, purpose, amount, existing) and to properties (savings, property, housing, third parties) but nothing about job. Finally, G_3 includes only attributes related to credit. From the plot,

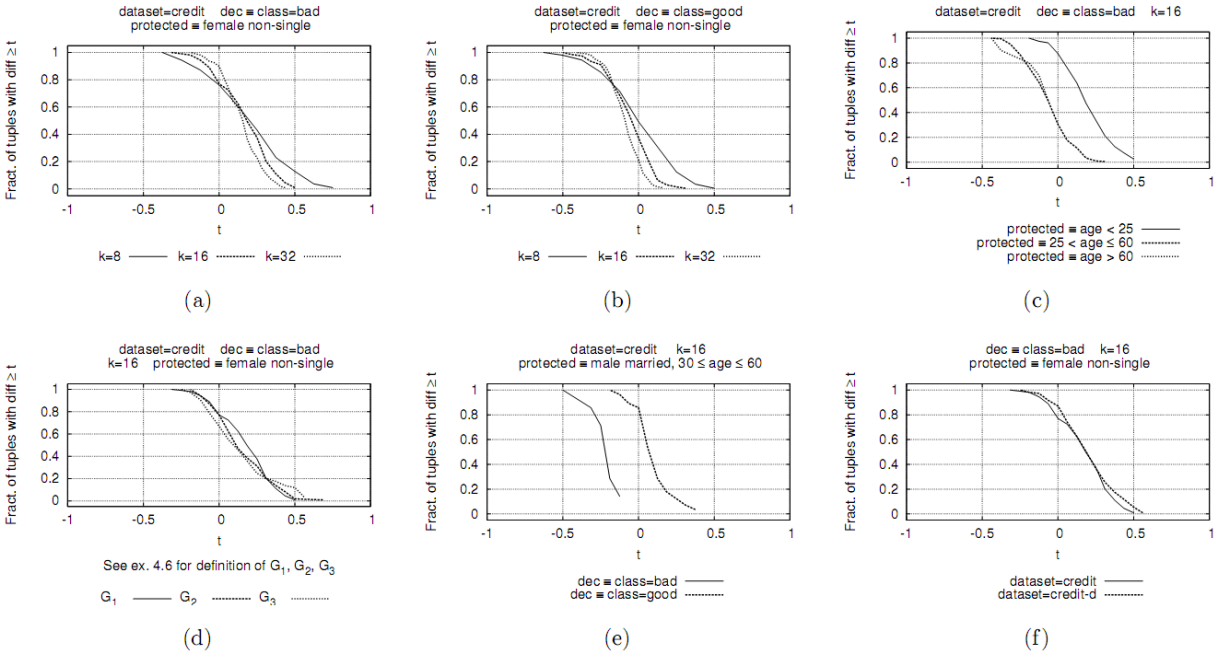


Figure 70: Cumulative distributions of $diff()$ for datasets *credit* and *credit-d*.

we can conclude that the distributions are not dramatically sensible to the set of attributes, although by restricting the set of attributes, high values of $diff()$ tend to be even higher.

Fig. 70(e) highlights the dual concept of *favoritism*, namely discrimination in favor of a group, where the "$protected$" group under analysis here consists of married men in the age range between 30 and 60. Compared to Fig. 70(a,b), the distributions for the bad and the good decision are swapped, in the sense that there is no bias against the group in bad-debtor classification decisions, and there is bias in favor of the group members in good-credit classification decisions.

In the end, Fig. 70(f) shows the distribution of $diff()$ for the *credit-d* dataset, which is obtained by discretizing interval-scaled attributes in *credit*. Protected group, target attribute and attributes in G are kept the same as in Fig. 70(a). Technically, discretization affects the distance function, possibly resulting in different k-NN sets. The plot, however, closely resembles the distribution of the original dataset, with a slight shift towards higher values.

In the $adult$ dataset the target attribute income is not properly a "decision" but a range of income (lower or equal than \$50K , and higher than \$50K). Here, the objective of the discrimination analysis is to discover whether people in a protected group suffer from low income not for their own specific characteristics but rather due to membership to the group - a form of the so-called *social* discrimination. Fig. 71(a,b,c) show the distributions of $diff()$ for non-white people with low income at the variation of three parameters of the analysis: (a) the number k of nearest neighbors; (b) the additional constraint $maxd$ on maximum allowable distance of the nearest neighbors; and (c) the level of significance in statistical validation. As one would expect, the distributions tend to shrink as k increases (already observed in the above example) and as the confidence level increases. Fig. 71(c) shows that very few cases could be brought in a court of law with the support of strong statistical arguments. Finally, putting a maximum distance threshold when looking for neighbors results in a longer right tail distribution, as shown in Fig. 71(b). The maximum distance parameter affects those tuples r that are somehow

