

CUSTOMS TARIFF CODE CLASSIFICATION

This application claims priority benefit of Singapore Patent Application No. 10201406401W, filed on 8 October, 2014.

FIELD OF THE INVENTION

[0001] This invention relates to improved methods of classifying goods, and more particularly to improved customs tariff code classification systems.

BACKGROUND OF THE INVENTION

[0002] The Harmonized Commodity Description and Coding System, also known as the Harmonized System (HS) of tariff nomenclature, is an internationally standardized system of names and numbers used to classify goods. HS codes are used in essentially all international trade. HS classification codes are used in many ways, including setting custom tariffs, collection of trade statistics, determination of rule of origin, and monitoring of controlled goods, for example. The HS codes are organized into a series of sections. HS Codes are first assigned to sections, and then to a specific chapter, heading, and subheading such as a national code, in that order, as necessary.

[0003] The HS nomenclature assigns up to a total of 12 digits at the tariff-rate (legal) level. To ensure harmonization, the first 6 digits of the HS are used universally. Each country may then append additional 2 to 6 digits to the original 6 to suit its own

1 tariff and statistical needs. Singapore, for example, uses the ASEAN Harmonized
2 Tariff Nomenclature (AHTN) HS code, which is 8 digits, and this is common across
3 ASEAN countries.

4
5 [0004] HS classification is a difficult process as a sizeable percentage of goods are
6 difficult to classify, especially low volume goods. Assignment of HS codes for a
7 given good or item has traditionally been done by a domain expert. Some efforts
8 have been made to attempt to automate this classification process. For example,
9 one approach for automated HS classification uses a centralized classification
10 database created by a crowdsourcing technique. Classification information is
11 created by a crowd (that is, a collection of users/subscribers) and typically comprises
12 historical declaration data corresponding to known items, and such data is shared
13 among other subscribers of a classification database. However, the quality of the
14 classification information created by crowd can be less than desired. Verification of
15 classification database information relies on the domain experts who simply cannot
16 practically validate all of the thousands of classification records created. Further,
17 retrieval of the classification information is still difficult, and users generally need to
18 make a significant effort to read through several past records suggested by the
19 system to classify goods.

20
21 [0005] Another attempt to automate HS classification involves creation of
22 classification rules. The domain expert defines a set of sufficient conditions for a
23 document or description to be labelled with a given category. This approach,
24 however, is labor intensive and requires the hiring of many experts.

1 [0006] Generally classification data generated by a crowd may contain errors and
2 simply imprecise rule descriptions. Classification rules created by experts are more
3 accurate but generally expensive and hard to achieve good coverage. Thus, there is
4 a need for a system which can automatically learn the knowledge from largely
5 available historical declaration data and yet tolerate inherent imprecision of crowd
6 based data while also taking advantage of HS codes and the related rules created by
7 experts.

8 9 SUMMARY OF THE INVENTION

10
11 [0007] In accordance with a first aspect, a method for customs code tariff
12 classification of items using harmonized system (HS) classification codes comprises
13 receiving a query about an item, using an ad hoc rule, wherein if the ad hoc rule is
14 satisfied, then an HS code is assigned to that item, and using a machine learning
15 rule, and if the machine learning rule is satisfied, then an HS code is assigned to that
16 item.

17
18 [0008] From the foregoing disclosure and the following more detailed description of
19 various embodiments it will be apparent to those skilled in the art that the present
20 invention provides a significant advance in the technology of tariff code classification
21 systems. Particularly significant in this regard is the potential the invention affords for
22 providing an elegant tariff code classification system adapted to handle the wide
23 range of goods seen at ports of entry. Additional features and advantages of various
24 embodiments will be better understood in view of the detailed description provided
25 below.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] Fig. 1 shows a schematic block diagram of a system for classification of tariffs using the harmonized system (HS) code for goods.

[0010] Fig. 2 shows a flow chart of a rule processing module of the system of Fig. 1.

[0011] Fig. 3 shows an example of a screen shot of a rule file.

[0012] Fig. 4 shows a schematic flow chart of a training module.

[0013] Fig. 5 shows an example of a screen shot of a training file, part of the system where keywords associated with particular classification codes may be associated.

