DISCRIMINATORY ASSOCIATION ANALYSIS ON SEMI-STRUCTURED DATA

Binh Luong Thanh[1] and Franco Turini[2]
[1]*IMT Institute for Advanced Studies Lucca – Piazza S. Ponziano, 6, 55100 Lucca, Italy*
[2]*Dipartimento di Informatica, Università di Pisa - Corso Italia 40, I-56125 Pisa, Italy*

ABSTRACT

Data mining has been applied to the discovery of illegally discriminatory treatments caused by protected-by-law attributes such as race, gender, age, etc. In this paper, we propose an improvement for the previous work of exploring discrimination in semi-structured business data. The main idea is that discrimination represented in the form of association rules is judged by opposite patterns whose components are almost the same except a single sensitive attribute and the decision (admission to school, acceptance to a position, etc.) However, the previous work requires numerous efforts and it has not been modeled in a systematic way. In order to solve this limitation, semantic analysis is integrated in the discrimination mining process showing better results in comparison with the previous work in experimental outcome.

KEYWORDS

Discrimination, association analysis, HTS, semantic analysis.

1. INTRODUCTION

Discrimination has been studied for a long time by economists, sociologists and legislators in society, i.e., studies in racial profiling and redlining [16], [17], personnel selection [5], [8] and mortgage granting [9], etc. One of common causes of discrimination is gender, for example, women were often paid less or had narrowed opportunities to promotion or bonus, e.g. unjust treatments in the Merrill Lynch case 2004 [15], Wal-mart case 2007 [11], discriminatory tariffs in Totes-Isotoner case [1] and so on. These cases and others motivate a need of implementing knowledge discovery models, particularly association analysis to extract proofs of relations between the discriminatory treatments and sensitive attributes e.g., gender, age, and race.

However, data are not often well-structured or clean due to noises, or mistakes of user, also the arguable treatments are not binary, hence, direct mining on such data is impossible. This paper introduces an improved framework of [10] for discrimination discovery on semi-structured business data, which are business data that not completely structured as attribute/value pairs, but contain textual parts that can be split into smaller components. The previous framework can mine discrimination from this data in the form of association rules supported by a common sense knowledge base. Nevertheless, the main step of finding discrimination based on opposite patterns whose components are almost the same (except a single protected-by-law attribute, e.g., race, gender and the decision) requires numerous efforts and it has not yet been modeled in a well systematic way. In order to solve this limitation, the improved framework deploys semantic analysis, borrowed from the idea of Part of Speech-PoS of text mining [22], [6], during the opposite patterns check supported by a semantic description of input data. This approach shows better results in discovering opposite patterns, it also helps to reduce the number of database scans. The problem of discrimination-aware data mining was introduced in [2] about classifiers that might execute racial discrimination. Generally, there are three non-mutually exclusive strategies towards discrimination discovery. The first one is to adapt the preprocessing approaches of data sanitization [4], [21] and hierarchy-based generalization [18], [23]. Along this line, [7] adopts a controlled distortion of the training set. The second one is to post-process the produced classification model, in which [12], [13] propose a confidence-altering approach for classification rules inferred by the CPAR algorithm [25]. The third one is to modify the classification-learning algorithm by internally integrating discrimination measures, which is the motivation for our research.

The remainder of this paper is structured as follows. Background of association mining is summarized in Section 2. A case study of the HTS problem is represented in Section 3. Section 4 defines the problem and represents the discriminatory association analysis framework. Experimental results are shown in section 5. Finally, conclusion and future development are presented in section 6.