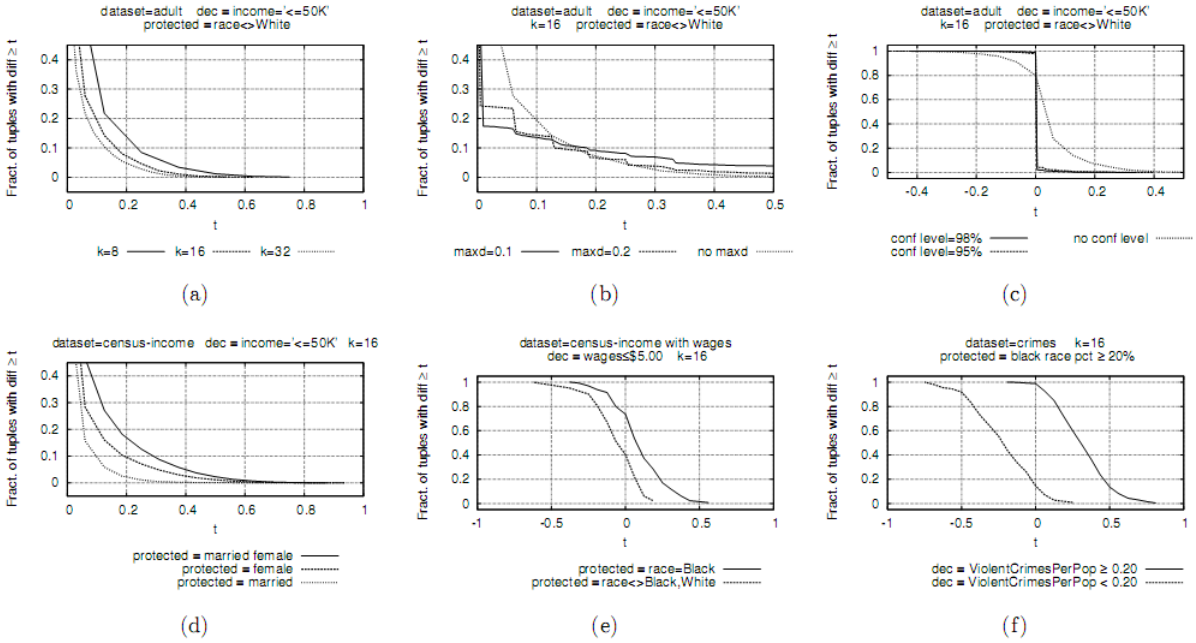


Figure 71: Cumulative distributions of $diff()$ for datasets *adult*, *census-income* and *crimes*.

isolated (ether because of their own attributes, or by effect of the *curse of dimensionality* on the distance function) by excluding "distant neighbors" from their k-NN set.

The dataset *census-income* also contains census data, with the same target attribute as *adult*. However, it is larger both in the number of tuples and in the number of attributes. In Fig. 71(d) we highlight the analysis of *multiple discrimination* (ENA07), namely discrimination on the grounds of two or more factors. The plot shows the distributions of $diff()$ for the low income class of three protected groups: women, married people, and married women. As it can be readily observed, society leads married women with low income to experience a higher difference from the unprotected group (not-married or men) than the difference experienced as women (w.r.t. men) or as married people (w.r.t. not married ones) alone.

Besides, the *census-income* dataset includes the attribute "*wages*" recording the wage per hour. We have selected the tuples with known values of "*wages*". For the resulting dataset, we can set "*wages*" as the (interval-scaled) target attribute and study the distribution of $diff()$. As for nominal target attributes, we concentrate on negative decisions. For interval-scaled attributes, the analogous of $dec(r) = \ominus$ is, for a certain threshold value t, $dec(r) \geq t$ if higher decision values mean more negative outcomes, and $dec(r) \leq t$ otherwise. Fig. 71(e) shows the distribution for wages up to \$5.00 for the protected groups of black people, and of other minorities (non-black & non-white people). It is readily checked that blacks with low wages observe a bias towards their group that is higher than the bias observed by the other minorities.

Finally, the target attribute in the crimes dataset is ViolentCrimesPerPop, the number of violent crimes in a community per 100K individuals. As already observed, it is not actually a "decision" taken by someone. Rather, here the interest is to understand whether certain conclusions can be drawn from the dataset that relate high values of ViolentCrimesPerPop to the percentage of minorities, e.g., blacks, in the community. Such conclusions could drive forms of statistical discrimination against communities with large presence of black people. Fig. 71(f) shows the distribution of $diff()$ for the protected communities having 20% or more

of blacks. Those of such communities that have a high number of crimes (ViolentCrimesPerPop ≥ 0.20) observe a significant difference in crimes with communities that are similar to them apart from having less than 20% of blacks. In other words, a high number of crimes is "discriminatorily" present in communities with a high percentage of blacks! To further support this, Fig. 71(f) shows the distribution for communities with a low number of crimes ($ViolentCrimesPerPop < 0.20$). Now, communities with 20% or more of blacks and a low number of crimes observe no bias for such low number of crimes due to the high presence of blacks.

7.6.2 Experiments of discrimination discovery

Consider the *credit* dataset, with the settings of Fig. 70(a) and $k = 32$. Using a C4.5 decision tree classifier (Qui93), the procedure DiscoveryN(*credit*, 0.10) yields the following model and evaluation measures. Since the purpose of the classifier is descriptive, precision and recall are calculated over the training set. Notice that a predictive procedure simply consists of checking $disc(t, r)$.

num_dependents $<= 1$
| credit_amount $<= 2631$: disc $=$ yes$(59.0/9.0)$
| credit_amount > 2631 : disc $=$ no$(44.0/15.0)$
num_dependents > 1 : disc $=$ no(6.0)

disc $=$ yes : Precision0.847Recall0.769

A pair (x, y) at a leaf node means that x tuples reach the leaf, y of which are incorrectly classified. y is omitted if it is 0. Intuitively, the bad debtor class (the decision) is *discriminatorily* assigned to a female non-single (the protected group) when she has zero or one dependents and asks for a credit amount up to $\$2,631$. This is a concise, clear, and global characterization of the cases when discrimination occurred. It covers 77% of the protected group (recall), and it is accurate at 85% (precision). On the same dataset, the RIPPER rule classifier (Coh95) yields a slightly

better recall.

