Auto-Categorization of HS Code Using Background Net Approach

Liya Dinga*, ZhenZhen Fan, DongLiang Chen

*aInstitute of Systems Science, National University of Singapore
25 Heng Mui Keng Terrace, Singapore 119615
*bCrimsonLogic, Singapore

Abstract

The Harmonized System of tariff nomenclature created by the Brussels-based World Customs Organization is widely applied to standardize traded products with Code, Description, Unit of Quantity, and Duty for Classification, to cope with the rapidly increasing international merchandise trade. As part of the function desired by trading system for Singapore Customs, an auto-categorization system is expected to accurately classify products into HS codes based on the text description of the goods declaration so to increase the overall usability of the trading system. Background Nets approach has been adopted as the key technique for the development of classification engine in the system. Experimental results indicate the potential of this approach in text categorization with ill-defined vocabularies and complex semantics.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Peer-review under responsibility of KES International

Keywords: Text categorization; Background net; HS code classification

1. Introduction

In order to cope with the rapidly increasing international trade around the world, the Brussels-based World Customs Organization (WCO) has created and administered the Harmonized Commodity Description and Coding System, also known as the Harmonized System (HS) of tariff nomenclature. Through using the HS code, it is expected to standardize traded products with Code, Description, Unit of Quantity, and Duty for Classification. Since coming into effect in 1988 this system has been adopted as a numeric language of international trade by more than...
200 countries worldwide, which covers 98% of international merchandise trade. To stay competitive in trade facilitation, trade companies worldwide have invested heavily to further automate and optimize the current trade process, replacing manual operations of HS classification with automated text categorization using machine learning techniques. However, trade studies show that, about 30% of declaration submission uses wrong HS code. This indicates clearly that accurate HS classification can be a highly challenging task to achieve.

As part of the function desired by trading system for Singapore Customs, a new categorization system has been designed to classify products into HS codes based on the text description of the goods, which is expected to improve the accurate handling of goods declaration, and increase the overall usability of the trading system. Background Nets (B-net) approach has been adopted as the key technique for the development of classification engine. A background net, as suggested by the name, captures useful semantics of background information through incremental learning of co-occurrence of words in the text to achieve robust classification in specific application domain with open and evolving vocabulary. The experiments have shown encouraging results.

The rest of the article is organized as follows. Section 2 provides more background information about the specific application domain as well as the challenges in solving this problem; Section 3 gives a brief introduction to the B-net approach with the focus on its representation, learning and inference; Section 4 describes the system design of the application and the experimental results for performance evaluation; finally, conclusions are given in Section 5.

2. Application Background

2.1. Why is HS Classification Important

HS code is a 6-digit international numerical code to represent and identify the goods for worldwide trade. In addition to the internationally standardized 6-digit code, each country is able to further add more digits to extend the code to 8, 10 or 12-digit for its own tariff and statistics purpose. HS Classification is the process of finding the most specific description in the harmonized system (HS) for the goods to be classified. For government, correct classification is required for three main purposes: 1. Calculation of duties, taxes and fees. 2. Determination of permits, license and certificates required. 3. Collection of trade statistics. For companies, correct classification can expedite the custom clearance process by avoiding unnecessary non-compliance which can cause shipment delays, increased number of inspections, fine, and other administrative penalties.

2.2. The Challenges

HS classification with satisfactory accuracy, however, is rather challenging to achieve. The difficulties are mainly due to three factors:

a. **HS Complexity.** The HS is a structured multipurpose nomenclature, organized into 21 Sections and 96 Chapters. WCO has developed a substantial number of defined rules, which include Notes, Subheading Notes, and Explanations of code structure to assist customs officers and other experts, but not common traders. These contribute to the difficulties associated with properly classifying products in the HS.

b. **Gaps in terminology.** There is always a gap between goods description in trade and their description in HS nomenclature. For example, traders would like to declare “MP3 player”, but they need to realize that it belongs to “85.19 - Sound recording or reproducing apparatus: - - Other apparatus: 8519.81 - - - Using magnetic, optical or semiconductor media.” Simple string search cannot help traders to locate the relevant HS codes because of the difference between the structured descriptions of HS nomenclature and the text descriptions during trade process.

c. **The evolving nature.** The 6-digit HS codes are revised every five years. And national HS codes change more frequent, sometimes several times per year. This requires a classification system to be robust and adaptable to continuously changing goods descriptions.