2. PRELIMINARIES

Association analysis is a process of discovering implicit interesting relationships in large datasets in the form of association rules [19]. Let $\Re$ be a non-empty relation over attributes $a_1, \ldots, a_n$, namely $\varnothing \subset \Re \subseteq dom(a_1) \times \ldots \times dom(a_n)$, $dom(a_i)$ is the domain of values of a, 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. An item is any a-item. A c-item is called a class item. Let I be the set of all items, 2^I denotes the set of all itemsets. An itemset $\mathbf{X} \subseteq I$, for a tuple $\sigma \in \Re$, it is said that σ verifies $\mathbf{X}$ if $\sigma \models \mathbf{X}$, namely for every $a = v$ in $\mathbf{X}$, $\sigma(a) = v$. The absolute support of an itemset $\mathbf{X}$ is the number of tuples in $\Re$ verifying $\mathbf{X}$: $asupp(\mathbf{X}) = |\{ \sigma \in \Re \mid \sigma \models \mathbf{X} \}|$, where $|\ |$ is the cardinality operator. The (relative) support of $\mathbf{X}$ is the ratio of tuples verifying $\mathbf{X}$ over the cardinality of R: $supp(\mathbf{X}) = asupp(\mathbf{X})/|\Re|$.

3. A CASE STUDY

The US Harmonized Tariff Schedule-HTS [20] is used to determine tariff for imported merchandise, including nomenclatures, descriptions for goods, and formulae for calculating tariff. Though it is carefully built and updated conforming to the law, [1] has uncovered an oddity that duties on men's and women's garments are different for no apparent reason; its calculation shows that the government imposes a 14 percent tariff on women's, but only 9 percent on men's on overall. According to [3], US importers have overpaid more than 1.3 billion dollars for discriminatory duties. Thus, a number of clothing importers i.e., Totes Isotones, Steve Madden filed a lawsuit alleging discrimination in tariff on the base of gender [1].

4. DISCRIMINATORY ASSOCIATION ANALYSIS

4.1 Problem statement

Let I be a database of itemsets referring to a set of attributes: $A = a_1, \ldots, a_n$, $dom(a_i) = \{d_{ij}\}$. A target attribute a_{tg} is used for classifying itemsets such as a decision (e.g., the admission to school, acceptance to a position), the cost of service, etc. Its value depends on other attributes but this dependence cannot be represented as a simple function, $dom(a_{tg}) = c_1, \ldots, c_m$. And *des* is a description attribute, containing a number of hidden elementary attributes and is expressed in a relatively free form, for instance, a description of a 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 type of data is significantly important since it contains overt knowledge but mining is possible if and only if each description des_i is represented by sub-items: $\{a_{i1}, \ldots, a_{ip}\}$, $dom(a_{ij}) = \{v_{ij1}, \ldots, v_{ijw}\}$. For example, the above description can be mapped into sub-items: *Gender = male*, *Name of goods* = {*suits*}, *Form of production* = {*not knitted, not crocheted*}, *Quantity = 36%*, *Materials* = {*synthetic fibers, wool, fine animal hair*}. These sub-items can be divided into sensitive group such as the *Gender* item and 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 which may cause discriminatory treatments, so-called *potentially discriminatory*-PD attributes [12], [13]. Its complementary set $\overline{S}$ is composed of *potentially non-discriminatory*-PND attributes $\{u_i\}$. The transactions of the database hereafter have the form of a three-tuple: $(\{u_i\}, \{v_i\}, c_i)$ where $\{u_i\}$ is the set of non-sensitive items, $\{v_i\}$ is the

set of sensitive items, and c_i is the target item. If the target attributes a_{tg} is observed to be strongly related to sensitive attributes S, it can be said that this correlation is discriminatory, or represented in association rules:

$$\{u_k\}, \{v_k\} \rightarrow c, \quad (1)$$
$$\{u_k\}, \{v_k'\} \rightarrow c'$$

where:

- $\{u_k\}$ is a subset of $\overline{S}$ 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.

Yet, the results of (1) are hard to achieve and probably do not provide valuable information if there is discrimination or which item(s) causes discrimination due to: i) if the target item is *not* binary then it is difficult to clarify in which case(s) discrimination really happens and which attribute(s) has the negative effect on the target attribute; ii) it is required numerous efforts to analyze *des* into sub-items to compare the effect of sensitive attributes through the association analysis, especially when the database is large. This paper proposes an improved framework of the previous work [10] for unveiling discriminatory associations in semi-structured business data, supported by semantic analysis as presented in Figure 1.