(credit_amount >= 3190) => disc = no(39.0/12.0)
#(installment_commitment <= 2)and(residence_since >= 3)
=> disc = no(10.0/2.0)
=> disc = yes(60.0/9.0)

disc = yes : Precision0.85Recall0.785

The model can be read as follows. Discrimination occurs in all cases, except when the credit requested is at least $\$3,190$, or when there are up to 2 installment payments (short term loans) of a resident since at least 3 years. Of course, changing a parameter of the approach may yield different models. The following decision tree is obtained for $k = 8$.

num_dependents <= 1
| installment_commitment <= 2
| age <= 23 : disc = no(9.0)
|| age > 23
||| employment = unemployed : disc = yes(1.0)
||| employment =< 1 : disc = no(8.0/1.0)
||| employment = 1 <= X < 4 : disc = yes(8.0/1.0)
||| employment = 4 <= X < 7 : disc = yes(1.0)
||| employment =>= 7 : disc = yes(4.0/2.0)
| installment_commitment > 2 : disc = yes(72.0/19.0)
num_dependents > 1 : disc = no(6.0/1.0)

disc = yes : Precision0.744Recall0.970

Discrimination occurs when there are zero or one dependents and mid to long-term loans, or short term loans to applicants older than 23 that are either unemployed or employed since at least 1 year.

	No pre-processing		0.10-correction	
classifier	accuracy	0.10-discr.	accuracy	0.10-discr.
C4.5	85.60%	4.24%	84.94%	1.07%
Naïve Bayes	82.46%	4.06%	82.33%	2.23%
Logistic	85.28%	6.61%	84.70%	0.61%
RIPPER	84.42%	5.24%	83.98%	3.94%
PART	85.20%	12.62%	84.00%	2.3%

Figure 72: Discrimination prevention on dataset adult

7.6.3 Experiments of discrimination prevention

Consider the dataset *adult* in Fig. 71, where the protected group consists of non-white people. By splitting the dataset into $2/3$ training and $1/3$ test sets, Fig. 72 shows accuracy and 0.10-discrimination for various types of classifiers, including decision trees (C4.5), Naive Bayes, logistic regression, and rule induction (RIPPER and PART). If no pre-processing correction is performed, we can observe that the dataset of predictions exhibit a 0.10-discrimination measure ranging from 4.24% to 12.62%. Overall, C4.5 exhibits the best trade-off between accuracy and non-discrimination. If the training set is pre-processed by a 0.10-correction, all the classifiers have a modest decrease in accuracy and a significant gain in non-discrimination. The logistic regression model exhibits now the best trade-off.

We have modelled the discrimination analysis problems by a variant of k-NN classification that implements the legal methodology of situation testing. Major advancements over existing proposals consist in providing: a stronger legal ground, overcoming the weaknesses of aggregate measures over undifferentiated groups; a global description of who is discriminated and who is not in discrimination discovery; a discrimination prevention method that is independent from the classification model at hand; an approach admitting interval-scaled and ordinal attributes and decisions.

Chapter 8

Conclusions and future work

Data mining is currently applied in a variety of domains, however, its models or more precisely classification models have been recently concerned of causing unfair treatments on the bases of protected-by-law attributes such as race, color, sex orientation. Hence, discrimination discovery and/or prevention in applying data mining is an urgent need.

In the meantime, one of the most popularly used methods in the exploration of discrimination in social sciences is *situation testing*. Its idea is considerably similar to association analysis in the aspect of discovering trends of groups of objects sharing a numbers of features where the transposition of those trends basically depends on changes in features of objects. Thus, a potential approach of discrimination discovery inspired from situation testing can be built with support of data mining, especially association analysis methodologies. In general, the situation testing approach based on pairs of individuals with the same features except for a sensitive one is implemented via the idea of finding such pairs by matching within the data set. The matching can be implemented in different ways: in the first one by constructing association rules and in the second one by kNN classification.

In the scope of this thesis, we propose an approach of discrimination

exploration at multiple conceptual levels. The approach is formalized as a 3-step framework for discrimination discovery on semi-structured business data. It directs how to preprocess this kind of data for discrimination discovery as pairs of attribute/value with semantic relations, which considerably supports discrimination exploration later. Subsequently, it can find possibly discriminatory dependence between the disparate treatments and protected-by-law attributes, e.g., color, ethnic origin, marital status, religion, disability, sex and represent them in the form of association rules. This process is performed through the notion of matching pairs of itemsets with different sensitive attributes and equal non-sensitive ones that are subject to different treatments-*matching items*. The concept of *matching items* is inspired by the testing pair of *situation testing*. Finally, rules at varied levels of generality are retrieved by reasoning over the hierarchy of ontology which also semantically enriches these associations by object properties between classes (concepts). Thus, they are more valuable and interesting than the flat association rules. Besides, the matching pair of situation testing can be approximated by a variant of the k-nearest neighbor (k-NN) classification, which labels each tuple in a dataset as discriminated or not. It is then feasible to perform discrimination exploration by build a classifier providing a global description of the conditions where discrimination happens.

The methodology was applied to varied datasets; one is a semi-structured garment dataset, one is a credit data set. Experimental results confirm the potential and flexibility of the approach, strongly verifying its ability in discrimination discovery.

Compared to previous studies, this approach has achieved the following advantages:

- Handling the case that the unequal decision is not binary.
- Proposing a general framework of discrimination discovery on semi-structured business data.
- Unveiling discrimination at different levels of generality.
- kNN discrimination discovery as situation testing.

We will briefly discuss each issue next.