[0014] Fig. 6 shows a flow chart of a training process of a "TRUE" algorithm for associating historically used keywords with particular classification codes.

[0015] Fig. 7 shows a flow chart of a preprocessing module.

[0016] Fig. 8 shows a representative flow chart of a training process for a power mean support vector machine (PMSVM) algorithm.

[0017] Fig. 9 shows a high level view of a process for classification of an item using a classification module.

[0018] Fig. 10 shows an expanded flow chart for the process for classification of an item of Fig. 9.

[0019] Fig. 11 shows an ad hoc subroutine or module for the classification of an item according to the process of Fig. 9.

[0020] Fig. 12 shows an algorithm or subroutine for associating historically used keywords for the classification of an item according to the process of Fig. 9.

[0021] Fig. 13 shows a power mean support vector machine subroutine or module for the classification of an item according to the process of Fig. 9.

[0022] Fig. 14 shows a high level schematic flow chart of a process to set up a classification service in one country.

[0023] It should be understood that the appended drawings are not necessarily to scale, presenting a somewhat simplified representation of various features illustrative of the basic principles of the invention. The specific design features of the customs tariff code classification system as disclosed here, including, for example, the specific number of display screens and the data presented at each screen, will be determined in part by the particular intended application and use environment. Certain features of the illustrated embodiments have been enlarged or distorted relative to others to help provide clear understanding. In particular, thin features may be thickened, for example, for clarity of illustration. All references to direction and

position, unless otherwise indicated, refer to the orientation illustrated in the drawings.

DETAILED DESCRIPTION OF CERTAIN EMBODIMENTS

[0024] It will be apparent to those skilled in the art, that is, to those who have knowledge or experience in this area of technology, that many uses and design variations are possible for the customs tariff code classification system disclosed here. The following detailed discussion of various alternate features and embodiments will illustrate the general principles of the invention with reference to a customs tariff code classification system suitable for use at a port of entry. Other embodiments suitable for other applications will be apparent to those skilled in the art given the benefit of this disclosure.

[0025] The present invention generally relates to a method and a system for automatic assignment or recommendation of HS codes by both ruled based and machine learning algorithms. Fig. 1 shows a schematic block diagram of a system for classification of tariffs using the harmonized system (HS) code for goods mentioned above. A model repository 101 is generated by a training module and stores machine learning rules generate by training , and is accessible by a classification module 109, a rule repository 102 can be access a rule repository module and can store ad hoc rules, a feature vector database 103 for storing feature vectors generated by the training module from historical data, an HS nomenclature and tariff database 104, an application server 106 containing the training module 107, rule processing module 108 and the classification module 109 (which receives

1 queries), a webserver 110 and a client computer 111 with a user interface 112, and
2 an administrative interface 113 are shown. The components can comprise several
3 computers including a cloud computer which may contain the algorithms and the
4 database of HS nomenclature and tariffs, for example, and are connected through a
5 network 114.

6
7 [0026] The method for customs code tariff classification and recommendation of
8 items using harmonized system (HS) classification codes comprises receiving a
9 query about an item. Typically the query is input by a user and may have a wide
10 range of accuracy and terminology. Next, using an ad hoc rule discussed in greater
11 detail below, assigning or recommending an HS code to that item if the ad hoc rule is
12 satisfied. In accordance with a highly advantageous element, a machine learning
13 rule may also be used. If the machine learning rule is satisfied, then an HS code is
14 assigned to that item. The machine learning rule may be generated by applying an
15 algorithm to a set of historical data. Generally user input words from queries which
16 can be converted to tokens and in turn to feature vectors comprising one of more
17 features. Historical data can also be converted into feature vectors.

18
19 [0027] The query may be compared with rule descriptions corresponding to the HS
20 codes using the ad hoc rule, and the query may be compared with a feature vector
21 generated from historical data using the machine learning rule. The ad hoc rule may
22 be used first, and if the query does not satisfy the ad hoc rule, then the machine
23 learning rule may be used.

