Adaptive Product Classification for Online Marketplace Based on Semantic Analysis

T. F. V. Medeiros and F. G. Cozman

***Abstract*— Each e-Commerce website describes and classifies their products according to what works best for themselves. But anyone willing to merge the catalog of several digital stores must face the problem of ontology matching and reclassification. In this context, this paper proposes an adaptive automatic classification system which relies on the textual description of products to find the correct category it belongs to and after human revision of possible errors, the system improves to better classify the next batch of new products. This is a work in progress, no results are shown yet.**

***Keywords*— E-Commerce, Product Classification, Machine Learning, Semantic Analysis, Adaptive Technology.**

I. INTRODUCTION

INTERNET is now part of our daily life in every aspect, social and economic, cultural and entertainment. Particularly for online retailers, dealing through the World Wide Web offers various advantages over brick and mortar stores, such as the virtually infinite shelf space. But this advantage causes a burden for the customer who has difficulty finding the wanted product among such a vast catalog.

To solve this problem three approaches are common:

- Search engines
- Catalog ontologies
- Recommender Systems

The second approach consists in building a tree of labels which applies to every product. This way, the user is able to navigate through the site narrowing the range of products at each branch chosen on the tree arrangement.

This is a great solution until two or more retailers must communicate, because their ontologies don't match perfectly even if they offer exactly the same products.[8]

Two common situations in which the merge of product catalogs is necessary are E-Marketplace, where many small vendors share the same digital roof; and Comparison Shopping sites which collect the prices of the same product in many stores and put them side by side to help the customers choose the best option.

In both of those business models, they need a unified product classification. And to make the problem harder, it is not static, new products come every day.

When the problem is not focused in one product domain, like movies or clothes or refrigerators, there is no common structured attributes except the product name and textual description, price and original seller or manufacturer.

We propose a system that extracts semantic features from the short textual description of the products and classifies each one into the appropriate category. After the automatic classification, the uncertain items are revised by humans that guarantee every product is found in this right place.

For academic purposes, the system is initially designed to use a previously labeled dataset which is broken in even batches of products, simulating the situation of successive increments in the product catalog. The system must be able to learn and be more accurate in each new batch.

In this first attempt, only the top level of the ontology is considered, leaving the problem of taking into account hierarchical to future works.

The reminder of this paper is divided in a **Brief Outline** of the system workflow, proposed **Application** and **Conclusion**.

II. BRIEF OUTLINE

The functioning of the system is as follows:

1. At first, an initial set of labeled products must be presented to train the classifier so that it shows a tolerable classification accuracy.
2. Thereafter for each new block of product that arrive and must be labeled into the fixed set of categories, this loop ensues:
   a. The current classification rules are applied.
   b. Experts validate the automatic classification, relabeling some products if necessary
   c. The adaptive layer examines the errors made and change the classification rules to avoid them.

Figure 1 illustrates this workflow.

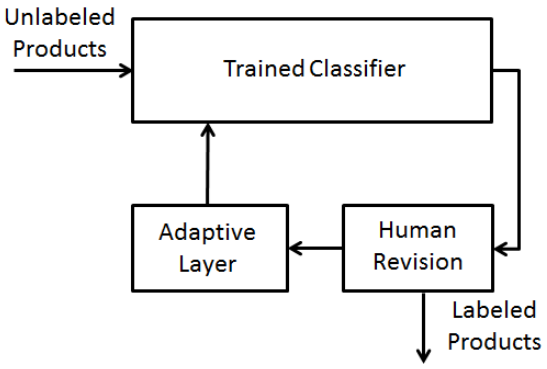

Figure 1 - System Workflow

T. F V. Medeiros, Universidade de São Paulo (USP), São Paulo, São Paulo, Brasil, tacio.medeiros@usp.br
F. G. Cozman, Universidade de São Paulo (USP), São Paulo, São Paulo, Brasil, fgcozman@usp.br

The classifier box is divided in 4 steps as shown in Figure 2. Each step can be made with more than one technique, and most of them require the setting of parameters. This makes the complete classifier very hard to calibrate for best performance using the provided dataset of products.

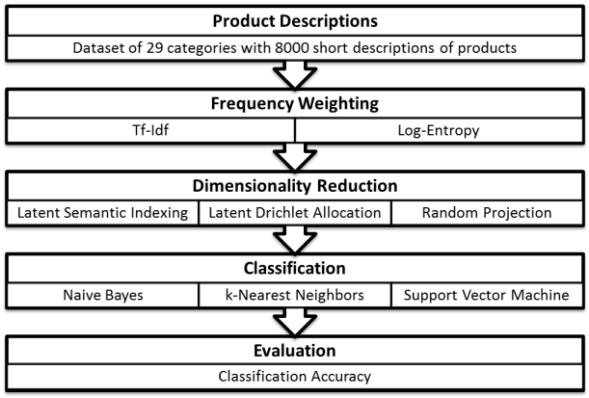

Figure 2 - Classifier steps

III. APPLICATION

The system is being implemented in Python programming language using the topic modelling library Gensim[9] for the Natural Language Processing part and using Orange[10] for the Machine Learning part. The adaptive layer will control the use of both libraries adjusting parameters.

Dataset

The original dataset is a dump of all products

A balanced subset was sampled from the original dataset for this first attempt of a classification system.

The resulting dataset is composed of 29 top level categories, each containing 8000 product descriptions. Adding up to a total of 232000 items.