8.1 Handling the case that the unequal decision is not binary

Often the available works use the assumption that the treatment is a binary attribute. However, this assumption does not always hold. For instance, the *accepted loan* item may have several levels for the amount of accepted money such as over $50,000$\$, $10,000$\$ to $50,000$\$, and 0 instead of saying "Rejected" or "Accepted" for the loan clients apply for. In this case, it it difficult to judge which case(s) discrimination really happens for the decision as different applicants have different profiles. A summation of these profiles for a common background is not always feasible. Additionally, it is required a lot of complex calculation to specify which attribute(s) has the negative effect causing the unequal values of the decision. To solve this matter, an intermediate binary means is proposed, the *discriminatory-indicator*. This parameter can be activated when ascertaining itemsets or more specifically *individuals* having the same background information but different (or opposite) sensitive attributes-*matching items*. If their target items are different, the *discriminatory-indicator* can be set as "*Yes*", and "*No*" otherwise. Its value is set via a "*matching analysis*", which is semantically supported by semantic rules extracted from the analysis of semi-structured data. Actually, its methodology is inspired from the method *situation testing*. The discrimination mining is then applicable on the whole dataset.

8.2 Proposing a general framework of discrimination discovery for semi-structured business data

The thesis proposes a 3-step framework for extracting discrimination information which is especially suitable to semi-structured business data. Particularly, it is supported by a common sense knowledge base of the

data domain, which is structured as a hierarchy of components (if any) and attributes of the sub-items extracted from part of semi-structured data. From the structure of the common sense knowledge base, thesauri are built to help classify words into categories and support the extraction of sub-items (the *Parsing*). Then semi-structured data will have the well-structured form of pairs of attribute/value. This knowledge domain is formalized by a general ontology for discrimination discovery. Moreover, in order to maintain the semantic relationships between sub-items, we represent these semantic relationships by means of semantic rules. They are generated by sub-items through syntactic analysis step added to the parsing. Typical semantic rules have the form of sequences of items of category thesauri, specifying their relationships. This idea is borrowed from Part-of-Speech (PoS) technique, in which extracted categorized terms from the original semi-structured data are sequentially checked through rules. Semantics relations added to the outcome of the system will help the result to be more comprehensive. Also, the number of the frequent patterns are considerably reduced, hence, it is easier for the analysis.

8.3 Unveiling discrimination at different levels of generality

Since flat association rules are considered to be weakly descriptive, i.e. the retrieved knowledge is insufficiently either general or detailed. Hence, generalization of these association rules is a reasonable solution for rich semantic association mining. We realize the discrimination mining at different levels of abstraction on ontology to obtain discriminatory rules from the specific to the high abstract levels, satisfying different views of discrimination checking. The TBox of the ontology represents the hierarchical structure of the data domain, whereas individuals of the ABox are well-structured data transferred from the relational databases. To support discrimination mining on ontology, Potentially Discriminatory (PD) Patterns (rules) are semi-automatically defined from components of the ontology (there might be support from expert users). They are used in

the reasoning service over the Abox for discovering instances satisfying these patterns. The data mining engine then searches obtained PD instances for generalized discriminatory association rules (highly frequent patterns) by using suitable measures to remove weak rules (low frequent patterns). For the purpose of quantifying the level of discrimination, we also propose a number of measures applied on the discriminatory association rules. These measures are used to obtain a precise vision of how discriminatory a non-binary decision system will be when a sensitive attribute(s) (not value) appears in certain contexts. Measures are also used to investigate which PD values are more favored (benefiting a more positive outcome treatment for the whole itemset) in comparison with others.

8.4 kNN discrimination as situation testing

There are certain limitations in the use of aggregation measures over undifferentiated sets of people, and the restriction to nominal attributes and decisions and to local models of discrimination in current method of discrimination mining. In order to overcome both these legal and technical drawbacks, we propose another approach of kNN discrimination discovery, which is still inspired by the legal experimental procedure of situation testing. Practically, it looks for pairs of people with similar characteristics apart from membership to a protected-by-law group. We approximate situation testing by a variant of the k-nearest neighbor (k-NN) classification, which labels each tuple in a dataset as discriminated or not. Discrimination discovery then reduces to build a classifier providing a global description of the conditions where discrimination occurs. Particularly, for each member of the protected-by-law group with a negative decision outcome (an individual who might claim to be a victim of discrimination) we seek testers with similar, legally admissible characteristics, apart from being or not in the protected group - k-nearest neighbors. If considerably varied treatments between the testers of the protected group and the ones of the unprotected group are recorded, it can be ascribed the negative treatment to a bias against the protected group. Similarity is modelled via a distance function. Main advancements in

comparison to existing proposals include the provision of: i)a stronger legal ground, overcoming the weaknesses of aggregate measures over undifferentiated groups; ii)a global description of who is discriminated and who is not in discrimination discovery; iii)a discrimination prevention method that is independent from the classification model at hand; iv)an approach admitting interval-scaled and ordinal attributes and decisions.

8.5 Future work

In the future, we will follow the approach of using association analysis with the support of a hierarchical structure for the purpose of revealing multiple point of views of discrimination. The relations among attributes of the background context are probably formalized by certain distance measures of nodes on the hierarchy, espeically when semantics can be added to the calculation of these measures. A formalization of semantic dependence exploiting tree theory might be then a potential direction. And if possible, de-discriminate those disparities by new measures and methods of measurement. This approach might also provide a way of preventing discrimination caused by both multiple potentially discriminatory attributes and potentially non-discriminatory attributes at different levels of abstraction.