[0028] The expert-knowledge for assigning the HS classification can be acquired by processing of the rule file created by experts. The rule processing module can work with different types of rules. Codes, key words, and free text are allowed to be inputted as the query as long as the rules of the system allow such entries. Each rule can then be mapped with different levels of HS code such as 2 digits (chapter level), 4 digits (heading level), 6 digits (sub-heading level) and full digit (national level) HS code. Since HS codes comprise one or more levels, the HS code assigned or recommended by the system can vary depending on the level of specificity of the query. Expert-knowledge acquired through processing of rule file can be used to predict the HS code from the given product description.

[0029] Use of the machine learning algorithm advantageously helps obtain classification knowledge from the crowd by use of historical declaration data. The knowledge extraction process can be further enhanced by using two separate algorithms. Querys can be accurately classified by using strong key words which are highly associated with one particular HS code. The machine learning rule may be generated by more than one algorithm. One algorithm may be a "TRUE" algorithm which detects strong keywords from the training or historical dataset. The True Algorithm assumes that if a single word or a multi-word phrase appears only in one HS code (or mostly appears in that HS code), the algorithm takes it as a rule that that phrase is associated with a particular item. By using this rule the algorithm can classify the correct HS code with very high prediction accuracy. However, some of the querys may be complex and not contain strong keywords. Therefore, a second algorithm may be used, such as a Power mean SVM algorithm to handle complex querys and classification errors in the training dataset.

[0030] In machine learning, support vector machines or networks (SVMs) are supervised learning models with associated learning algorithms that analyze data and recognize patterns. SVM is used for classification and regression analysis. It has been found that power mean SVM, in terms of classification accuracy, is preferable as it outperforms state-of-the-art additive kernel implementations, and has major advantages over linear SVM. A more detailed description of a suitable representative power mean support vector machine algorithm may be found in *Power Mean SVM for Large Scale Visual Classification* (2012) by Jianxin Wu, presented at The IEEE Int'l Conference on Computer Vision and Pattern Recognition, the disclosure of which is herein incorporated by reference in its entirety. Different models which represent different levels of HS code can be trained by use of the PmSVM algorithm. Optionally a plurality of trained models can be used to predict HS code (which can be different levels) from the given query.

[0031] Fig. 2 shows a flow chart of the rule processing module 108 of the system of Fig. 1. The process, which an ad hoc process making use of an ad hoc rule has a rule file 201 uploaded through the administrative interface 113 from a client computer 111. Fig. 3 shows an exemplary format of a display of one of the computers of the system disclosed herein, and some examples of the rule file 201 which can contain HC code 301, description 302, type 303, minimum similarity 304 (expressed on a 0 to 1 scale, with no similarity at 0 and a perfect match at 1), and remarks 305 columns. Independent minimum similarities for the rule descriptions the queries can be set for each rule. Each row in the rule file 201 represents a rule which can be used to predict the HS code given in respective HS code columns 301 of each row.

1 Different level of HS code such as 2 digits, 4 digits, 6 digits or full length HS code
2 can be placed in the HS code columns 301 based on specificity of the description
3 given in the description 302 column. Three types of rule which can be specified in
4 the type column 303 are allowed in the rule file 201. The types include code which
5 can be used to detect code based description such as CAS number or CA product
6 code, key words and free text.

7
8 [0032] A method for customs code tariff classification of items using harmonized
9 system (HS) classification codes comprises receiving a query for an item, generating
10 a first similarity score by comparing the query with rule descriptions corresponding to
11 the HS codes in a first way, generating a second similarity score by comparing the
12 query with the rule descriptions corresponding to the HS codes in a second way,
13 applying a rule for each of the similarity scores, and if the rule is satisfied, then
14 assigning an HS code to the query. The rule descriptions can be part of a rule file
15 accessed by the rule processing module, and the rule file also has HS codes
16 corresponding to each rule description. Preprocessing, described in greater detail
17 below is preferably preformed so that the query, rule descriptions and historical data
18 are at least formed into tokens. The first similarity score can be calculated by
19 dividing a sum of the number of tokens of the query that match the tokens of the rule
20 descriptions over the number of tokens in the query, for example. The second
21 similarity score is calculated by dividing a sum of the number of tokens of the query
22 that match the tokens of the rule descriptions over the number of tokens in the rule
23 descriptions, for example.

