Attribute-Value Specification in Customs Fraud Detection

A Human-Aided Approach

Norton T. Roman*
FACCAMP
Guatemala 167, Jardim America
13231-230 Campo Limpo Paulista, SP (BRAZIL)

Cristiano D. Ferreira*
University of Campinas
Albert Einstein, 1251
13084-971 Campinas, SP (BRAZIL)

Luis A. A. Meira†
Federal University of Sao Paulo
Talim, 330
12231-280 Sao Jose dos Campos, SP (BRAZIL)

Rodrigo Rezende*
University of Campinas
Albert Einstein, 1251
13084-971 Campinas, SP (BRAZIL)

Luciano A. Digiampietri*
School of Arts, Sciences and Humanities
Arlindo Bettio, 1000
03828-000 São Paulo, SP (BRAZIL)

Jorge Jambeiro Filho‡
Brazil's Federal Revenue
Santos Dummont, Km 66
13055-900 Campinas, SP (BRAZIL)

ABSTRACT

With the growing importance of foreign commerce comes also greater opportunities for fraudulent behaviour. As such, governments must try to detect frauds as soon as they take place, if they are to avoid the profound damage to the society frauds may cause. Although current fraud detection systems can be used on this endeavour with reasonable accuracy, they still suffer with the inconsistencies and ambiguities of unstructured databases, especially in customs. To deal with this kind of problem, we propose a twofold approach: building a brand new structured database, keeping it as clean as possible; and mining the current database for the desired information. Then, as a first contribution, we present a methodology for mining product attribute-value pairs in unstructured text datasets, bringing more structure to the current customs database. Next, as our second contribution, we introduce a system for building a structured database for the Brazilian customs and keeping it with as few redundancies as possible. Both systems aim at building datasets capable of improving the accuracy of fraud detection systems.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; J.1 [**Administrative Data Processing**]: Government

1. INTRODUCTION

Foreign commerce, *i.e.*, the exchange of goods amongst countries, has long played a lead role on the political and economical systems of nations worldwide, with commercial treaties dating back to 509 B.C. [4]. Although important and potentially beneficial to the world economy, with greater commerce comes also greater opportunities for fraudulent behaviour, which can remain concealed within the vast amount of import and export operations. Frauds, defined by the Compact Oxford English Dictionary[1] as "wrongful or criminal deception intended to result in financial or personal gain", can take many forms. Within customs, they may come up as undervaluation of goods, misclassification, underpayment of taxes due [6, 14, 16] and invoice falsification [17], to name but a few.

Even though the mere evasion of taxes may itself represent significant economic consequences, specially to countries whose revenue strongly depends on taxation of imports [6], frauds' damaging effects, and specifically customs frauds, go beyond duty evasion, lying at the heart of many other illicit actions. To start with, customs frauds are responsible for increasing corruption [7], with deep political, social and economic negative effects to the society [6]. Also, some fraudulent operations, such as overvaluations for example, may open up a free channel for money transfer amongst countries. This channel can then be used to conceal money laundering schemes[2] [2], serving to organised crime groups and even to terrorist financiers [17]. Last, but not least, the use of false invoices and other trade manipulations may disguise serious smuggling and drug trafficking operations [5].

Given the considerable hazard that customs frauds represent, it becomes paramount for governments to try to avoid them by taking preventive measures. Notwithstanding, even known prevention may be the best way to tackle this problem, the fact that fraudsters are adaptive and the advent of new technologies that, unfortunately, may be used to this end too [2], makes it hard to successfully stop frauds from taking place. Hence, along with preventive procedures, governments should also apply some fraud detection mechanism, *i.e.*, a way to identify frauds as soon as possible once their prevention has failed [2].

The problem with this approach, however, is that:

1. There is a great deal of data to be analysed, which rules

[1]{nortontr,crferreira,rcrezende, luciano.digiampietri}@gmail.com
[2]augusto.meira@unifesp.br
[3]jorge.filho@jambeiro.com.br

[1]http://www.askoxford.com/dictionaries/?view=uk
[2]The so called trade-based money laundering.

out having them inspected by hand. Also, the usually high number of attributes that must be accounted for reduces the performance of automatic systems [15];

2. From this amount of data, only a very small portion represents actual frauds, rendering automatic fraud detection models based on machine learning less accurate [2, 15];

3. It is hard for automatic techniques to reflect the entire knowledge of domain experts. On this account, some existing systems (*e.g.* [8, 14, 16]) do actually try to codify some of this expertise in a computational way, although they fail in capturing the experts' intuition in depth;

4. Current product classification systems, such as the Harmonized System of Nomenclature (HSN)[3] or the Mercosul Common Nomenclature (NCM)[4], are rather subjective and ambiguous, making it hard to assign category codes to products [5]; and

5. Usually, customs import forms do not care about the detailed description of goods, recording manufacturers, exporters, importers, countries, product category etc, but leaving the product description as a free text field. As such, it requires a lot of effort to mine specific attributes (like the resolution of digital cameras, for instance), in order to compare different products.