References

[AAK+00] Epstein A.M., J.Z. Ayanian, J.H. Keogh, S.J. Noonan, N. Armistead, P.D. Cleary, J.S. Weissman, J.A. David-Kasdan, D. Carlson, J. Fuller, D. March, and R. Conti. Racial disparities in access to renal transplantation. *New England Journal of Medicine*, 343:1537–1544, 2000. 48

[AC77] D. J. Aigner and G. G. Cain. Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review*, 30:175–187, 1977. 7

[AFN98] Peter Arrijn, Serge Feld, and André Nayer. Discrimination in access to employment on grounds of foreign origin: the case of belgium. *International Migration Papers*, 1998. 8

[Agr02] A. Agresti. *Categorical Data Analysis*. Wiley-Interscience, 2002. 141

[AP01] J.G. Altonji and C.R. Pierret. Employer learning and statistical discrimination. *Quarterly Journal of Economics*, 2001. 43, 48

[Arr73] K. Arrow. The theory of discrimination. In Ashenfelter and A. Rees, editors, *In Discrimination in Labor Markets*. Princeton University Press, 1973. 44

[AS94] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proc. of VLDB 1994*, pages 487–499. Morgan Kaufmann, 1994. 16

[ATA93] Rakes A., T.Imielinski, and A.Swani. Mining association rules between sets of items in large databases. In *SIGMOD '93 Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, volume 22. Springer, 1993. 16

[Aus10] Australian Legislation. (a) Equal Opportunity Act – Victoria State, (b) Anti-Discrimination Act – Queensland State, 2010. `http://www.austlii.edu.au`. 97

[Aut08] Equal Treatment Authority. Cases of authority, 2008. http://www.egyenlobanasmod.hu/zanza/72-2008.pdf. 8

[BCP07] M. Bell, I. Chopin, and F. Palmer. *Developing Anti-Discrimination Law in Europe*. European Network of Legal Experts in Anti-Discrimination, 2007. http://ec.europa.eu/employment_social/funda_mental_rights. 7

[Bec57] G. S. Becker. *The Economics of Discrimination*. University of Chicago Press, 1957. 44

[Ben07a] Jr. Bendick. Situation Testing for Employment Discrimination in the United States of America. *Horizons Strategiques*, 5:17–39, 2007. 92

[Ben07b] Marc Bendick. Situation testing for employment discrimination in the united states of america. *Horizons stratgiques*, 2007. 8

[BM02] M. Bertrand and S. Mullainathan. Are emily and brendan more employable than lakisha and jamal? a field experiment on labor market discrimination., 2002. 43

[BMNPS03] Franz Baader, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider. *The Description Logic handbook: Theory, implementation, and applications*. Cambridge University Press, 2003. 29, 30

[Bol] Susan Bolovic. Are social networks profiles putting us at risk for business discrimination? 43 FR 38295, http://www.fastcompany.com/blog/susan-solovic/womenachievementdeservepermanentplace-recognitionsupportnationalwomensh7. 39

[Buc06] Buchanan B.G. Brief history of artificial intelligence., 2006. 43 FR 38295, http://www.aaai.org/AITopics/bbhist.html. 3

[CG00] K.Y. Chay and M. Greenstone. The convergence in black-white infant mortality rates during the 1960s. *American Economic Review*, 90, 2000. 43

[Cli03] C. Clifton. Privacy preserving data mining: How do we mine data when we aren't allowed to see it? In *Proc. of ACM KDD 2003, Tutorial*, 2003. http://www.cs.purdue.edu/homes/clifton. 6, 48

[Coh95] W. W. Cohen. Fast effective rule induction. In *Proc. of ICML 1995*, pages 115–123. Morgan Kaufmann, 1995. 152

[cou08] Swedish courts. Escape bar and restaurant v. the ombudsman against ethnic discrimination, 2008. http://www.domstol.se/domstolar/hogstadomstolen/avgoranden/ 8

[CR00] Goldin C. and C. Rouse. Orchestrating impartiality: The impact of blind auditions on female musicians. *American Economic Review*, 90:715–741., 2000. 43

[CU01] Freshtman C. and Gneezy U. Discrimination in a segmented society: an experimental approach. *The Quarterly Journal of Economics.*, 116:351–377, 2001. 43

[CV10] Toon Calders and S. Verwer. Three naive bayes approaches for discrimination-free classication. In *Proc. ECML/PKDD, 2010*, 2010. 48, 54, 56

[CZSG03] Xiaoming Chen, Xuan Zhou, Richard B. Scherl, and James Geller. Using an interest ontology for improved support in rule mining. In *DaWaK*, pages 320–329, 2003. 35

[dFE08] Claudia d'Amato, Nicola Fanizzi, and Floriana Esposito. Distance-based classification in owl ontologies. In *KES (2)*, pages 656–661, 2008. 36

[Dun03] M.H Dunham. *Data mining introductory and advanced topics*. Prentice Hall, 2003. 3

[DW96] Neal Derek and Johnson William. The role of premarket factors in black-white wage differences. *Journal of Political Economy*, 104:869–895, 1996. 43

[Dym06] G. A. Dymski. Discrimination in the credit and housing markets: Findings and challenges. *Handbook on the Economics of Discrimination*, pages 215–259, 2006. 7

[EdF10] Floriana Esposito, Claudia d'Amato, and Nicola Fanizzi. Fuzzy clustering for semantic knowledge bases. *Fundam. Inform.*, 99(2):187–205, 2010. 36

[El] EUR-lex. Online library for eus law. 43 FR 38295, http://eur-lex.europa.eu/en/index.htm. 7, 41, 42

[Ell05] E. Ellis. *EU Anti-Discrimination Law*. Oxford University Press, 2005. 7, 41, 43, 136

[ENA07] ENAR. European Network Against Racism, Fact Sheet 33: Multiple Discrimination, 2007. http://www.enar-eu.org. 151

[Eur10] European Union Legislation. (a) Racial Equality Directive, (b) Employment Equality Directive, 2010. http://ec.europa.eu/-employment_social/fundamental_rights. 7, 42, 46, 97

[Fd07] Nicola Fanizzi and Claudia d'Amato. A hierarchical clustering method for semantic knowledge bases. In *KES (3)*, pages 653–660, 2007. 36

[FLMCO04] Fernandez-Lopez, Mariano, Corcho, and Oscar. *Ontological Engineering*. Springer, first edition, 2004. 26