4.2 The proposed framework

4.2.1 Parsing

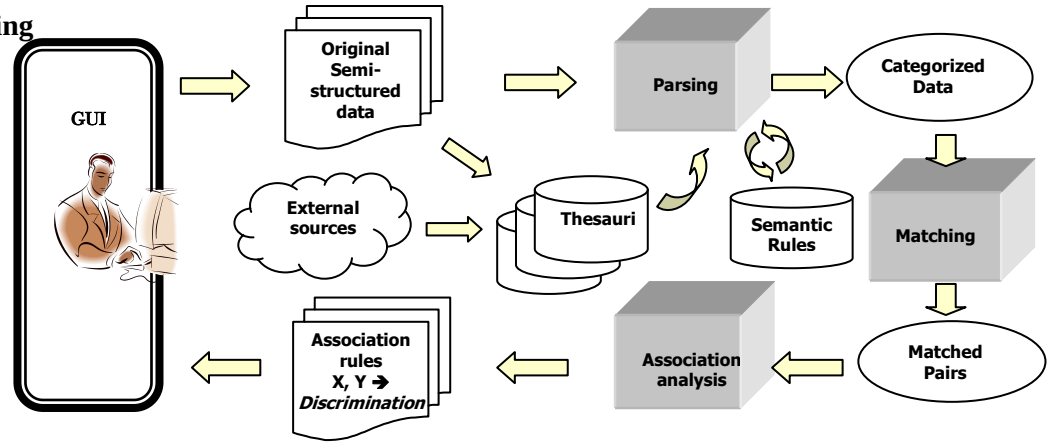

Figure 1. Framework of discriminatory association analysis.

The original data as illustrated in Figure 2a will be firstly combined with external sources of knowledge, e.g. WordNet's thesaurus [24], Roget's thesaurus [14] 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 this structure, thesauri are built to classify extracted terms (sub-items) from the *des*. There are two kinds of thesauri [10]:

- *Form thesauri*: for normalizing extracted words, e.g. thesaurus of synonyms, abbreviations. Any word extracted from a description matches a relevant word of an entry of the thesaurus, e.g. "*not, excluding*", an entry of the negative thesaurus, it will be replaced by its corresponding term, "*not*".
- *Category thesauri*: clarify categories of terms, i.e. thesaurus of materials of goods: *silk*, *fur*, etc.

Therefore, redundant information such as *stop words*, i.e., "*to*", "*the*", "*of*" is removed; only meaningful terms which can be categorized into a particular thesaurus are kept, which means the original description item is transformed into pairs of attribute/value as in Figure 2b. However, these data cannot convey the semantic relationships between sub-items; for example, 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 borrowed from PoS technique of natural language processing. For instance:

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

This rule states that the *materials 1* belonging to the object level (*name*) have level 1; whereas the *materials 2* belonging to the component of object (*parts* item) have level 2 but not others if the checked sub-itemset satisfies the order of the rule. In order to implement semantic analysis, we introduce the following

Semantic_Rule_Forming algorithm that 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:

```
Algorithm Semantic_Rule_Forming (dataset of category termed descriptions D)
BEGIN
  R = ∅ // 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_i
R_i = r_i ∪ R'//if r_i satisfies initial rule(s) of R', add both to the combinatorial semantic R_i
If R_i ⊄ R Then R ← R_i // if R_i is new to R then add R_i to R.
  end
END.
```

ID	Description	Quantity2	Code	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	Limitation	Negative
male	*suit*	*fine animal hair*	*wool*	*synthetic fiber*	*knitted*	*crocheted*	*36%*	*more*	*not*

(b)

Gender	Name of goods	Material			Form		Quantity	Limitation
male	*suit*	*fine animal hair*	*wool*	*synthetic fiber*	*(not)knitted*	*(not)crocheted*	*36%*	*more*

(c)

Figure 2. Example of transformation from semi-structured data to well-structured data.