In this paper we describe some strategies adopted to tackle the aforementioned problems within the HARPIA[5] project. Idealised as a partnership between some Brazilian universities and the Brazilian Federal Revenue, the HARPIA project aims at the use of artificial intelligence for customs fraud detection. In its current version, depicted in Figure 1, the project comprises, amongst other things, the following modules:

- A catalogue of products, described in Section 3.2;
- A registry of foreign exporters [5];
- A tool for the identification of suspicious operations before customs clearance (Carancho [14] in the figure);
- A system for identification of such operations after clearance (ANACOM), still under development;
- A module for controlling small imports made through the express mailing service (under development); and
- A system for making structured queries to the current customs database, described in Section 3.1.

Instead of presenting the existing mechanisms and those that are currently being developed for fraud detection, this work focuses on the steps necessary to make them more effective. More specifically, we address the issues of (i) what to do with current databases, as described on Item 5 above, *i.e.* how to deal with an imports/exports database where product attributes must be mined from unstructured text containing product descriptions; and (ii) how to improve this database, in order to reduce its subjectivity and ambiguity, by dynamically building a fully structured database with product attributes. Hence, within Figure 1, we cover both Attribute-Value Extractor and Catalogue systems.

The rest of this paper is organised as follows. Section 2 presents an overview on the state of the art of current fraud detection technologies, along with their advantages and drawbacks. It also points out how the measures described in this paper could improve the use of such techniques. Next, on Section 3, we present our strategies for dealing with the current imports database and how to improve it to get better results from automatic fraud detection systems. Finally, Section 4 presents a conclusion to this work, alongside the identified venues for future research.

2. RELATED WORK

Given the costs involved and the workload necessary to detect frauds with normal audit procedures [2, 5], current approaches to this problem usually rely either on machine learning techniques or on statistical analysis of vast amounts of data. Amongst the main techniques, we identify the use of Neural Networks [12, 13], Bayesian Belief Networks [12, 13], Decision Trees [12, 15], Rule-based Systems [16], and Statistical Outlier Detection [10, 14, 18]. Overall, these approaches can be further classified as belonging to one of two categories. The first one, comprising the supervised methods for fraud detection, consists of feeding some learning algorithm both with samples of clear and fraudulent data [2], so it can automatically detect some sort of pattern. This pattern can then be used to classify new coming records.

The main advantage of such methods lies on the fact that, once trained in a usually small set of labelled data, they can be used in a fully automatic fashion to classify a much bigger set of unlabelled data (which may correspond, for example, to a stream of import operations that are analysed as soon as registered by the importers). Also, it is possible for this kind of algorithm to detect some pattern that passed unnoticed by the customs officers, thereby detecting fraudulent operations that otherwise would be wrongfully cleared. On the other hand, the use of labelled data requires (a) one to be sure about the true classification of each training record [2], under the risk of misleading the learning algorithm; and (b) the existence of sample data on both clear and fraudulent sets, something that may turn out not so easy to obtain [11]. Moreover, since the method relies on types of fraud already known by the authorities [2], it may fail to detect newly created strategies for deceiving customs.

The second category, namely, the unsupervised approach, corresponds to those methods capable of running without being previously trained on a set of labelled data. In general, unsupervised methods either seek out the dataset, looking for records that somehow lie too far from some statistical model [2], or make use of a set of predefined rules (set up by domain experts) to classify each operation. As such, besides working around the main drawbacks of supervised techniques and saving resources that otherwise would be needed

[3]World Customs Organization. http://www.wcoomd.ora/
[4]Mercosul/Mercosur – Southern Common Market. http://www.mercosur.int/msweb/
[5]Risk Analysis and Applied Artificial Intelligence.