3. Background Net Approach

Machine learning approaches have been applied in capturing and learning relatively stable and long-term criteria for text categorization. One common and useful approach is to use a provided set of keywords under the well-
known vector space model (VSM)\textsuperscript{11-13}. In VSM, a document or a keyword set is represented as a feature vector in a universal feature space. The task of text categorization is then considered as a process of computing the similarity between feature vectors, and the result returned is the most likely categories based on the criteria provided. Such model with features obtained through statistical methods can be insufficient in representing documents with rich but ill-defined semantic information. In order to represent a document more accurately matching its content in the form of feature vector, techniques from information science and machine learning fields have been proposed. Term weighting is a statistical method used to evaluate how important a term is to a document in a collection\textsuperscript{14-15}. Feature selection and extraction is often performed for text categorization, to transform the original feature space into a smaller feature space with the extracted important features\textsuperscript{12,16}. However, selecting an appropriate set of features remains a difficult task\textsuperscript{17}, and a feature set that’s too small or too large will often lead to a poor performance.

The key characteristics that distinguish background net approach from other representative methods in text categorization (such as graph-based model\textsuperscript{18-21}, fuzzy set model\textsuperscript{22,23}, or fuzzy-rough hybrid approach\textsuperscript{24}) are:

1) Instead of using a predetermined fixed set of features, a freely expanding set of words together with the co-occurrence between words is used to represent a text document, so as to avoid the tedious and time-consuming process of feature selection and also to better cope with an open and changing domain of content;
2) Co-occurrence of words is learned in an incremental manner to relieve the demand of large sample data for construction of classifiers;
3) A graphic representation with an expanding set of vertices is employed to represent an accumulation of background information from a specific content domain of concern;
4) Different evaluation measures (symmetrical similarity and unsymmetrical acceptance) are applied to serve different purposes of article selection or classification, which helps reduce the impact of irrelevant information from background domain, leading to increased robustness;
5) Inference on background net is carried out in a “focus spreading” manner rather than over the entire network, so to ensure a reasonable complexity of computation.

3.1. Representation

**Definition 3.1.**\textsuperscript{14} A background net $N = <V, E>$ is a weighted undirected graph, with vertex set $V$ representing a set of terms or a group of terms as a vertex:

$$V = \{v_i \mid v_i = \text{Symbol}(\text{term}_i), \text{term}_i \in U, i = 1, ..., q\}$$

where, $U$ is a set of $q$ terms obtained from article(s), and $\text{Symbol}(\text{term}_i)$ is the symbolic representation of $\text{term}_i$; and edge set $E$ representing the relation indicating contextual association between two terms:

$$E = \{e_{i,j} \mid e_{i,j} = (v_i, v_j), v_i, v_j \in V, i, j = 1, ..., q, i \neq j\}$$

where, each edge $e_{i,j}$ is associated with a weight $w(e_{i,j}) = w(e_{j,i}) = \text{Count}(v_i, v_j)$, $w_{i,j}$ for short, defined as the count of co-occurrence of $v_i$ and $v_j$ in basic units of article(s).

Definition 3.1 is general, and applies to a single article as well as a set of articles that contains the former as a special case. It does not provide a specific restriction for the partition of article in capturing contextual association between $v_i$ and $v_j$. Following the normal practice in document processing\textsuperscript{11,12}, sentence is used as a basic unit to partition an article in our current development, but other methods may also be possible.

**Example 3.1.** Given an article consisting of three sentences, each with three words: $A_1$: $t_1, t_2, t_3 \mid t_2, t_3, t_4 \mid t_1, t_2, t_5$. Fig. 1 shows the corresponding background net.

![Fig. 1. Background net of Example 3.1](image-url)
semantic relations between terms under certain domain knowledge defined by knowledge engineers. It is also different from latent semantic indexing (LSI), for which the main purpose is feature extraction from a term-document matrix in information retrieval.

3.2. Learning from Data

3.2.1. Incremental Learning

Given a collection of text records with each assigned a category label, a supervised learning process is carried out in an incremental manner taking the sample documents from the collection one by one. For a particular application with finite number of categories, one background net is constructed for each category through learning. Initially a background net of category \( c \) is set to be \( N^{(c)} = \langle \emptyset, \emptyset \rangle \). A single record of category \( c \) for learning is first represented as a background net \( N^{(a)} \) and then merged to \( N^{(c)} \) to get updated \( N^{(c\text{-new})} \).