[0033] The particular ad hoc rule used in the method disclosed above can be one of several different types of rules. For example a rule may assign an HS code only if one of the following apply: the first similarity score and the second similarity score are equal to 1; either the first similarity score or the second similarity score is equal to 1; one of either the first similarity score and the second similarity score is greater than a preset level; and both of the first similarity score and the second similarity score is greater than a preset level. The minimum correlation can be less than an exact match (that is, some amount less than 1.0 for each rule description the query is compared with). Also, the ad hoc rule, once created, may be stored in the rule repository 102 accessible by the rule processing module 108. Other ways for calculating similarity scores other rules will be readily apparent to those skilled in the art given the benefit of this disclosure.

[0034] Minimum correlation or similarity 304 is useful especially when the rule type 303 is free text. For example, only 80% of the words (4/5) of the query match a rule description, and only 60% (3/5) of the number of words of the query match the number of words of the rule description. Such minimum similarity of two variables used in combination has advantageously been found to be useful in establishing classification codes for items. Where a minimum similarity is not specified in a rule, the system may use default minimum similarity scores.

[0035] Referring back to Fig. 2, at block 203 uploaded rules can be processed and converted into the format that can be inputted into a text search engine library such as Apache Lucence indexer, for example. In the block 204 processed rules can be indexed by the Apache Lucence indexer to the rule repository 101. Fig. 4 shows a

1 schematic flow chart of a machine learning training module 107. The module starts
2 when an administrator uploads 402 a training data file 401 via administrative user
3 interface 113. At block 403, an administrator can manage the training data and
4 prepare a dataset for use in training. At block 404 the administrator can schedule
5 training. The scheduler can start the training process according to the training plan
6 defined in block 404. At block 405 the "True" algorithm can be used and
7 subsequently followed by use of the PMSVM to train for a plurality of models which
8 represent the different levels of HS code 406, 407, 408 and 409, for example. More
9 generally, a system and method for customs code tariff classification of items using
10 harmonized system (HS) classification codes, comprises receiving the query,
11 comparing the query with rule descriptions corresponding to the HS codes using a
12 machine learning rule, and if the machine learning rule is satisfied, then an HS code
13 is assigned to that item. The machine learning rule is generated by applying an
14 algorithm to a set of historical data, and the algorithm of the machine learning rule
15 comprises at least one of a TRUE algorithm and a power mean support vector
16 machine algorithm.

17
18 [0036] Fig. 5 depicts an example of a format of a training file of one of the computers
19 showing the training dataset in a manner similar to the display of the ad hoc data set.
20 Fig. 6 show a flow chart the training process 405 of the "TRUE" algorithm for
21 associating historically used keywords with particular classification codes. Training
22 dataset of historical data can be loaded in block 601 to train the TRUE algorithm. It
23 has been found that many queries from declaration data include noise (ambiguous
24 statements). This is the case because the declarants tend to use item descriptoins

1 from commercial invoices to make the declaration. Such noises are preferably
2 cleaned in pre-processing step 602 before further processing.

3
4 [0037] Preprocessing of the query may be done prior to producing a feature vector.
5 The step of preprocessing can comprise, for example, separating joint alphabetical
6 and non-alphabetical characters of the query, removing punctuation of the query,
7 tokenizing the query, removing letter case of the query, and singularizing the query,
8 and would typically include all of these steps. Fig. 7 describes the pre-processing
9 step 602 of Fig.6 in greater detail. At block 701 joint alphabetical and non-
10 alphabetical characters can be separated to support better tokenization. Punctuation
11 can be removed at block 702 and tokenization of words takes place at block 703.
12 Case letter folding, converting all words into all upper or lower case can be done at
13 block 704. A singularization process 705 converts tokenized words into singular
14 form.