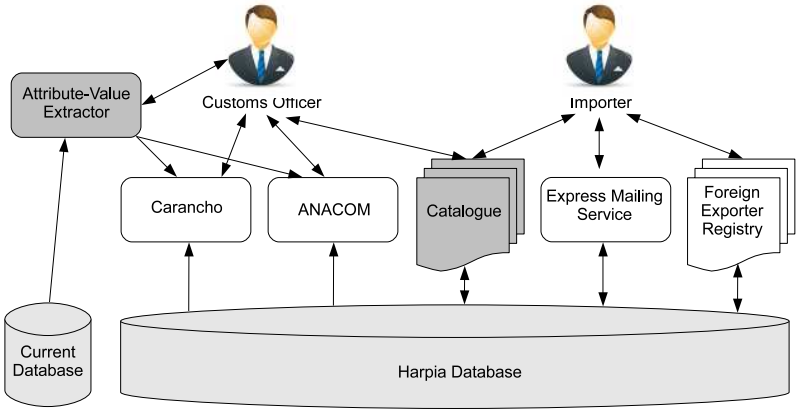

Figure 1: Current modules of the Harpia Project.

for labelling the training data, methods under this approach also have the advantage of (i) partially accounting for the expertise developed by customs officers, in the case of rule-based systems and some outlier detection approaches; and (ii) being able to adapt to changes in the importers' behaviour, in the case of statistics-based systems [5].

Notwithstanding, rule-based systems demand constant upkeep, in order to keep rules updated to reflect new fraudulent behaviours [5]. Outlier detection systems, in turn, require the majority of data to be clear, so that frauds can be seen as outliers (this, however, is not so hard a requirement, given that clear operations are actually the vast majority [2, 15] within customs). Also, it is worth mentioning that outlier detection techniques, although robust, cannot faithfully identify frauds but, instead, only suspicious operations [2], *i.e.*, operations that considerably astray from the rest.

Whatever the chosen method, it is always necessary to compare some specific operation against a set of related operations. To this end, one must have at hand some description for them, in the form of a set of attribute-value pairs, for example. Such a requirement, however, cannot be easily accomplished due to the lack of standards for product attributes, in both customs and the electronic commerce world [1]. Also, and specially within customs, product descriptions usually take the form of a free text typed in by the importer. This further complicate matters since, in order to get to the values for some product's attributes, one has to rely on existing text mining tools (such as the ones described in [3] and [8]). The problem with such systems is that, although useful, they usually depend on a training set of labelled data (*e.g.* [3, 8]), which reduces their adaptability to new circumstances[6].

In the next section, we present our approach to the problem of identifying attribute-value pairs. Our ultimate goal is twofold: (i) to deal with existing imports/exports databases, so they can better suit current fraud detection techniques; and (ii) to build a structured database for imported/exported goods, in order to provide automatic systems with a clearer way to compare different products.

[6]In the case of [8], however, this problem is reduced by having the user correct the results and provide new training data. Still, there is no apparent way the user can explicitly input rules or patterns to the system.

3. CLEARING THE PATH FOR FRAUD DETECTION

Given the importance of the current database for customs, and the need to improve it with more structured information, we decided to split up the problem into two separate tasks. At first, we developed a system for extracting attribute-value pairs from free-text descriptions of products (Section 3.1). Then, we developed a full-fledged catalogue to record all import/export operations through customs. As it will be seen in Section 3.2, this catalogue is responsible not only for keeping the database clear (*i.e.* with minimum redundancy), but also for providing it some structure, so as to allow for a more efficient use of current fraud detection techniques. Figure 2 sketches out this approach.

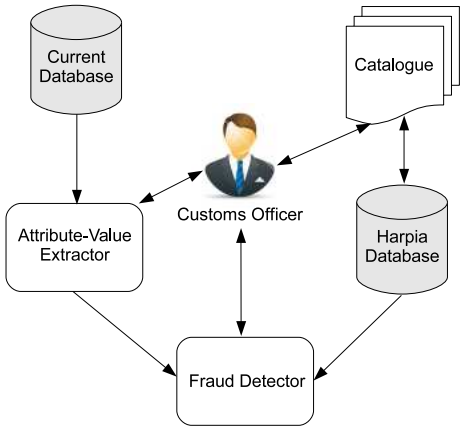

Figure 2: Approaches for dealing with current and the Harpia databases.