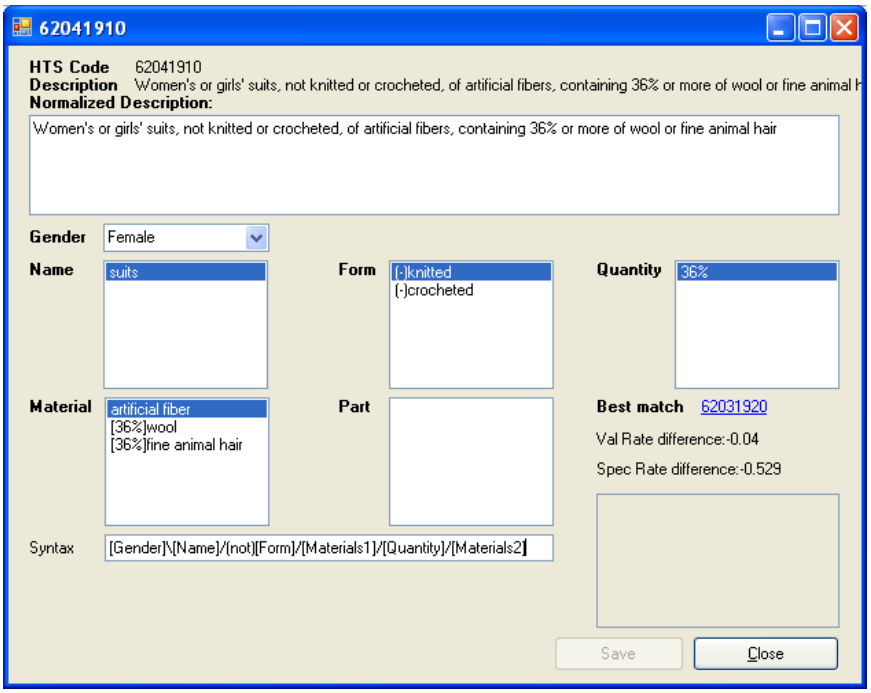

Figure 3. Analyzing HTS data.

First, users direct the system what rules will be by adding simple initial rules, which are experiences of users in a specific domain. Rules are defined as sequences of items of *category thesauri*, specifying their relationships. For instance, an initial rule for a clothing database "*(number "%" + materials) + forms*" which 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, complicated

rules are generated by combining these initial rules in tandem or with other items of category thesauri when scanning the database. In the end, the returned result is a well-defined structure with semantics constraints as shown in Figure 2c. A typical example for this process is provided in Figure 3.