Assume \( N^{(c)} = \langle V^{(c)}, E^{(c)} \rangle \) for category \( c \), and \( N^{(a)} = \langle V^{(a)}, E^{(a)} \rangle \) for a new case of learning, without loss of generality. After learning, the updated \( N^{(c\text{-new})} = \langle V^{(new)}, E^{(new)} \rangle \), where \( V^{(new)} = V^{(a)} \cup V^{(c)} \), and the weight \( w^{(new)}_{i,j} \) of each edge \( e^{(new)}_{i,j} \in E^{(new)} = E^{(c)} \cup E^{(a)} \) for \( v^{(new)}_i, v^{(new)}_j \in V^{(new)} \) can be determined by (3).

\[
W_{ij}^{(new)} = \frac{k \times W_{ij}^{(c)} + (l^{(a)} - l^{(c)}) \times W_{ij}^{(new)}}{k + 1}
\]

where, \( k \geq 1 \) is the number of documents learned, \( 0 < l \leq 1 \) the confidence factor,

\[
W_{ij}^{(c)} = \begin{cases} 0 & e^{(new)}_{i,j} \in E^{(c)} \\ 1 & \text{otherwise} \end{cases},
\]

\[
W_{ij}^{(new)} = \begin{cases} 0 & e^{(new)}_{i,j} \in E^{(a)} \\ 1 & \text{otherwise} \end{cases},
\]

and

\[
\mu(e^{(new)}_{i,j}) = \begin{cases} \frac{W_{ij}^{(a)}}{M^{(c)}} & e^{(new)}_{i,j} \in E^{(c)} \\ 1 & \text{otherwise} \end{cases},
\]

\[
M^{(c)} = \max_{i,j : e^{(c)}_{i,j} \in E^{(c)}} W_{ij}^{(c)}.
\]

The above (3) ~ (5) are for conceptual definitions, and simplification on calculation has been made in actual implementation to reduce the computation complexity, which is linear to the number of documents and linearithmic or loglinear to the number of terms.

The learning processing is described as in Algorithm-1.

---

**Algorithm-1: Learning(D)**, the input \( D = \{d_i | i = 1, 2, \ldots, k\} \) is a set of training documents, \( k \) is the number of documents currently available, and each document \( d_i \) with a given confidence factor \( \lambda_i (i = 1, 2, \ldots, k), 0 < \lambda_i \leq 1 \).

1: \textbf{Initially}, the category’s background net is \( N^{(c)} = \langle \emptyset, \emptyset \rangle \).
2: \textbf{For each} training document \( d_i, i = 1, 2, \ldots, k \)
3: \textbf{Construct} the article’s background net \( N^{(a)} \) for \( d_i \) by Definition 3.1
4: \textbf{Update} category’s background net \( N^{(c)} \) by:
5: \( \emptyset^{(a)} \leftarrow V^{(a)} \cup V^{(c)} \)
6: \( E^{(a)} \leftarrow E^{(c)} \cup E^{(a)} \)
7: \( W_{ij}^{(c)} \leftarrow W_{ij}^{(c)} + (l^{(a)} - l^{(c)}) \times W_{ij}^{(new)} \)
8: \textbf{Update} the weights \( W_{ij}^{(new)} \) to \( W_{ij}^{(new)} / k \), where \( v^{(c)}_i, v^{(c)}_j \in V^{(a)} \)
9: \textbf{Return} \( N^{(c)} \)

The Algorithm-1 is general and can be applied for different application purposes, including test categorization, and personalized document retrieval. The confidence factor refers to the level of preference, confidence, certainty, or importance of document in particular application. In typical classification or categorization applications where sample records are considered with full certainty, the confidence factor is set to be 1 and so can be ignored in calculation.

3.2.2. Association Degree

A background net \( N = \langle V, E \rangle \) captures the contextual association between terms \( v_i \) and \( v_j \) for \( v_i, v_j \in V \), and \( i \neq j \).

The weight \( w_{i,j} \) of edges \( e_{i,j} \in E \) is the count of co-occurrence of terms \( v_i \) and \( v_j \) in a same partition. Based on this
count, for a given term \( v_i \), we can determine an association degree to indicate at what level \( v_i \) is associated to \( v_j \) for \( i \neq j \). The reasoning on background net for text categorization is achieved by comparison of concepts based on the association degree of related terms.