3.1 Human-Aided Attribute Value Detection

Product classification, entity resolution, undervaluation detection, and other data analysis tasks employed in the HARPIA project depend on the products' attributes. However, they are not always available. Even when they are present, they can be spread over a plain text description. To minimise

these problems, we propose a methodology for mining attributes and their respective values from text descriptions. Our approach also allows for the creation of new attributes from those that were already detected. For instance, one could extract the value of *width* and *length* from some paper sheets description and then use this information to make up a third attribute – *area* – as the product of them. Thus, as it will be seen in what follows, the main difference between our system and current systems for mining attribute values in text (*e.g.* [3, 8, 16]) lies on the fact that we do not try to automatically figure out rules for attribute value selection but, instead, we leave it to the customs officer (*i.e.*, the user) to codify all his/her expertise into the system.

To do so, we build up attribute-value pairs by applying a sequence of *Interpretation Rules*. An Interpretation Rule defines how to get, either from existing attributes or from raw text, the information needed to define the value for some other attribute. Since different rules may result in different values, the user can assign a priority level to each of them, so as to have the system apply the interpretations in the order s/he thinks would better suit the target attribute. The system then constructs the pair with the value returned by the first successful rule.

Figure 3 illustrates the definition of an attribute as well as the set of its constituent interpretation rules. In this figure, the customs officer determines whether the attribute will hold a string or some numerical value. Along with the attribute name and description, the officer can list the *Interpretation Rules* responsible for giving the attribute its value. Section 3.1.1 describes in details the different kinds of Interpretation Rules available in this system.

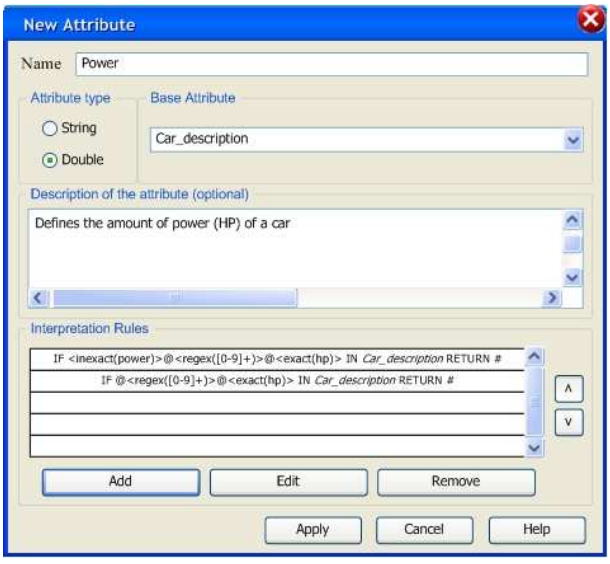

Figure 3: An attribute and its interpretation rules.

3.1.1 Interpretation Rules

There are three kinds of Interpretation Rules, to wit, *Algebraic Expressions*, *Conditional Rules*, and *Text Search*. Along with their return values, each rule is also allowed to output a special value – UNDEFINED, which means the desired value could not be determined by applying the rule to some specific input. Next, we describe each type of rule, its purpose and possible outputs.

Algebraic Expressions. This Interpretation Rule is defined by an algebraic expression and the set of attributes that composes it. As such, it can only be used over numerical values, outputting the (numerical) result of applying the expression to the input values. Supported operations are[7]:

- Addition, where the values of two attributes are added to form the value for a third one;
- Subtraction, where the value of some attribute is subtracted from the value of another;
- Multiplication, where the values of two attributes are multiplied;
- Division, where the value of an attribute is divided by the value of another; and
- Exponentiation, where the value of some attribute is raised to the power defined by the value of another.

To illustrate the use of Interpretation Rules, consider the attributes, *width* and *length*. Now, suppose one wishes to infer an attribute *diagonal*, corresponding to $\sqrt{width^2 + length^2}$. This can be achieved through the Algebraic Expression:

$$diagonal = pow(pow(width, 2) + pow(length, 2), 1/2)$$

Conditional Rules. Similar to [16], we define a Conditional Rule as a *boolean expression* and a return value, with the following structure:

IF *boolean expression* THEN RETURN *return value*

Within this rule, the *boolean expression* admits the logical operators *OR*, *AND*, and *NOT*, along with the comparison operators $=$, $\neq$, $>$, $<$, $\leq$ and $\geq$, which should be applied over the input attributes. The *return value*, in turn, is the value that will be returned should the *boolean expression* hold true. If, however, this expression is not satisfied, UNDEFINED is returned.