4.2.2 Matching

The main purpose of the framework is checking possibly discriminatory behaviors caused by PD attributes whose task is finding "*matching*" itemsets that have the same background information but different sensitive attributes. For example, the data in Figure 2a 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 almost the same sub-items except the gender (*female* vs. *male*). It is often difficult to discover matching pairs even when descriptions are analyzed as in Figure 2c since it is required exhaustively scanning the database, which is sometimes infeasible, especially when dataset is huge. We propose an approach as in Figure 4 for finding matching itemsets, the core task of discrimination discovery, using semantic rules that requires to scan neither the whole dataset nor every attribute as [10]; only itemsets satisfying the same semantic rules of the given itemsets are checked which helps to increase both performance and precision.

```
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_l in AttributeSet
                          If u_li ≠ u_lik Then
                          begin
                                 isMatch = false
                                 break
                          end

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

Figure 4. Matching algorithm based on semantic rules.

We recall the main concepts of this approach [10] for discrimination association analysis.

DEFINITION 1. *A discriminatory indicator θ_i for the target item caused by the PD attribute s_i on a given context $\{u_k\}$ is defined as following: given two tuples:* $(\{u_k\},\{v\},c)$, $(\{u_k\},\{v'\},c')$.

$$\textit{The value of } \theta_i \textit{ is defined as: } \theta_i = \begin{cases} 1 & \textit{if} \quad c \neq c' \\ 0 & \textit{if} \quad c = c' \end{cases}$$

where:

- $\{u_k\} \subseteq \overline{S}$ defines a background context
- $\{v\},\{v'\}$: sensitive attributes with different values.
- c,c': target attribute a_{tg} with different values.

It can be explained that for a given itemset, the discriminatory indicator is activated when another itemset that is different in values of PD attribute(s) and target attribute on the same context is found. 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. Thus, 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 situations may happen:

- there are several itemsets satisfying the triple: $(\{u_k\},v_i, \theta_i = 1)$

In this case, the following association is generated:

$$\{u_k\}, v_i \rightarrow \textit{Discrimination} \qquad (2)$$

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.

- there are several itemsets satisfying the triple: $(\{u_k\},\{v_i\},\{\theta_i = 1\})$

In this case the following associations are generated, one for each PD item in the set $\{v_i\}$

$$\{u_k\},\{v_i\} \rightarrow \textit{Discrimination} \qquad (3)$$

(2) and (3) are called *discriminatory association rules*. In the two cases, the multiple-valued target item is replaced by a *binary* item, the *discriminatory indicator* θ from which it is possible to discover discrimination or PD attribute(s) negatively affect target attribute is discriminatory commonly by association analysis.

4.2.3 Association analysis

Finally, the discriminatory association mining is implemented to compute the confidence of each of the possibly discriminatory associations mentioned in the previous step. Any association rules mining algorithm can be applied, e.g., Apriori [26], FPGrowth [27], etc. Given a user-specified α-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 discriminatory treatment. If no convincing argument is given for the negative effect of such attributes on the discriminatory decision, they should be considered wrong/illegal and removed in the decision-making process. Examples of possibly archived association rules will be presented in the Section 4.

5. EXPERIMENT

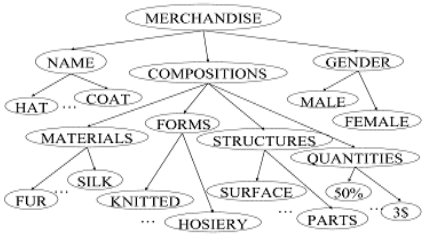

Figure 5. Distribution of generalized discriminatory rules in HTS

The framework was applied to the mentioned HTS dataset, which consists of 824 categories of garment products. This dataset is semi-structured since there is a rich information-attribute as presented in Figures 2a. A common sense knowledge base has been built as shown in Figure 5 from the original data. Additionally, thesauri of synonyms, abbreviations, negative meanings are also built. Each apparel is cleaned and standardized as in Figure 2b and hierarchically structured as in Figure 2c: data are categorized into the name of the apparel, its gender (for which sex group it is produced), material, form of production, quantities of each material, and quantities for its components (structure). It is obvious that the tariff attribute is the target attribute, the gender attribute is the PD attribute and all the others form the background context. Based on semantic rules, matching pairs are found easier and more precise than the basic approach of [10] as shown in Table 1.

Table 1: Comparison between two approaches

Parameters	Without semantic analysis	With semantic analysis
Number of extracted matching pairs	345	290
Number of exactly analyzed	275	289
Accuracy	79.71%	99.66%

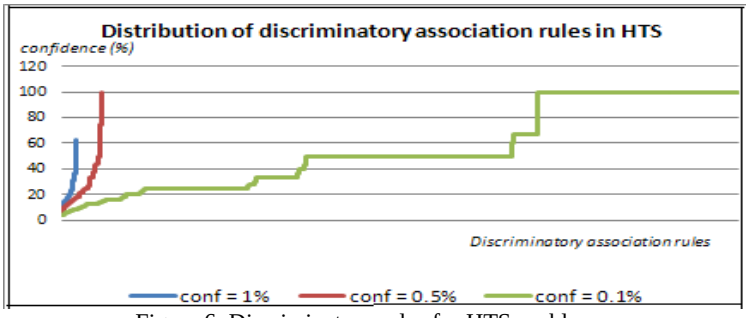

Figure 6. Discriminatory rules for HTS problem

Discriminatory rules mined at different minimum supports are represented in Figure 6, showing strong correlation between discriminatory tariffs and gender, examples of discriminatory rules are as the following:

```
form = "knitted", "crocheted", material = "fine animal hair", "wool"
➔ discrimination (conf = 73.53%)

form = "not_knitted", "not_crocheted", material = "fine animal hair", "wool"
➔ discrimination (conf = 50%)
```

In order to measure the level of different tariffs between male and female products, we use discrimination measures proposed in [10]. Let $\bar{s}, s \rightarrow \theta_i = 1$ (discrimination) and $\bar{s} \rightarrow \theta = 1$ be association rules with confidences correspondingly ρ and φ. We have:

- Absolute difference:
  $abs\_lift = \varphi - \rho$ (4)
- Relative difference:
  $rev\_lift = (\varphi - \rho)/\varphi$ (5)
- Ratio gain of difference:
  $rg\_lift = \varphi/\rho$ (6)

where:

- $\bar{s} \subseteq \bar{S}$ are PND items
- $s \subset S$ is a set of PD items

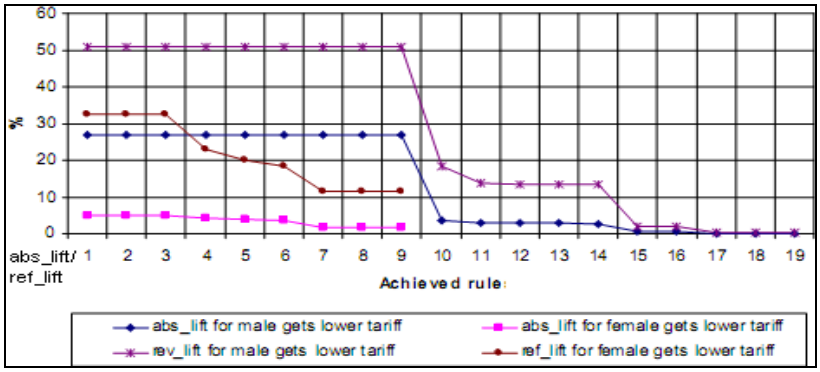

Figure 7a. Distribution of discrimination in tariff between male and female in HTS w.r.t. *abs_lift* and *rev_lift*.

When $|S_i| = 1, \bar{s}, \neg s \rightarrow \theta_i = 1$ has the confidence 1 - ρ where $\neg S_i$ is the opposite value of S_i:

- Absolute opposite difference:
  $abs\_lift = 1 - 2\rho$ (7)
- Relative opposite difference:
  $rev\_lift = (1 - 2\rho)/(1 - \rho)$ (8)
- Ratio gain opposite difference:
  $rg\_lift = (1 - 2\rho)/(1 - \rho)$ (9)

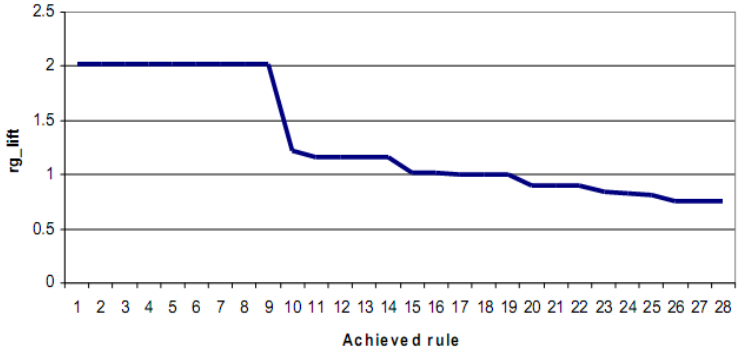

Figure 7b. Distribution of discrimination in tariff between male and female in HTS w.r.t. *rg_lift*.

Experiment shows that for a given kind of apparels, male often receives lower tariff than female as illustrated in Figure 7a and Figure 7b. Moreover, the difference of male/female rates is not high if female is imposed lower tariff that is opposite to the situation when male retrieves lower tariff. The common causes of discriminatory tariffs are the MFN rate type 7, "knitted", "crocheted", "not knitted", and "not crocheted" for the form of production, "silk", textile, "wool" for materials with a relatively high confidence for different values of minimum support. In general, it can be said that apparels for male are favored in the tariff.

6. CONCLUSIONS

Data mining has shown its critical usefulness in analysis, it is then expected to support discrimination discovery as well. This paper introduces improvements of previous work for discrimination discovery on business data in which strong correlations between sensitive attributes and discriminatory treatments are represented as association rules. The new framework uses semantic analysis supporting rules generation that

helps to increase performance and precision of the mining process. In the future, we will focus on representing discrimination at varied generalized levels using formal structures e.g. ontology.

REFERENCES

[1.] Barbaro, M., 2007. In apparel, all tariffs aren't created equal. *New York Times* April 28th 2007. New York, USA.

[2.] Clifton C., 2003. Privacy preserving data mining: How do we mine data when we aren't allowed to see it? *Tutorials at The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Washington, USA.

[3.] Gersper, M., and T. Gould, 2008. Gender and Age discrimination costs U.S. importers billions. *US Custom House guide*.

[4.] Hintoglu, A.A. et al, 2005. Suppressing data sets to prevent discovery of association rules. *Proc. The fifth IEEE International conference on Data Mining*. Houston, Texas, USA, pp. 645-648.