[FLP03] J. L. Fleiss, B. Levin, , and M. C. Paik. *Statistical Methods for Rates and Proportions*. Wiley, 2003. 141

[FPSS96] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17:37–54, 1996. 2, 3

[Gas84] J. L. Gastwirth. Statistical methods for analyzing claims of employment discrimination. *Industrial and Labor Relations Review*, 38:75–86, 1984. 7

[GCP03] Susan Gauch, Jason Chaffee, and Alexander Pretschner. Ontology-based personalized search and browsing. *Web Intelligence and Agent Systems*, 1:1–34, 2003. 35

[Ger02] Benoît Gerald. Data mining. In Blaise Cronin, editor, *Annual Review of Information Science and Technology*, volume 36. Information Today Inc. on behalf of ASIS&T, 2002. 2

[GH06] L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys*, 38(3), 2006. 5, 16

[Goe10] B. Goethals. Frequent itemset mining implementations repository, 2010. http://fimi.cs.helsinki.fi. 16

[groa] Roget group. Rogets thesaurus alphabetical index. http://thesaurus.com/roget/. 80

[grob] WordNet group. The lexical database wordnet. http://wordnet.princeton.edu/ . 80

[Gru93] T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199–220, 1993. 26

[H.A03a] David H.Autor. The economics of discrimination - evidence. In *Lecture note: The economics of discrimination*, MIT 14.661, 2003. 43

[H.A03b] David H.Autor. The economics of discrimination - theory. In *Lecture note: The economics of discrimination*, MIT 14.661, 2003. 43

[Han01] D. J. Hand. Modelling consumer credit risk. *IMA Journal of Management mathematics*, 12:139–155, 2001. 5, 6, 48

[HCXY07] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, 2007. 16

[HH97] D. J. Hand and W. E. Henley. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society, Series A*, 160:523–541, 1997. 6, 48

[HISK05] A. A. Hintoglu, A. Inan, Y. Saygin, and M. Keskinöz. Suppressing data sets to prevent discovery of association rules. In *Proc. of IEEE ICDM 2005*, pages 645–648. IEEE Computer Society, 2005. 10, 48, 49

[HL03] Holzer H.J. and J. Ludwig. Measuring discrimination in education: Are methodologies from labor and housing markets useful? *Teachers College Record*, 115:1147–1178, 2003. 43

[HMS01] David Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining*. The MIT Press, first edition, 2001. 2

[HN04] H. J. Holzer and D. Neumark, editors. *The Economics of Affirmative Action*. Cheltenham: Edward Elgar, 2004. 7

[HPSB+04] Ian Horrocks, Peter F. Patel-Schneider, Harold Boley, Said Tabet, Benjamin Grosof, and Mike Dean. Swrl: A semantic web rule language combining owl and ruleml, 2004. `http //www.w3.org/Submission/SWRL/`. 30, 31

[HS01] Ian Horrocks and Ulrike Sattler. Ontology reasoning in the shoq(d) description logic. In *Proc. of the 17th Int. Conference on Artificial Intelligence (IJCAI 2001)*, 2001. 24, 29

[HT00] Ian Horrocks and Stephan Tobies. Reasoning with axioms: Theory and practice. In *Proc. of the 7th international conference on principles of knowledge representation and reasoning(KR 2000)*. Morgan Kaufmann, 2000. 24

[HY95] Jiawei H. and Y.Fu. Discovery of multiple-level association rules from lager databases. In *Proc. 21st VLDB Conferene Zurich, Swizerland*, 1995. 17

[J.07] Millman J. Largest discrimination case in history: Wal-mart case's appeal denied. *Diversity Inc Magazine. February 7th 2007.*, 2007. 7

[JH00] D. Jurafsky and Martin J. H. *Speech and Language Processing*. Prentice Hall, second edition, 2000. 79

[JM00] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice Hall., 2000. 21

[JM01] Han J. and Kamber M. *Data mining: concepts and techniques*. Morgan Kaufmann, first edition, 2001. 2, 3

[KC09] Faisal Kamiran and Toon Calders. Classification without discrimination. In *IEEE Int.'l Conf. on Computer, Control & Communication (IEEE-IC4)*. IEEE press, 2009. 6, 48, 49, 54

[KC10] Faisal Kamiran and Toon Calders. Classication with no discrimination by preferential sampling. In *Proceedings of the 19 th Machine Learning conference of Belgium and The Netherlands*, 2010. 48, 55

[KCP09] Faisal Kamiran, Toon Calders, and M. Pechenizkiy. Building classiers with independency constraints. In *IEEE ICDM Workshop on Domain Driven Data Mining*. IEEE press, 2009. 48, 54, 55, 99

[KCP10] Faisal Kamiran, Toon Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. In *IEEE International Conference on Data Mining*. IEEE press, 2010. 48, 54, 57

[KDn] KDnutggets. Data mining applications. 43 FR 38295, `http://www.kdnuggets.com/polls/2008/datamining-applications.htm`. 5

[Kno86] R. Knopff. On proving discrimination: Statistical methods and unfolding policy logics. *Canadian Public Policy*, 12:573–583, 1986. 97

[Kro93] R. Krovetz. Viewing morphology as at inference process. In *Proceedings OF the Sixteenth Annual internationally ACM SIGIR Conference on Research and development in information retrieval.*, 1993. 22

[Ler91] N. Lerner. *Group Rights and Discrimination in International Law*. Martinus Nijhoff Publishers, 1991. 136

[LL99] M. LaCour-Little. Discrimination in mortgage lending: A critical review of the literature. *Journal of Real Estate Literature,*, 7:15–49, 1999. 7