15
16 [0038] Referring back to Fig. 6, feature vector preparation can be done at block 603
17 and feature vectors created can be stored at feature vector database 103. The
18 TRUE algorithm uses multi-gram features typically extending from unigram (one
19 word) to pentagram (five words) for both training and classification process. The
20 TRUE algorithm can induces the machine learning rule at block 604 from generated
21 features. During the rule induction process True algorithm can consider the following
22 criterion: Feature frequency on a class, which refers to the number of training
23 records which contains one particular feature under a certain HS code; Feature
24 coverage on a class, which is a ratio between the number of training records which
25 contains one particular feature under a certain HS code and total number of training

1 records under that HS code; Feature confidence on a class, which is a ratio between
2 the number of training records which contain one particular feature under a certain
3 HS code and the total number of training records which contains that feature;
4 Normalized feature confidence on a class which refers to a class is a ratio between
5 the normalized frequency of training records which contains one particular feature
6 under a certain HS code and the normalized frequency of training records which
7 contains that feature in the whole training dataset (the normalized frequency is an
8 adjusted frequency produced by multiplication of original frequency and
9 normalization factor of respective HS code wherein normalization factor of a HS
10 code is a ratio between the total number of training records for that HS code and a
11 constant number); and Class distributional factor of a feature which is a ratio
12 between the numbers of HS codes associated with one particular feature and the
13 total number of HS codes available in the whole training dataset. By using the
14 aforementioned criterion the TRUE algorithm nominates the rule candidates and
15 stores them in a temporary file 605. Subsequently, in block 607 the TRUE algorithm
16 can prune the rule candidates to a final rule set. The main concept used in rule
17 redundancy pruning 607 is that if a certain HS code can be predicted by a certain
18 lower-gram feature then it is not necessary to use higher-gram features which
19 contain that particular lower-gram feature. For example, if the unigram feature
20 "iphone" can be used to predict the HS code, then it is not necessary to use bigram
21 feature "iphone_6s" to predict that HS code. Thus, such a redundant feature will be
22 pruned at block 607. Once the final rule set is settled, in the block 607 TRUE
23 algorithm will compiled the rules into a model and store the trained rule in the model
24 repository 101.

25

[0039] Referring back to Fig. 4, after TRUE algorithm training process at block 405, the different models of the PMSVM will be trained at block 406, 407, 408 and 409. The model trained in block 406, 407, 408, 409 will be used to predict 2 digits (chapter level) HS code, 4 digits (heading level) HS code, 6 digits (sub-heading level) HS code, full digits (national level) HS code respectively. Fig. 8 depicts the flow diagram of the steps to be taken to train the PMSVM algorithm for the processes of steps 406, 407, 408, 409. PMSVM training process follows the similar procedure as TRUE algorithm. The machine learning rule can be generated by applying the power mean support vector machine algorithm to historical data, for each of different levels of the HS code. Starting with training dataset loading 801, pre-processing 602, feature vector preparation 802. Advantageously, it has been found that only unigram and bigram features can be generated at block 802 for the PMSVM algorithm. This has the effect of reducing the training time. In block 803, the PMSVM algorithm can train the different models using generated features. Trained models can be stored in the model repository 101 at block 804.

[0040] Fig. 9 depicts the high level view of a classification process where the user may enter the query 901 or query from the user interface 112. HS Classification module 109 can process the query and give the appropriate classification result 902 depending the information which can be extracted from the given query. Classification result 902 may include single full digit (national level) HS code, single six digits (sub-heading level) HS code, single four digits (heading level) HS code, single 2 digits (heading level) HS code and the suggestions of up to five full digit (national level) HS codes, for example. If the classification result is not full digit (national level) HS code, the user may drill down to the national level HS code using