[5.] Holzer, H. et al, 2004. Black job applicants and the hiring officer's race. *Industrial and Labor Relations Review*, 57(2), pp.267–287.

[6.] Jurafsky, D., Martin J. H., 2000. *Speech and Language Processing*, Second Edition. Prentice Hall.

[7.] Kamiran, F., and T. Calders, 2009. Classifying without Discriminating. *Proc. 2nd IEEE International Conference on Computer, Control and Communication - IC4 2009.* Karachi, Pakistan, pp. 1-6

[8.] Katz, M.J., 1991. The Economics of Discrimination: The three Fallacies of Croson. *Yale Law Journal.* pp. 1000-1033.

[9.] LaCour-Little, M., 1999. Discrimination in mortgage lending: A critical review of the literature. *Journal of Real Estate Literature*, 7, pp. 15-50.

[10.] Luong, T. B., F. Turini, 2010. Association analysis of semi-structured data for discrimination discovery in business. *Proceedings of The sixth conference on Data Mining- DMIN 2010*. Las Vegas, Nevada, USA, pp. 193-199.

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

[12.] Pedreschi, D. et al, 2008. Discrimination aware data mining. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Las Vegas, Nevada, USA, pp. 560-568,

[13.] Pedreschi, D.et al, 2009. Integrating induction and deduction for finding evidence of discrimination. *Proc. of The Twelfth International Conference on Artificial Intelligence and Law-ICAIL 2009*. Barcelona, Spain, pp. 157-166.

[14.] Roget's Thesaurus alphabetical index. http://thesaurus.com/roget/

[15.] Sachdev, A., 2004. Merrill Lynch to pay $2.2 million in damage in sexual discrimination case. *Chicago Tribune newspaper*. April 24th 2004.

[16.] Squires, G.D., and S. O'Connor, 2001. *Colour and Money: Politics and Prospects for Community Reinvestment in Urban America.* Albany, NY: Suny Press.

[17.] Squires, G.D., 2003. Racial profiling, insurance style: Insurance redlining and the uneven development of metropolitan areas. *Journal of Urban Affairs*, 25(4), pp. 391-410.

[18.] Sweeney, L., 2002. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty and Fuzziness in Knowledge-Based Systems* 10(5): 571-588.

[19.] Tan, P.N. et al, 2006. *An introduction to data mining*, first edition. Pearson Addison Wesley.

[20.] *US Harmonized Tariff Schedules* 2008. http://www.usitc.gov/tata/hts/

[21.] Verykios, V.S. et al, 2004. Association rule hiding. *IEEE Transactions on Knowledge and Data Engineering* 16(4), pp. 434-447.

[22.] Voutilainen, A. 1995. A syntax-based part of speech analyse. *Proc. of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*. Dublin, Ireland, pp. 157-164.

[23.] Wang, K. et al, 2005. Template-based privacy preservation in classification problems. *Proc. The fifth IEEE International conference on Data Mining*. Houston, Texas, USA, pp. 466-473.

[24.] The lexical database WordNet. http://wordnet.princeton.edu/

[25.] Yin, X., and J. Han, 2003. CPAR: Classification based on Predictive Association Rules. *Proc. SIAM International. Conference on Data Mining 2003* San Francisco, USA, pp. 331-335.

[26.] Agrawal, R., Srikant R. 1994. Fast algorithms for mining association rules in large databases. *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB. Santiago, Chile, pp. 487-499.

[27.] Han, J. et al, 2004. Mining frequent patterns without candidate generation. *Proceedings of the Conference of Data Mining and Knowledge Discovery 8*: pp. 53-87.