The descriptions are written in Portuguese, and have been stripped of accented characters and standardized to lowercase. Those descriptions were made by different retailers and do not follow any particular pattern, often not even proper grammar.

Pre-processing

Each product description is broken in an unordered list of words. Furthermore, each pair of consecutive words is also recorded and appended to the list. So each product is now represented by its list of unigram and bigram words.

Before jumping to converting every distinct word to a dimension, some tasks are performed in order to group or disregard some words in order to reduce at firsthand the size of the dictionary.

Entity recognition

Many documents in this dataset contain a product code within the text. Each code is unique and only have meaning for the retailer, but it is recognizable as a code and not all categories have them, so all words code-like are replaced for the placeholder _CODE_ which adds some information for the semantic processing to follow. It is the only Entity Recognition performed in this version of the system.

Frequency cutoff

Words that appear only once or twice in the whole dataset or too often in different documents do not help and can be discarded. The lower and upper threshold are parameters which have to be calibrated.

Stemming

Stemming consists in removing suffixes and prefix and conjugation from words so that only the radical remains, this way a further reduction of the dictionary dimension is achieved.

Frequency Weighting

At this step, each document (product description) is stored as a vector of words. The dimension of this vector is the size of the vocabulary of this dataset.

Now, instead of a simple count of words for each document, some statistical transformations can be performed to enhance the discriminative power of those values. This process is called Frequency Weighting[7].

Among the possible techniques, the most common are Tf-Idf and Log-Entropy.

Dimensionality Reduction

For the next step, of classification, the dimension of the resulting Term-Document Matrix is still too high. So more dimension reduction must be performed.

The most common method is Latent Semantic Indexing (LSI), which is similar in nature to the even more commonly used in any Machine Learning application Singular Value Decomposition (SVD).

Other dimension reduction techniques are Latent Drichlet Allocation (LDA) and Random Projection (RP).

This step is also responsible to group words which are synonyms and distinguish polysemy.

Classification

The initial collection of texts is now in the form of a matrix where the number of rows is the number of products and the number of columns was arbitrarily chosen in the Dimensionality Reduction step.

We have chosen the following classifiers for this experiment which are readily available in the Orange framework:

- Naïve Bayes
- k-Nearest Neighbors (KNN)

- Support Vector Machine (SVM)

Evaluation

The fundamental metric for evaluating this system is the classification accuracy.

On a related work, GoldenBullet[1], it is also used the accuracy of the 10 best class candidates. Since there is a human validation after the automatic classification, the system is tolerant, at least in the beginning, to make a small mistake. If the correct class is among the first 10 predicted ones, it is considered a success, raising considerably the accuracy.

IV. Follow Up

The next steps in this project are to implement the adaptive layer and then making the classification hierarchical, descending in the product categories tree.

This learning program will be modeled as a Cellular Learning Automata as described in [5].

V. Acknowledgement

The first author was funded by the Buscapé Company.

The second author acknowledges support by the Buscapé Company and CNPq.

References

[1] Ding, Y.; Korotkiy, M.; Omelayenko, B.; Kartseva, V.; Zykov, V.; Klein, M.; Schulten, E. & Fensel, D. GoldenBullet: Automated Classification of Product Data. in *E-commerce Business Information Systems*, 2002

[2] Mattos, A.; Kampen, M.; Carriço, C.; Dias, A. & Crivellaro, A. Caseli, H.; Villavicencio, A.; Teixeira, A. & Perdigão, F. (Eds.). E-commerce Market Analysis from a Graph-Based Product Classifier. *Computational Processing of the Portuguese Language*, Springer Berlin Heidelberg, 2012, 7243, 291-297

[3] Beneventano, D. & Magnani, S. A Framework for the Classification and the Reclassification of Electronic Catalogs. *Proceedings of the 2004 ACM Symposium on Applied Computing*, ACM, 2004, 784-788

[4] Jackson, Q. Z. & Landgrebe, J. D. Design Of An Adaptive Classification Procedure For The Analysis Of High-dimensional Data With Limited Training Samples. 2001

[5] Esmaeilpour, M.; Naderifar, V. & Shukur, Z. Cellular Learning Automata Approach For Data Classification. *International Journal Of Innovative Computing Information And Control*, ICIC. INTERNATIONAL TOKAI UNIV, 9-1-1, TOROKU, KUMAMOTO, 862-8652, JAPAN, 2012, 8, 8063-8076

[6] Cereda, P. R. M. & Neto, J. J.. Mineração Adaptativa de Dados: Aplicação à Identificação de Indivíduos. *Workshop de Tecnologia Adaptativa*, 2012

[7] Solka, J. L.. Text Data Mining: Theory and Methods. *Statistics Surveys*, 2008, *2*, 94-112

[8] Ding, Y.; Fensel, D.; Klein, M.; Omelayenko, B. & Schulten, E.. The role of ontologies in ecommerce. *Handbook on ontologies, Springer*, 2004, 593-615

[9] Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

[10] Demšar, J.; Zupan, B.; Leban, G. & Curk, T. Boulicaut, J.-F.; Esposito, F.; Giannotti, F. & Pedreschi, D. (Eds.) Orange: From Experimental Machine Learning to Interactive Data Mining. *Knowledge Discovery in Databases PKDD 2004*, Springer, 2004, 3202, 537-539