**Definition 3.2.** The 1-step association degree of term \( v_i \) to term \( v_j \) with a background net provided is defined as the degree of direct contextual association from term \( v_i \) to \( v_j \) in fixed one step:

\[
\text{Degree}_{1}(v_i, v_j) = \frac{1}{\sum_{k=1}^{m} w_{i,k}}
\]

When \( \text{Degree}_{1}(v_i, v_j) = 0 \), that means \( v_i \) and \( v_j \) have no direct contextual association in 1-step, there can be two possible situations: 1) an indirect contextual association exists from \( v_i \) to \( v_j \) through other term(s), or 2) they do not have an association within finite steps. In the first case, if there are \( k_1 \) \((k_1 > 1)\) terms \( v_{i,1}, v_{i,2}, \ldots, v_{i,k_1} \) making up \( k_1 \) 2-step associations \( v_{i,q}^1-v_{i,q}^2-v_{i,q}^3 \), \((q = 1, 2, \ldots, k_1)\) then the 2-step association degree from \( v_i \) to \( v_j \), \( \text{Degree}_{2}(v_i, v_j) \), is defined as the maxima among \( k_1 \) values obtained through multiplications of 1-step association degrees of \( v_{i,q}^1-v_{i,q}^2 \) and \( v_{i,q}^1-v_{i,q}^2 \), \((q = 1, 2, \ldots, k_1)\). More generally, when \( \text{Degree}_{1}(v_i, v_j) = 0 \), considering \( m > 1 \) and \( r < m \), if \( m \)-step association can be established as an association chain from \( v_i \) to \( v_j \) via \( v_{i,r}, \ldots, v_{i,m} \in V \), the corresponding \( m \)-step association degree is defined as

\[
\text{Degree}_{m}(v_i, v_j) = \max_{v_i \in V} \left( \text{Degree}_{1}(v_i, v_j) \times \text{Degree}_{2}(v_j, v_i) \right)
\]

where, \( V^r \) is \( r \)-ary Cartesian product over vertex set \( V \) for \( r \) times. \( \text{Degree}_{m}(v_i, v_j) \) takes the maxima among all multiplications of 1-step association degrees along the possible association chains from \( v_i \) to \( v_j \).

The full-step association degree \( \text{Degree}_{\text{full}}(v_i, v_j) \) from term \( v_i \) to term \( v_j \), \( \text{Degree}(v_i, v_j) \) for short, is defined as \( \text{Degree}_{\text{full}}(v_i, v_j) \), where \( M \) is the maximal number of steps of all possible association chains that can be established between \( v_i \) and \( v_j \).

**Example 3.2.** Given a background net \( A_1 = <V_1, E_1> \) represented as an adjacency matrix \( M_1 \) shown below:

\[
M_1 = \begin{bmatrix}
2 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 0 & 0 \\
1 & 1 & 2 & 1 & 0 \\
1 & 0 & 1 & 2 & 1 \\
1 & 0 & 0 & 1 & 1
\end{bmatrix}
\]

and \( V_1 = \{v_1, v_2, v_3, v_4, v_5\} \), the 1-step association degree and full-step association degree are shown in TABLE 1 (a) and (b), respectively.

| 1-Step | 1 | 2 | 3 | 4 | 5 |
|-------|---|---|---|---|---|
| \( v_1 \) | 1.00 | 0.25 | 0.25 | 0.25 | 0.25 |
| \( v_2 \) | 0.50 | 1.00 | 0.50 | 0.00 | 0.00 |
| \( v_3 \) | 0.33 | 0.33 | 1.00 | 0.33 | 0.00 |
| \( v_4 \) | 0.33 | 0.00 | 0.33 | 1.00 | 0.33 |
| \( v_5 \) | 0.50 | 0.00 | 0.00 | 0.50 | 1.00 |

(a) 1-Step Association Degree

| Degree | 1 | 2 | 3 | 4 | 5 |
|-------|---|---|---|---|---|
| \( v_1 \) | 1.00 | 0.25 | 0.25 | 0.25 | 0.25 |
| \( v_2 \) | 0.50 | 1.00 | 0.50 | 0.00 | 0.00 |
| \( v_3 \) | 0.33 | 0.33 | 1.00 | 0.33 | 0.00 |
| \( v_4 \) | 0.33 | 0.00 | 0.33 | 1.00 | 0.33 |
| \( v_5 \) | 0.50 | 0.13 | 0.17 | 0.50 | 1.00 |

(b) Full-Step Association Degree

It is not always helpful to consider full-step association in inference over B-net, and an appropriate \( 1 \leq m^A < M \), called association factor, needs to be selected according to the complexity of contextual information in particular application. In this HS code categorization, we have set association factor to be one, i.e., to make use of 1-step association.

### 3.2.3. Concept of Term

With contextual information captured in a background net, a single word associated with other words is expected to represent richer information than a symbolic term itself. We use the expression “concept of term” to reflect this characteristic.