To exemplify the use of Conditional Rules, consider the attributes *Manufacturer* and *Horsepower* of some car. Now, suppose one knows that all cars by manufacturer Y that have from 70 to 90 horsepower are of the type X. Also suppose these two last values are kept in the attributes *Manufacturer* and *Horsepower*, respectively. Thus, based on this information, the following Conditional Rule could be formulated:

IF *Manufacturer* = Y AND *Horsepower* $>$ 70 AND *Horsepower* $<$ 90 THEN RETURN X

[7]Although the operations are primarily defined over attribute values, they can as well be used on numerical constants.

i.e., if the system identifies the manufacturer of some car as Y and its horsepower as lying between 70 and 90, then the target attribute will get X as its value.

Text Search. The goal of Text Search is to find a *search pattern* in some text attribute. There are two variations:

1. IF FIND *search pattern* IN *text attribute* THEN RETURN *return value*
2. IF FIND *search pattern* IN *text attribute* THEN REMOVE

The first variation defines that if the *search pattern* matches some substring in the *text attribute*, then *return value* is returned. The second variation, in turn, works like a filter. All the occurrences of *search pattern* are removed from the attribute. Since the rules with lower priority values are run on the filtered version of the attribute, this variation has as its main consequence the removal of the *search pattern* from further consideration.

As it can be seen in its definition, this kind of Interpretation Rule is composed by three elements. The first one, *text attribute*, corresponds to the attribute in which the *search pattern* will be run. *Search pattern*, in turn, is split up in three other patterns, to wit, the *head*, the *body*, and the *tail* patterns, with *head* and *tail* patterns comprising matches that must occur before and after the *body*, respectively (see Figure 4). Each of these patterns may correspond to one of the following types:

- *Exact pattern*, which looks for substrings that perfectly match the pattern;
- *Regular expression*, representing a plain regular expression; and
- *Inexact pattern*, looking for substrings that may or may not perfectly match the pattern (in the case of a partial match, the system calculates the probability of the substring corresponding to a misspelled form of the pattern).

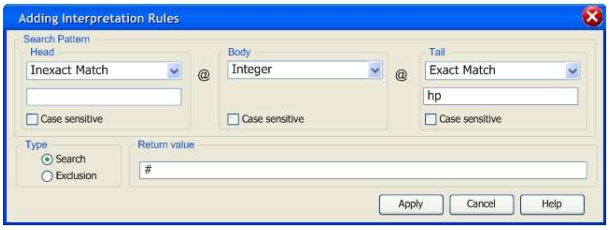

Figure 4: Defining a Text Search Rule.

Finally, the *return value* is the value (string type only) that will be returned in the case of a match with the *search pattern*. As an example, consider the following *search pattern* (notice that both head and tail patterns can be empty):

- *head pattern*: empty
- *body*: Regular expression $\rightarrow [0-9]+$
- *tail pattern*: Exact pattern $\rightarrow$ *hp*

and the return value #, which means that the rule should return a verbatim copy of the substring that matched the *body*. Consider also a text attribute with value "Car model X with 98 hp,...". By running the above rule on this text, one would have the *body* match "98" and the *tail pattern* match "hp". As a consequence, the whole *search pattern* would match positively to "98 hp" and, since the *return value* = #, 98 is returned by the rule.

3.2 Towards a Cleaner Database

The Brazilian Federal Revenue's current protocol for foreign commerce need the local importer to fill in, amongst others, two form fields when importing some product, to wit, its NCM (Mercosul Common Nomenclature) code and its description, produced as a free text. An NCM code consists of a numerical value inside a table with approximately 10,000 elements, each of them corresponding to a specific class of products. A product's free description, in turn, is a way for the importer to characterise this product by presenting its specifications as a text. Choosing specific NCM codes is a primary factor for determining the taxes due. As such, the Brazilian Federal Revenue (*BFR* henceforth) has at high priority to inspect whether the product was correctly classified.

One major limitation of the current protocol lies on the fact that there is no sole identifier for products. There most certainly are hundreds of thousands of distinct products that must be classified within little more than 10 thousand codes at the NCM table. Hence, a great number of products are bound to get the same code when they are classified, even though they are not possibly related. This lack of precision, along with existing ambiguities and subjectivities at the NCM table, are factors that bring complexity to inspection by the BFR officers. Another limitation resides on the unstructured text used to describe the products. In this text, importers might describe the same product, say a car, in a different manner at each new import operation. As an example, consider two possible descriptions for the same product:

NCM	*Free Description*
87032310	Gol G4 1.6 75Hp
87032310	Gol Generation IV HP=75Hp 1,599cc

with both texts relating to the NCM 87032310, *i.e.*, "cars with an internal combustion engine, $1,500 < cm^3 \leq 3,000$, up to 6 passengers".