1 the user interface 112. The related tariff information for the national level HS code
2 can also be seen on the user interface 112.
3
4 [0041] Fig. 10 shows an expanded flow chart for the classification process. The
5 system can attempt to classify with rules 1001, TRUE algorithm 1004 and different
6 models of PMSVM 1007, 1011, 1014, 1017. Fig. 11 depicts the flow diagram of
7 classification process by rule 1001. In a manner similar to the training module, the ad
8 hoc module can have queries subject to a preprocessing step 602. Rule matching
9 can be conducted in the subsequent step 1101. If given query matches with a single
10 rule or multiple rules pointed to a single HS code, then the system will return that HS
11 code as a prediction result 1003. Otherwise, the rule engine will hand over the
12 prediction process to the TRUE algorithm 1004. Fig. 12 depicts the flow diagram of
13 classification process by TRUE algorithm 1004. Preprocessing can be done at block
14 602. The TRUE algorithm classification process 1004 and TRUE algorithm training
15 process 405 must use the same feature vector preparation step 603. In the step
16 1201, the TRUE algorithm predicts an HS code by matching between generated
17 feature vectors and the feature vectors in the trained TRUE algorithm model. The
18 following describes representative acceptance criterion for the TRUE algorithm 1202:
19 Generated features of the given description can be matched with at least two
20 features from the trained TRUE algorithm model and those matched features each
21 point to a single HS code; and there must not be more than a single HS code
22 nominated. If both conditions are met, the TRUE algorithm will return a single full
23 digit (national level) HS code as a prediction result 1006. Otherwise, the TRUE
24 algorithm will hand over the prediction process to PMSVM algorithm to predict with
25 different levels of HS code model 1007, 1011, 1014, 1017.

1
2 [0042] Fig. 13 depicts a flow chart of the classification process by PMSVM algorithm
3 1007, 1011, 1014, 1017 in accordance with one embodiment. In block 602, the
4 system will clean the noises from the query. Next in block 802, feature vector for the
5 given query can be generated as was discussed above at Fig. 4 in steps 406, 407,
6 408, 409. The relevant PMSVM model can be retrieved at step 1301 from the model
7 repository 101. Subsequently in step 1302, the algorithm ranks or sorts HS code by
8 the relevancy score or similarity score generated by the PMSVM algorithm. Later in
9 step 1303, the algorithm returns ranked HS code list with associated relevancy
10 scores to the main classification process in Fig. 10. Referring back to Fig. 10, the
11 PMSVM algorithm takes over classification process starting from the step 1007.
12 First, the algorithm starts by classifying with full digit (national level) HS code model
13 1007. In the step 1008, the system checks highest relevancy score generated from
14 step 1007. If the score is higher than a predefined minimum acceptance score or
15 threshold, the module proceeds to step 1009 to decide whether to predict a single
16 HS code directly or to make a suggestion. If highest relevancy score has the clear
17 distance 1009 from the others, that is, there does not appear to be confusion with
18 other HS codes, the system will predict the HS code associated with highest
19 relevancy score 1010. Otherwise, the system will proceed to prepare a suggestion
20 list 1020 which contains up to five HS codes ranked by their relevancy scores
21 generated by the PMSVM algorithm at step 1007. Referring back to step 1008, if
22 highest relevancy score generated from step 1007 does not pass the threshold, the
23 system can proceed to classify with broader PMSVM models in step 1011 (6 digits
24 model), 1014 (4 digits model) and/or 1017 (2 digits model). Once a certain PMSVM
25 model by passes the pre-defined threshold, the system will stop classifying broader

1 PMSVM models and predict the HS code with the highest relevancy scores
2 generated by that particular model. The system can prepare a suggestion list 1020 if
3 the system cannot make the prediction even after classification with the broadest 2
4 digits PMSVM model or if there is confusion in step 1009. The following criterion can
5 be used to prepare a suggestion list at step 1020: the suggestion candidates should
6 be the top 5 HS codes ranked by relevancy scores generated by full digits PMSVM
7 model 1007; the relevancy scores of the suggestion candidates should be within the
8 predefined distance from the highest relevancy score; and the suggestion candidates
9 must pass pre-defined rejection line.

10 If there are some suggestions generated by step 1020, the system will return the
11 suggestion list 1022 to the user. Otherwise, the system will not predict an HS code
12 for the given query.