**Definition 3.3.** The concept of a term \( v \), \( \text{Concept}(v) \) in a given background net \( N = <V, E> \) is defined as a fuzzy set \( \tilde{S} \):
where,
\[
\rho^{(N)}(v, v_j) = \text{Degree}^{(N)}(v, v_j) \quad v, v_j \in V, v \neq v_j,
\]
and the superscript \(N\) indicates the background net \(N\) under discussion. The Degree refers generally to association degree without specific association factor indicated. The concept of term \(v\) is defined as a fuzzy set and represented through the contextual association degree from the term \(v\) to other terms \(v_i (v_i \in V)\), while the term \(v\) itself only serves as a label of the concept. It is also important to note that the fuzzy set of concept of term \(v\) can be varied when different association factor is selected.

**Example 3.3.** The concepts of \(v_1\) and \(v_2\) in B-net \(A_1\) of Example 3.2, based on full-step association, can be represented as fuzzy sets \(c_1\) and \(c_2\), respectively:
\[
\begin{align*}
\text{c}_1 & = \text{Concept}^{A_1}(v_1) = \frac{1}{v_1} + \frac{0.25}{v_2} + \frac{0.25}{v_3} + \frac{0.25}{v_4} + \frac{0.25}{v_5}, \\
\text{c}_2 & = \text{Concept}^{A_1}(v_2) = \frac{0.5}{v_1} + \frac{1}{v_2} + \frac{0.5}{v_3} + \frac{0.17}{v_4} + \frac{0.13}{v_5}.
\end{align*}
\]

### 3.3. Inference

#### 3.3.1. Inference

Inference on background net is carried out through steps of concept comparison between B-nets, typically representing background information of specific domain, input article for classification, or base article for personalized retrieval. To support different types of application, similarity and acceptance measures have been proposed \(^4\). When categorization is aimed, an input record is first represented as a B-net, which is then compared with corresponding B-net that captures application domain through learning from sample data. In this case, and acceptance measure is used for the evaluation.

#### 3.3.2. Similarity and Acceptance

Similarity is a common idea that refers to the closeness of two things of comparison. The similarity measure of two concepts is at a symmetric basis. While the acceptance measure is not symmetric in the sense that it measures how well a concept in the guest background net be accepted by the concept involved in the host background net, with both having the same symbolic representation. The two background nets should not be treated as at the same level of discussion when the former is representing an input article, and the latter a background domain. Fig. 2 gives a conceptual illustration. When the main concern is on how much a guest net \(N_2\) can be accepted by the host net \(N_1\), an acceptance measure should be used. With the page limit, we only provide the definition of the acceptance measure for background net approach, but leave the details of similarity measure to [ref. 4].

![Fig. 2. The similarity and acceptance measures](image)

The acceptance measure of two background nets is evaluated based on the acceptance of individual pairs of concepts that have the same symbolic term throughout two background nets.

**Definition 3.4.** The acceptance of \(N_2 = <V_2, E_2>\) based on background net \(N_1 = <V_1, E_1>\), is defined as
\[
\text{Acceptance}^{(N)}_{\text{Net}}(N_2) = \sum_{\text{Concept}(v_1), \text{Concept}(v_2)} \left( \frac{\text{Acceptance}^{(N)}_{\text{Net}}(N_2)}{|V_2|} \right)
\]

As shown in Definition 3.3, the concept of a term in a particular background net of application domain relates to the distribution of sample data. When the availability of a representative distribution of sample data is a concern, a realistic way is to adopt a simpler binary acceptance of two concepts.
Definition 3.5. Given term $v$, the binary acceptance of concept $c_2 = \text{Concept}(v)$ for $v \in V_2$ in $N_2 = <V_2, E_2>$, based on $c_1 = \text{Concept}(v)$ for $v \in V_1$ in $N_1 = <V_1, E_1>$ is defined as

$$\text{Binary Acceptance}^{(N_1)}_{\text{Concept}}(c_1, c_2) = \frac{|\text{Supp}(c_1) \cap \text{Supp}(c_2)|}{|\text{Supp}(c_2)|}$$

(11)

where, $\text{Supp}(c_1)$ and $\text{Supp}(c_2)$ refer to the support of fuzzy set $c_1$ and $c_2$, respectively.

Combining (10) and (11), we can easily obtain the binary acceptance of $N_2 = <V_2, E_2>$ based on $N_1 = <V_1, E_1>$.