As it can be seen, a product whose description takes the form of a text produced by its importer does not allow, in principle, the automatic identification of NCM misclassification. Within the current system, auditing is carried out manually by some BFR officer. Besides, identified misclassifications cannot be extended to similar products, given the lack of unique ways to identify them. Notice that it is not a trivial task to automatically decide that "Gol G4 1.6 75Hp" and "Gol Generation IV HP=75Hp 1,599cc" refer to the same product.

Consider, on the other hand, a scenario where we have some structured description for this same product. Since the attribute "displacement" in a car is a determinant factor for its correct classification within the NCM, any inconsistency, such as a 5,000cc vehicle classified as NCM=87032310 for example, would be easily detected by automatic means. The system could even prevent the user from misclassifying some product. This structuring, besides aiding on the correct classification of products, allows for fewer ambiguities in the data analysis task by experts. Instances of such ambiguities are "3M Silver Tape", "Peugeot 206" and "HP Printer", which could taken, respectively, for some measurement ("3m long"), some integer value, as opposed to the model of a car, and an indication that the importer was describing the horsepower of some car, instead of the printer's manufacturer.

These examples, however, present only well known products and values. In a more realistic set-up, texts produced by importers/exporters may contain less obvious ambiguities. On the other hand, within structured systems many ambiguities disappear, as shown in the table below:

Company	*Model*	*Product*	*Length*
3M		Tape	
		Tape	3m
Peugeot	206		
HP		Printer	

To comply with these questions, we are currently developing, as a part of the HARPIA project, a Catalogue of Products. This catalogue aims at uniquely identifying products traded across the Brazilian borders. To do so, the system presents the user with a search for similar products, along with a spelling correction module, so as to help preventing the user from mistyping some information. Besides, the system keeps all attribute information in a structured way. Each local importer/exporter is responsible for keeping his/her own list or products, without being able to access other local databases. In what follows, we will present an example of this catalogue on the run.

Every time a local importer/exporter wishes to trade a new product s/he must register it in the catalogue, which will then assign a unique identifier to it. If, on the other hand, the user is about to trade an already registered product, s/he can perform a search within the system, in order to get this product's identifier.

Figure 5 shows the catalogue's fluxogram. Within the system, the user starts with a free text search, in an attempt to find similar products. This search is processed by a search engine based on the Lucene [9] open source library, kept by Apache Software Foundation[8]. After this search, the importer/exporter can then retrieve the identifier for registered products and use them in his/her commercial transactions.

If, however, it turns out that the system finds no match to the user's search (as with some new product), the user must provide it with the product's NCM code. Next, the system retrieves from its database a template corresponding to products under this NCM. This template, which must be

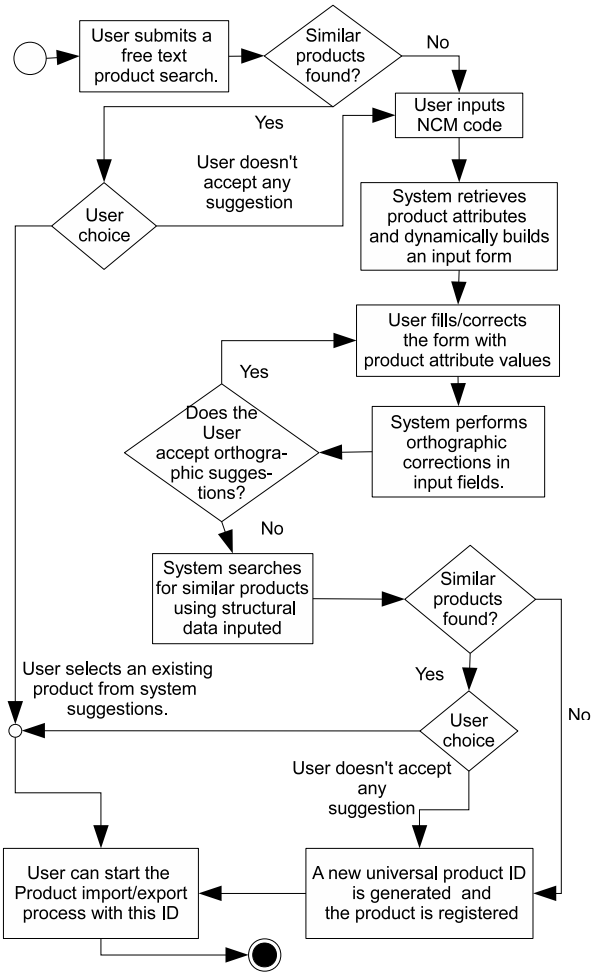

Figure 5: Catalogue's usage.

filled up by the user, was constructed following guidelines provided by customs officers and private consultants. By using pre-defined templates the amount of inconsistencies in the database was reduced. After the submission of the information needed, the system runs a spelling checking on the template fields.