[LO08] Zhuhadar Leyla and Nasraoui Olfa. Personalized cluster-based semantically enriched web search for e-learning. In *Proceeding of the 2nd international workshop on Ontologies and information systems for the semantic web*, ONISW '08, pages 105–112, 2008. 36, 37

[LO10] Zhuhadar Leyla and Nasraoui Olfa. Augmented ontology-based information retrieval system with external open source resources. In *Proceedings of the 2010 Seventh International Conference on Information Technology: New Generations*, ITNG '10, pages 144–149, 2010. 36, 37

[Mak06] T. Makkonen. *Measuring Discrimination: Data Collection and the EU Equality Law*. European Network of Legal Experts in Anti-Discrimination, 2006. `http://ec.europa.eu/employment_social/fundamental_rights`. 7

[Mak07] T. Makkonen. *European handbook on equality data*. European Network of Legal Experts in Anti-Discrimination, 2007. `http://ec.europa.eu/employment_social/funda_mental_rights`. 7, 42

[MBGV07] D. Martens, B. Baesens, T. Van Gestel, and J. Vanthienen. Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, 183(3):1466–1476, 2007. 5

[MG08] Gersper M. and T. Gould. Gender and age discrimination costs u.s. importers billions. *In: US Custom House guide*, 2008. 74, 75, 121

[Mic07a] Barbaro Michael. Clothing makers allege sex discrimination in u.s. tariffs. *International Herald Tribune.*, 2007. 7, 74, 75, 121

[Mic07b] Barbaro Michael. In apparel, all tariffs aren't created equal. *New York Times*, 2007. 74, 75, 121

[M.J91] Katz M.J. The economics of discrimination: The three fallacies of croson. *Yale Law Journal*, 1991. 7, 10, 48

[MSS04] Boris Motik, Ulrike Sattler, and Rudi Studer. Query answering for owl-dl with rules. In *Journal of Web Semantics*, pages 549–563. Springer, 2004. 24

[New98] R. G. Newcombe. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, 17:873–89, 1998. 141

[NHBM98] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz. UCI repository of machine learning databases, 1998. http://archive.ics.uci.edu/ml. 76

[oNS04] Committee on National Statistics. Measuring racial discrimination, 2004. 40, 43, 60

[OZ03] S. R. M. Oliveira and O. R. Zaiane. Protecting sensitive knowledge by data sanitization. In *Proc. of the 3rd IEEE International Conference on Data Mining*, 2003. 49

[Phe72] E.S. Phelps. The statistical theory of racism and sexism. *American Economic Review*, 62:659–661, 1972. 44

[Por80] M.F. Porter. An algorithm for suffix stripping. *Program.*, 14(3):130137, 1980. 22

[PRT07] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. Technical Report 07-19, Dipartimento di Informatica, Università di Pisa, September 2007. http://compass2.di.unipi.it/TR. 48, 58, 76, 127

[PRT08] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proc. of ACM KDD 2008*, pages 560–568. ACM, 2008. 6, 10, 48, 58, 68, 76, 97, 99, 127, 136, 140, 146

[PRT09a] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Integrating induction and deduction for finding evidence of discrimination. In *Proc. of ICAIL 2009*, pages 157–166. ACM, 2009. 10, 48, 58, 99, 127

[PRT09b] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Measuring discrimination in socially-sensitive decision records. In *Proc. of the SIAM SDM 2009*, pages 581–592. SIAM, 2009. 6, 48, 58, 95, 97

[PRT10] D. Pedreschi, S. Ruggieri, and F. Turini. Dcube: Discrimination discovery in databases. In *SIGMOD 2010.* ACM, 2010. 10, 61

[PSG+05] Bijan Parsia, Evren Sirin, Bernardo Cuenca Grau, Edna Ruckhaus, and Daniel Hewlett. Cautiously approaching swrl., 2005. http //http://www.mindswap.org/papers/CautiousSWRL.pdf. 31

[PW99] M. J. Piette and P. F. White. Approaches for dealing with small sample sizes in employment discrimination litigation. *Journal of Forensic Economics*, 12:43–56, 1999. 7

[Qui93] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993. 152

[Ror09] I. Rorive. Proving discrimination cases - the role of situation testing, 2009. Centre For Equal Rights & Migration Policy Group, `http://www.migpolgroup.com/publications.php`. 42, 46, 92

[RPA03] Navigli Roberto, Velardi Paola, and Gangemi Aldo. Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems*, 18:22–31, January 2003. 34

[RPT10] S. Ruggieri, D. Pedreschi, and F. Turini. Data mining for discrimination discovery. *ACM Trans. on Knowledge Discovery from Data*, 4(2):Article 9, 2010. 68, 95

[RR02] P. A. Riach and J. Rich. Field experiments of discrimination in the market place. *The Economic Journal*, 112:480–518, 2002. 7

[SA94] R. Srikant and R. Agrawal. Fast algorithms for mining generalized association rules. In *Proc. of 20th VLDB94*, pages 487–499. Morgan Kaufmann, 1994. 17

[Sac04] Sachdev. A.: Merrill lynch to pay $2.2 million in damage in sexual discrimination case. *Chicago Tribune newspaper. April 24th 2004*, 2004. 7

[SC03] Stefan Schlobach and Ronald Cornet. Explanation of terminological reasoning a preliminary report. In *Proc. of International Workshop on Description Logics. (DL2003)*, 2003. 24, 29

[SL96] Fortin Scott and Liu Ling. An object-oriented approach to multilevel association rule mining. In *Proceedings of the fifth international conference on Information and knowledge management*, CIKM '96, pages 65–72, 1996. 34

[SMB07] Ahu Sieg, Bamshad Mobasher, and Robin Burke. Learning ontology-based user profiles: A semantic approach to personalized web search. *The IEEE Intelligent Informatics Bulletin*, 8:7–18, 2007. 35