4. HS Code Categorization System Using Background Nets

As mentioned in subsection 2.2, HS code classification lacks a complete set of rules, due to the continuous addition of new goods, and the changing goods descriptions.

After observing the transaction dataset, we notice that a substantial part of records with different codes use some common words, which causes confusion to classifiers. The Table 2 below shows such cases:

| HS Code | Record Description | Brand name | Quantity | Unit of Measurement |
|---------|--------------------|------------|----------|---------------------|
| 22082050 | NAPOLEON VSOP | NAPOLEON | 12 | LITRE |
| 22087000 | CAMUS NAPOLEON COGNAC | NAPOLEON | 16.8 | LITRE |
| 90191010 | MASSAGE CHAIR | OSIM | 101 | NUMBER |
| 90330010 | OSIM MASSAGE EQUIPMENT PARTS | OSIM | 31 | NUMBER |

Note that “NAPOLEON VSOP” is a brandy which is different with cognac but both of them come from “NAPOLEON”. Both MASSAGE chair and part belong to massage equipment but under different categories. Moreover the text descriptions provided by international traders are from a rather open vocabulary, which makes vector based techniques incapable of capturing semantics.

The B-net approach has been adopted for solving the HS code classification problem for two main reasons: (a) It uses co-occurrence based network instead of feature based vector, to capture semantics of the domain and also to be more tolerant with open vocabulary; (b) It uses an incremental learning scheme rather than batch learning, to cope with the changing domain and also to allow learning to start with small sample data.

4.1. Data Preparation

Datasets available are real transaction entries from service providers of current trading system, and those records with HS codes starting with 22 (Chapter 22) and 90 (Chapter 90) are selected for learning and testing as they are more likely to be misclassified in practice. The records from transaction database are converted into common standard format. Record description is used as input for learning and HS Code is the single label of category. Fig. 3 shows an example of structured data record after cleaning.

| HS Code | Record Description | Brand name | Quantity | Unit of Measurement |
|---------|--------------------|------------|----------|---------------------|
| 22042111 | PAUL JABOULET AINE HERMITAGE LA CHAPELLE ROUGE 2003 (14.5%) | PAUL | 9 | LITRE |

Fig. 3. Example of structured data record

Data pre-processing is carried out after cleaning, which includes: removal of numbers, punctuations, and contents within brackets; converting all words into upper case; deleting stop-words using a standard list; and removal of data duplications for each category. Strings with joint alphabetical and non-alphabetical characters are segmented and treated separately, and records with missing attributes are removed.

After all the above steps, the removal of noisy data is carried out. Here noisy data refers to the records with the same product description yet appearing under multiple categories with different HS codes. Table 3 shows the number of data records of Chapter 22 and Chapter 90 before and after the data pre-processing.

| Chapter, Record, Category | Before | After |
|---------------------------|--------|-------|
| Chapter 22: Beverages, spirits and vinegar Records | 210,623 | 40,861 |
| Categories | 52 | 52 |
| Chapter 90: Optical, photographic, cinematographic, measuring, checking, precision, medical or surgical instruments and apparatus; parts and accessories thereof Records | 343,476 | 83,830 |
| Categories | 204 | 204 |
4.2. System Design

The classification system is expected to either work independently as a stand-alone system or to run as a component of current trade systems. Fig. 4 shows the top-level processing flow for a stand-alone classification and as a classifier in trade system, respectively.

Fig. 4. (a) Stand-alone classification system; (b) Classifier in trade system

For this particular application, some relevant B-net settings have been applied:
- Basic unit: the entire text description in a record is treated as basic unit
- Confidence factor: $\lambda = 1$ is set for all the text records
- Association factor: $m^A = 1$ is chosen
- Comparison measure: binary acceptance

4.2.1. Learning of Classifiers

Learning is a one-pass processing. The semantics of each category is captured by one particular B-net, and learning of multiple classifiers for multiple categories can be carried out separately or in parallel. Using incremental learning, a new sample record will only update one B-net that represents the corresponding category, though the effect is global when multiple B-nets are applied to evaluate the same input description during inference.

4.2.2. Acceptance Threshold and Uncertainty Tolerance

There are two requirements of HS Classification system from industry:

a) With noisy environment, a rejection rate around 20% can be acceptable;

b) With ambiguous cases, the system should still provide suggestions with possible uncertainty.