For the spelling checking task, we have implemented a module that calculates the chance every word in a description has of being new or just a variation of an already known word. To make this decision, the system uses a Bayesian probabilistic method based on N-grams and a word frequency model. Such a model is sensitive to the context of the importer/exporter, since it keeps for each of them a table of known words and frequencies.

After the user has filled the template and the system has finished running the spelling checking, it carries out a new search for similar products. This time, however, its search engine takes into account the fact that both the database and the newly produced template are presented in a structured way. If the engine happens not to find similar products, or if it finds some related ones but the user insists that the product s/he is trying to register is not one of the sug-

[8]http://lucene.apache.org

gested ones, then the system updates the database with this new product, presenting the user with a unique identifier for it. This identifier can then be used on future transactions by this importer/exporter.

Figure 6 illustrates the organisation of the products database. Each importer/exporter has a separate dataset containing the products s/he has already registered. All the datasets together build up the Catalogue of Products, along with a module for distributing the search and correction tasks amongst a number of different computers, and a common database, responsible for aiding in the distribution of the catalogue. Sets of known words and frequencies are separated according to the importer/exporter, being kept within their own database, and thereby creating an adequate context for orthographical suggestions, based on the behaviour and particularities of each importer/exporter.

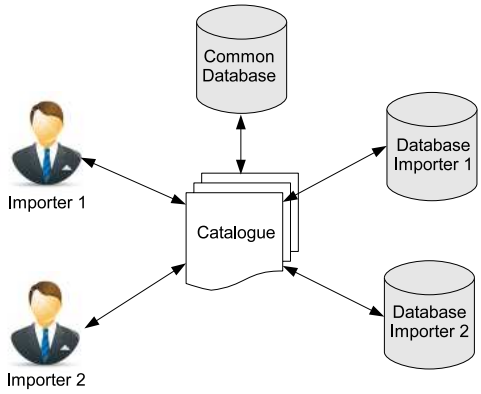

Figure 6: Database separation amongst importers.

As such, the Catalogue of Products presents the advantage of structuring product descriptions as attributes, allowing for (i) a more accurate use of data analysis tools, (ii) a cleaner dataset, *i.e.* with little redundancies, achieved with the aid of the search and spelling correction modules; and (iii) the creation of unique identifiers for products, further increasing the precision of data analyses.

4. CONCLUSION AND FUTURE WORK

To improve the performance of current fraud detection techniques it becomes necessary to increase the quality of the datasets used by such systems. Nevertheless, existing unstructured databases, specially in customs, cannot be neglected, since they are responsible for keeping all import/export records so far. To work around theses problems, a twofold approach is proposed, in order to both deal with the existing datasets and start building a brand new one, designed to better suit the customs' needs.

In this paper we described two systems designed to handle both the old and new databases within the Brazilian customs. Our main contributions to the field are (i) the description of a methodology for mining attribute-value pairs from text descriptions of products, in order to bring more structure to the current customs database; and (ii) the introduction of a system developed for building a structured database for products and keeping it as clean as possible (*i.e.*, with minimum redundancy). Both new database and attribute mining system, by allowing a better comparison of different products through their attributes, are intended to be directly queried by fraud detection systems in the near future.

As for future work, we are currently implementing some fraud detection techniques to query the datasets produced by the systems described in this paper. Next, we intend to design a report-making system capable of producing human-readable reports. These reports will then be used as evidence for fraud in trails. As such, our long-term goal is not only to help customs officers to detect frauds, but also to provide them with the means to build up their case before the court.

Regarding the Catalogue of Products, we are planning to build a suite in order to bring together all the tools developed so far, to wit, the Catalogue of Products, the Foreign Exporter Registry, the Carancho module and the Attribute-Value Extractor. Such a tool would present the user with statistical and graphical resources, along with configurable rules, capable of improving the inspection task in a semi-automatic way.