13
14 [0043] Fig. 14 shows a high level schematic flow chart of the process to to set up a
15 classification service in one country. The process flow can be initiated when there is
16 a requirement to enable HS Classification service in a particular country, such as at
17 a port of entry. Historical declaration data and tariff data for that particular country
18 for can be acquired 1401. Subsequently in block 1402, acquired historical data can
19 be cleaned and analyzed on both qualitative and quantitative aspects. Next at step
20 1403, customization of the structure of the data, preparation of a storage structure
21 for a tariff database 104, model repository 101, rule repository 102, and feature
22 vector database 103 can be performed. At step 1404, cleaned historical data and
23 tariff data can be loaded to a system developed as discussed herein. Subsequently
24 at step 1405, an administrator can select the best parameters for PMSVM by
25 analyzing the performance of the system, for example. Once the PMSVM algorithm

1 is set, customization of HS Classification web-service and web application can be
2 done at step 1406 and 1407 to align with requirements of the particular country. As
3 the final step 1408, HS Classification web-service and web application can be
4 deployed at application server 106 and web server 110 respectively. After
5 successful deployment, the HS classification service and administrative services
6 such as rule indexing 108 and historical data training 107 will be accessible through
7 normal user interface 112 and administrative interface 113. Other types of beneficial
8 information which may be presented on display screens of both the administrators
9 and the users will be readily apparent to those skilled in the art given the benefit of
10 this disclosure.

11
12 [0043] From the foregoing disclosure and detailed description of certain
13 embodiments, it will be apparent that various modifications, additions and other
14 alternative embodiments are possible without departing from the true scope of the
15 invention. The embodiments discussed were chosen and described to provide the
16 best illustration of the principles of the invention and its practical application to
17 thereby enable one of ordinary skill in the art to use the invention in various
18 embodiments and with various modifications as are suited to the particular use
19 contemplated. All such modifications and variations are within the scope of the
20 invention as determined by the appended claims when interpreted in accordance
21 with the breadth to which they are fairly, legally, and equitably entitled.

CLAIMS

What is claimed is:

1. A method for customs code tariff classification using harmonized system (HS) classification codes, wherein the HS classification codes includes a plurality of HS codes, comprising in combination the steps of:
 - loading a training dataset, the training dataset including a plurality of training records;
 - generating a machine learning rule, wherein generating the machine learning rule includes the steps of:
 - applying a first algorithm to the training dataset, wherein the first algorithm:
 - nominates rule candidates using the criteria of feature frequency, feature coverage, feature confidence, normalized feature confidence, and class distributional factor;
 - prunes the rule candidates to a final rule set by pruning redundant higher-gram features;
 - compiles the final rule set into a first model; and
 - stores the first model; and
 - subsequent to applying the first algorithm, applying a second algorithm to the training dataset to train and store a second model, wherein the second algorithm is a power mean support vector machine (PMSVM) algorithm;
 - receiving a query for an item;
 - comparing the query to a rule file of an ad hoc rule, and if comparing the query to the rule file predicts one of the plurality of HS codes, then assigning the one of the plurality of HS codes to the item;

1
2 if comparing the query to the rule file fails to predict one of the plurality of HS
3 codes, then comparing the query to the first model, and if comparing the query to the
4 first model predicts one of the plurality of HS codes, then assigning the one of the
5 plurality of HS codes to the item; and

6 if comparing the query to the first model fails to predict one of the plurality of
7 HS codes, then comparing the query to the second model, and if comparing the query
8 to the second model predicts one of the plurality of HS codes, then assigning the one
9 of the plurality of HS codes to the item.

10
11 2. The method of claim 1

12 wherein using the criteria of feature frequency includes considering the
13 number of training records in the training dataset that contains one particular feature
14 under a certain HS code;

15 wherein using the criteria of feature coverage on a class includes considering
16 a ratio between (a) the number of training records in the training dataset that contain
17 one particular feature under a certain HS code and (b) a total number of training
18 records in the training dataset under that HS code;

19 wherein using the criteria of feature confidence includes considering a ratio
20 between (a) the number of training records in the training dataset which contain one
21 particular feature under a certain HS code and (b) the total number of training records
22 in the training dataset which contains that feature;

23 wherein using the criteria of normalized feature confidence includes
24 considering a ratio between (a) the normalized frequency of training records in the
25 training dataset which contains one particular feature under a certain HS code and (b)

1 the normalized frequency of training records in the training dataset which contains
2 that feature in the training dataset; and

3 wherein using the criteria of class distributional factor includes considering a
4 ratio between (a) the number of HS codes associated with one particular feature and
5 (b) the total number of HS codes available in the training dataset.