The rejection rate is affected by two predefined parameters, the acceptance threshold, and the uncertainty tolerance. The acceptance threshold indicates the minimum acceptance expected for classification. Typically an input record may be classified into multiple categories with corresponding acceptances assigned. A category assigned with an acceptance less than the threshold will be treated as irrelevant for classification. An input case will be rejected when all the categories under consideration are confirmed as “irrelevant” to it. A “rejected” case is then sent to be handled by manual processing.

On the other hand, multiple candidate categories with high acceptance at the same time may also cause confusion in decision and eventually result in poor performance. To handle such situation, we first rank the candidate categories that are considered “relevant” according to their acceptance in decreasing manner, and then check the difference between the acceptances of categories that are ranked at first and second. If the difference is less than the uncertainty tolerance, the corresponding case will also be rejected.
4.3. Experiments

4.3.1. Performance Measure

There are commonly used performance measures in text categorization and retrieval, such as precision, recall, and their harmonic mean F1. Considering in our particular application of HS code classification, traders may use various ways to describe their goods and the descriptions in records appear as free text, the most relevant performance measure is the overall accuracy of the suggested categories.

4.3.2. Dataset Partition

For experiments, the records after pre-processing are divided into 60% for training set and 40% test set. Different partitions are formed in such way: data under each category is randomly divided into 5 sets (each with 20%) and then 3 sets (60% in total) are used for training and remaining 2 sets (40% in total) for testing. So 10 combinations of data sets are formed to undergo 10 cycles of evaluation.

4.3.3. Experimental Results

Table 4 shows the average accuracy of 10 cycles for Chapter 22 and Chapter 90, respectively with different settings. Note that for Chapter 90, although a lower acceptance threshold could pass more categories as “relevant” at the first stage, it also caused more potential confusion leading to more rejections later. A higher rate of rejection left only those results with better classification (higher acceptance and less confusion) for decision, so eventually resulted in higher classification accuracy. The “Top $k$” ($k = 1…3$) columns in the table indicate the rate of correct category obtained within the first $k$ candidates according to their acceptance measures.

| Settings                     | Rejected Rate | Top 2       | Top 3     |
|------------------------------|---------------|-------------|-----------|
| Min Acceptance = 0.1, Uncertainty Tolerance = 0.1 | 0.0780 | 0.9377 | 0.9782 | 0.9856 |
| Min acceptance > 0, Select Top 1 only | 0.9145 | 0.9145 | 0.9145 | 0.9145 |
| Chapter 22                   |               |             |           |
| Min Acceptance = 0.1, Uncertainty Tolerance = 0.1 | 0.4138 | 0.7488 | 0.7915 | 0.8330 |
| Min acceptance = 0.4, Uncertainty Tolerance = 0.1 | 0.2065 | 0.6055 | 0.7053 | 0.7562 |
| Min acceptance > 0, Select Top 1 only | 0.5301 | 0.5301 | 0.5301 | 0.5301 |
| Chapter 90                   |               |             |           |

Compared with Chapter 22 data, the classification accuracy for Chapter 90 data is much lower. Further analysis on both the data sets and the results has led to the following possible causes:

a. Short record description. Many of the record descriptions in Chapter 90 data have small number of terms or words, often less than three words. Short record description lacks sufficient information for the system to learn enough associations for classification, especially when it also lacks descriptive words. And 84.14% of the short records are actually classified wrongly due to matching too many categories;

b. High level description. Many records are described at very high level, such as “Electronic equipment and parts” which is the text directly captured from HS nomenclature. While a learned B-net “summarizes” the category with such kind of description data, any test data with more detailed and specific descriptions will have issues matching against this category.

5. Conclusion

We have attempted to solve the challenging task of auto-categorization of HS code using background net approach, and obtained primarily encouraging results. Further efforts shall be spent on performance improvement by fully utilizing the B-net approach with fuzzy acceptance when representative distribution of sample data is assured, and with multi-step association to help categorization for short record description. Our future work will also be carried out to explore hierarchical learning through concept clustering in B-net along the line of deep learning.
Acknowledgements

The authors would like to thank Dr. Sio-Long Lo in Macau University of Science and Technology for the suggestions to this application; to Mr. Goh Ming Hui for the participation in the project development.