5. ADDITIONAL AUTHORS

Additional authors:
Andreia A. Kondo (Institute of Computing, UNICAMP, email: `andreia.kondo@gmail.com`),
Bruno C. Brandão (Institute of Computing, UNICAMP, email: `brunocedraz@gmail.com`),
Everton R. Constantino (Institute of Computing, UNICAMP, email: `constantino.everton@gmail.com`),
Helder S. Ribeiro (Institute of Computing, UNICAMP, email: `helder@gmail.com`),
Jacques Wainer (Institute of Computing, UNICAMP, email: `wainer@ic.unicamp.br`) and
Siome K. Goldenstein (Institute of Computing, UNICAMP, email: `siome@ic.unicamp.br`) .

6. REFERENCES

[1] S. Bergamaschi, F. Guerra, and M. Vincini. A data integration framework for e-commerce product classification. In *Proceedings of the First International Semantic Web Conference on The Semantic Web (ISWC '02)*, pages 379–393, London, UK, 2002.

[2] R. J. Bolton and D. J. Hand. Statistical fraud detection: A review. *Statistical Science*, 17(3):235–255, 2002.

[3] V. Borkar, K. Deshmukh, and S. Sarawagi. Automatic segmentation of text into structured records. *ACM SIGMOD Record*, 30(2):175–186, 2001.

[4] E. W. Bowen. Roman commerce in the early empire. *The Classical Weekly*, 21(26):201–206, May 14 1928.

[5] L. A. Digiampietri, N. T. Roman, L. A. A. Meira, J. J. Filho, C. D. Ferreira, A. A. Kondo, R. C. Rezende, E. R. Constantino, B. C. B. ao, H. S. Ribeiro, P. K. Carolino, A. Lanna, J. Wainer, and S. K. Goldenstein. Uses of artificial intelligence in the brazilian customs fraud detection system. In *Proceedings of the 9th Annual International Conference on Digital Government Research (dg.o 2008)*, Montréal, Canada, May, 18–21 2008.

[6] A. Doig and S. Riley. *Corruption and Integrity*

Improvement Initiatives in Developing Countries, chapter Corruption and Anti-Corruption Strategies: Issues and Case Studies from Developing Countries. UNDP, New York, 1998.

[7] K. M. Dye. *Performance Accountability and Combating Corruption*, chapter Corruption and Fraud Detection by Supreme Audit Institutions, pages 299–317. The World Bank, Washington, D.C., 2007.

[8] R. Ghani, K. Probst, Y. Liu, M. Krema, and A. Fano. Text mining for product attribute extraction. *ACM SIGKDD Explorations Newsletter*, 8(1):41–48, 2006.

[9] E. Hatcher and O. Gospodnetic. *Lucene in Action*. Manning Publications, December 2004.

[10] V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[11] M. Jans, N. Lybaert, and K. Vanhoof. Data mining for fraud detection: Toward an improvement on internal control systems? In *Proceedings of the 30th Annual Congress European Accounting Association (EAA2007).*, Lisbon, Portugal, 2007.

[12] E. Kirkos, C. Spathis, and Y. Manolopoulos. Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications: An International Journal*, 32(4):995–1003, 2007.

[13] S. Maes, K. Tuyls, B. Vanschoenwinkel, and B. Manderick. Credit card fraud detection using bayesian and neural networks. In *Proceedings of the 1st International NAISO Congress on Neuro Fuzzy Technologies*, Havana, Cuba, 2002.

[14] N. T. Roman, E. R. Constantino, H. Ribeiro, J. J. Filho, A. Lanna, S. K. Goldenstein, and J. Wainer. Carancho – a decision support system for customs. In *Proceedings of ECML PKDD 2006 Workshop on Practical Data Mining: Applications, Experiences and Challenges*, pages 100–103, Berlin, Germany, September 22nd 2006.

[15] H. Shao, H. Zhao, and G.-R. Chang. Applying data mining to detect fraud behavior in customs declaration. In *Proceedings of the First International Conference on Machine Learning and Cybernetics*, volume 3, pages 1241–1244, Beijing, China, November, 4-5 2002.

[16] A. K. Singh, R. Sahu, and K. Ujjwal. Decision support system in customs assessment to detect valuation frauds. In *Proceedings of the 2003 IEEE International Engineering Management Conference (IEMC-2003): Managing Technologically Driven Organizations: The Human Side of Innovation and Change*, pages 546–550, Albany, USA, November, 2-4 2003. IEEE.

[17] U.S. Department of State. *International Narcotics Control Strategy Report. Volume II: Money Laundering and Financial Crimes*, 2005.

[18] K. Yamanishi, J. Takeuchi, G. Williams, and P. Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery*, 8(3):275–300, 2004.