6
7 3. The method of claim 1 further comprising a step of preprocessing the query
8 and the training dataset, which comprises at least one of:

9 separating joint alphabetical and non-alphabetical characters;

10 removing punctuation;

11 tokenizing;

12 removing letter case; and

13 singularizing.

14
15 4. The method of claim 1 wherein the plurality of HS codes comprise one or more
16 levels, including a specific chapter, a heading, a sub-heading and national level
17 numbers, and the HS code assigned can vary by these levels depending upon the
18 query; and

19 classifying a full digit national level HS code first, and if a highest relevancy
20 score generated does not pass a threshold for assigning the HS code, then

21 classifying a 6 digit sub-level, and if the highest relevancy score generated
22 does not pass the threshold, then

23 classifying a 4 digit sub-heading level, and if the highest relevancy score
24 generated does not pass the threshold, then

25 classifying a 2 digit specific chapter.

1

2 5. The method of claim 1 wherein the first algorithm comprises an assumption
3 that a phrase that mostly appears in one of the plurality of HS codes is associated
4 with that one of the plurality of HS codes.

5

6 6. The method of claim 1 wherein a feature vector is generated from the training
7 dataset, and further comprising the step of storing the feature vector in a feature
8 vector database.

9

10 7. A method for customs code tariff classification of items using harmonized
11 system (HS) classification codes, comprising in combination the steps of:

12 receiving a query for an item;

13 preprocessing the query and rule descriptions into tokens;

14 generating a first similarity score by dividing a sum of the number of tokens of
15 the query that match the tokens of the rule descriptions over the number of tokens in
16 the query;

17 generating a second similarity score by dividing the sum of the number of
18 tokens of the query that match the tokens of the rule descriptions over the number of
19 tokens in the rule descriptions;

20 applying an ad hoc rule using both of the first similarity score and the second
21 similarity score; and

22 assigning an HS code only if one of the following apply:

23 the first similarity score or the second similarity score are equal to 1;

24 either the first similarity score or the second similarity score is equal to

25 1;

1 either the first similarity score or the second similarity score is greater
2 than a preset level; and

3 both the first similarity score and the second similarity score are greater
4 than a preset level.

5
6 8. The method of claim 7 further comprising the step of storing the ad hoc rule in
7 a rule repository accessible by a rule processing module.

8
9 9. The method of claim 1 further comprising the step of storing the machine
10 learning rule in a model repository generated by a training module and accessible by
11 a classification module.

12
13 10. The method of claim 1 wherein the machine learning rule is generated by
14 applying the power mean support vector machine algorithm to the training dataset, for
15 each of the different levels of the HS code.

16
17 11. The method of claim 1 wherein the second model is used to predict a full digit
18 national level HS code, and further comprising the steps of:

19 subsequent to applying the second algorithm, applying a third algorithm to the
20 training dataset to train and store a third model, wherein the third algorithm is a power
21 mean support vector machine (PMSVM) algorithm and the third model is used to
22 predict a 6 digit sub-heading level HS code;

23 subsequent to applying the third algorithm, applying a fourth algorithm to the
24 training dataset to train and store a fourth model, wherein the fourth algorithm is a

power mean support vector machine (PMSVM) algorithm and the fourth model is used to predict a 4 digit heading level HS code; and

subsequent to applying the fourth algorithm, applying a fifth algorithm to the training dataset to train and store a fifth model, wherein the fifth algorithm is a power mean support vector machine (PMSVM) algorithm and the fifth model is used to predict a 2 digit chapter level HS code.

12. The method of claim 11 further comprising the steps of:

if comparing the query to the second model fails to predict one of the plurality of HS codes, then comparing the query to the third model, and if comparing the query to the third model predicts one of the plurality of HS codes, then assigning the one of the plurality of HS codes to the item;

if comparing the query to the third model fails to predict one of the plurality of HS codes, then comparing the query to the fourth model, and if comparing the query to the fourth model predicts one of the plurality of HS codes, then assigning the one of the plurality of HS codes to the item; and

if comparing the query to the fourth model fails to predict one of the plurality of HS codes, then comparing the query to the fifth model, and if comparing the query to the fifth model predicts one of the plurality of HS codes, then assigning the one of the plurality of HS codes to the item.