References

1. World Customs Organization (WCO), What is the Harmonized System (HS)?, from http://www.wcoomd.org/en/topics/nomenclature/overview/what-is-the-harmonized-system.aspx
2. 3CE Technology, Is HS classification difficult?, from http://www.3ce.com/resources/faqs/
3. S. L. Lo, L. Ding, Learning and reasoning on background net – Its application to text categorization, ICIC Express Letters, vol.6, no.3, pp.625-631, 2012.
4. S. L. Lo, L. Ding, Learning and Reasoning on Background Nets for Text Categorization with Changing Domain and Personalized Criteria, The International Journal of Innovative Computing, Information and Control, Vol.9, No.1, pp. 47–67, 2013.
5. Holm Kappler, Reversing the trend: low cost and low risk methods for assuring paper duty payments, World Customs Journal, Vol.5, No.2, pp. 109 – 122, 2011.
6. L. Lv and Y. Liu, Research of English text classification methods based on semantic meaning, Information and Communications Technology, pp.689-700, 2005.
7. Q. Wu, E. Fuller and C. Zhang, Text document classification and pattern recognition, Proc. of International Conference on Advances in Social Network Analysis and Mining, pp.405-410, 2009.
8. X. Q. Yang, N. Sun, T. L. Sun, X.Y. Cao and X. J. Zheng, The application of latent semantic indexing and ontology in text classification, International Journal of Innovative Computing, Information and Control, vol.5, no.12, pp.4491-4499, 2009.
9. D. Bracewell, J. Yan and F. Ren, Single document keyword extraction for Internet news articles, International Journal of Innovative Computing, Information and Control, vol.4, no.4, pp.905-913, 2008
10. A. Leubner and W. Kießling, Personalized keyword search with partial-order preferences, Brazilian Symp. on Database Systems, pp.181-193, 2002.
11. R. Feldman and J. Sanger, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, New York, 2007.
12. C. D. Manning, P. Raghavan and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.
13. G. Salton, A. Wong and C. S. Yang, A vector space model for automatic indexing, Communications of the ACM, vol.18, no.11, pp.613-620, 1975.
14. M. Lan, C. L. Tan, J. Su and Y. Lu, Supervised and traditional term weighting methods for automatic text categorization, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.31 no.4, pp.721-735, 2009.
15. V. Lernattee and T. Theeramunkong, Effect of term distributions on centroid-based text categorization, Information Sciences, vol.158, no.1, pp.89-115, 2004.
16. S. Li, R. Xia, C. Zong and C. R. Huang, A framework of feature selection methods for text categorization, Proc. of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp.692-700, 2009.
17. R. Jensen and Q. Shen, Are more features better? A response to attributes reduction using fuzzy rough sets, IEEE Trans. on Fuzzy Systems. vol.17, no.6, pp.1456-1459, 2009.
18. M. Montes-y-Gomez, A. Lopez-Lopez, and A. F. Gelbukh, Information retrieval with conceptual graph matching, in Proc. 11th Int. Conf. Database Expert Syst. Appl., pp.312-321. 2000.
19. W. Jin and R. K. Sriprihari, Graph-based text representation and knowledge discovery, in Proc. 22nd ACM Symp. Appl. Comput., New York, pp.807-811, 2007.
20. A. Schenker, M. Last, H. Bunke, and A. Kandel, Classification of web documents using a graph model, Proc. of the Seventh International Conference on Document Analysis and Recognition, pp.240-244, 2003.
21. J. Y. Choi, Clustering analysis of collaborative tagging systems by using the graph model, Technical Report, Indiana University Bloomington, 2008.
22. D. A. Buell, An analysis of some fuzzy subset applications to information retrieval systems, Fuzzy Sets System, vol.7, no.1, pp. 35-42, 1982.
23. Y. Ogawa, T. Morita, and K. Kobayashi, A fuzzy document retrieval system using the keyword connection matrix and a learning method, Fuzzy Sets System, vol.39, no.2, pp.163-179, 1991.
24. H. H. Huang and Y. H. Kuo, Cross-lingual document representation and semantic similarity measure: A fuzzy set and rough set based approach, IEEE Trans. on Fuzzy Systems, vol.18, no.6, pp.1098-1111, 2010.
25. S. Kendal and M. Creen, An Introduction to Knowledge Engineering, Springer-Verlag, 2007.
26. L.A. Zadeh, Fuzzy sets, Information and Control, vol.8, no.3, pp.338-353, 1965.
27. G.J. Klir and T.A. Folger, Fuzzy Sets, Uncertainty, and Information, Prentice-Hall, 1988.
28. S.-L. Lo and L. Ding, Incremental Learning of Concept Cluster on Background Net for Fuzzy Granulation of Document Representation, ICIC Express Letters, Vol. 6, No. 3, pp. 699-704, 2012.
29. L. Deng and D. Yu, Deep Learning: Methods and Applications, Foundation and Trend in Signal Processing, vol.7, nos. 3-4, pp.197-387, 2013.