[SO01] Squires and G.D.and S. O'Connor. Colour and money: Politics and prospects for community reinvestment in urban america. *Journal of Urban Affairs*, 4, 2001. 7, 48

[Squ03] G. D. Squires. Racial profiling, insurance style: Insurance redlining and the uneven development of metropolitan areas. *Journal of Urban Affairs*, 25(4):391–410, 2003. 7

[SS98] Li Shen and Hong Shen. Mining flexible multiple-level association rules in all concept hierarchies. In *In DEXA '98*, volume 1460 of *Lecture Notes in Computer Science*. Springer, 1998. 17

[Swe01] L. Sweeney. *Computational Disclosure Control: A Primer on Data Privacy Protection*. PhD thesis, MIT, Cambridge, MA, 2001. 48, 49, 50

[Swe02] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty and Fuzziness in Knowledge-Based Systems*, 10(5):571–588, 2002. 10, 48, 49, 50

[TG04] Joana Trajkova and Susan Gauch. Improving Ontology-Based user profiles. In *Proceedings of RIAO*, pages 380–389, 2004. 35

[Tho00] L. C. Thomas. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16:149–172, 2000. 5, 6, 48

[TL01] Ming-Cheng Tseng and Wen-Yang Lin. Mining generalized association rules with multiple minimum supports. In *In DaWaK 2001*, volume 2114 of *Lecture Notes in Computer Science*. Springer, 2001. 17

[TL04] Ming-Cheng Tseng and Wen-Yang Lin. Generalized association rules with evolving taxonomies. In *Int. Computer Symposium*, 2004. 17

[TLN07] Xiaohui Tao, Yuefeng Li, and Richi Nayak. Ontology mining for semantic interpretation of information needs. In *KSEM'07*, pages 313–324, 2007. 36

[TMRMB09] Amir Tal, Galia Moran, Dan-Olof Rooth, and Jr. Marc Bendick. Using situation testing document employment discriminat. *Employee relations law journal*, 35, 2009. 8

[TSK06] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2006. 3, 16

[U.K10] U.K. Legislation. (a) Sex Discrimination Act, (b) Race Relation Act, 2010. `http://www.statutelaw.gov.uk`. 7, 46, 61, 97, 98, 142

[Uni08] United Nations Legislation. (a) Universal Declaration of Human Rights, (b) International Covenant for Civil and Political Rights, (c) International Covenant on Ecomomic, Social and Cultural Rights, (d) Convention on the Elimination of All forms of Racial Discrimination, (e) Convention on the Elimination of All forms of Discrimination Against Women, 2008. `http://www.ohchr.org`. 7, 39, 42

[Uni10] United Nations Legislation. (a) Convention on the Elimination of All forms of Racial Discrimination, (b) Convention on the Elimination of All forms of Discrimination Against Women, 2010. http://www.ohchr.org. 7, 42

[U.S78a] U.S. Federal Legislation. Uniform guidelines on employee selection procedure, 1978. 43 FR 38295, http://www.usdoj.gov-/crt/emp/uniformguidelines.html. 98

[U.S78b] U.S. Federal Legislation. Us harmonized tariff schedules 2008., 1978. 43 FR 38295, http://http://hts.usitc.gov. 72

[U.S10] U.S. Federal Legislation. (a) Equal Credit Opportunity Act, (b) Fair Housing Act, (c) Intentional Employment Discrimination, (d) Equal Pay Act, (e) Pregnancy Discrimination Act, (f) Civil Right Act, 2010. http://www.usdoj.gov. 7, 40, 142

[VEB+04] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni. Association rule hiding. *IEEE Transactions on Knowledge and Data Engineering*, 16(4):434–447, 2004. 10, 48, 49

[VK06] M. Vojtek and E. Kočenda. Credit scoring methods. *Journal of Economics and Finance*, 56:152–167, 2006. 5

[Vou95a] Atro Voutilainen. A syntax-based part of speech analyse. In *Proc. of the Seventh Conference of the European Chapter of the Association for Computational Linguistics.*, 1995. 22

[Vou95b] Atro Voutilainen. A syntax-based part of speech analyser. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics.Association for Computational Linguistics.*, 1995. 79

[VSC01] V. S. Verykios, Y. Saygin, and Chris Clifton. Using unknowns to prevent discovery of association rules. *Sigmod records*, 30(4):45–54, 2001. 49

[WA00] Ronald H. Welch and Carlos T. Angulo. Justice on trial: Racial disparities in the american criminal justice system. In *Washington, DC: Leadership Conference on Civil rights/Leadership Conference Education Fund*, 2000. 48

[WFY05] K. Wang, B. C. M. Fung, and P. S. Yu. Template-based privacy preservation in classification problems. In *Proc. of IEEE ICDM 2005*, pages 466–473. IEEE Computer Society, 2005. 10, 48, 49

[WH08] Chieh-Ming Wu and Yin-Fu Huang. An efficient data structure for mining generalized association rules. In *Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, 2008. 17

[WIZT04] S.M. Weiss, N. Indurkha, and F.J. Zhang T., Damerau. *Text mining: Predictive methods for analyzing unstructured information.* Springer, first edition, 2004. 4

[YH03] X. Yin and J. Han. CPAR: Classification based on Predictive Association Rules. In *Proc. of SIAM SDM 2003*, pages 331–335. SIAM, 2003. 10, 48

[YN03] Li Yuefeng and Zhong Ning. Ontology-based web mining model: Representations of user profiles. In *Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence*, WI '03, pages 96–103, 2003. 35

[YN04] Li Yuefeng and Zhong Ning. Capturing evolving patterns for ontology-based web mining. In *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '04, pages 256–263, 2004. 35