Data Mining and the Text Categorization Framework

Paola Cerchiello
University of Pavia, Italy

INTRODUCTION

The aim of this contribution is to show one of the most important application of text mining. According to a wide part of the literature regarding the aforementioned field, great relevance is given to the classification task (Drucker et al., 1999, Nigam et al., 2000).

The application contexts are several and multitask, from text filtering (Belkin & Croft, 1992) to word sense disambiguation (Gale et al., 1993) and author identification (Elliot and Valenza, 1991), trough anti spam and recently also anti terrorism. As a consequence in the last decade the scientific community that is working on this task, has profuse a big effort in order to solve the different problems in the more efficient way.

The pioneering studies on text categorization (TC, a.k.a. topic spotting) date back to 1961 (Maron) and are deeply rooted in the Information Retrieval context, so declaring the engineering origin of the field under discussion. Text categorization task can be briefly defined as the problem of assigning every single textual document into the relative class or category on the basis of the content and employing a classifier properly trained.

In the following parts of this contribution we will formalize the classification problem detailing the main issues related.

BACKGROUND

In a formal way, text categorization is the task of assigning a Boolean value to:

$$(d_j, c_i) \in D \times C$$

where D is a domain of documents and $C = [c_1, \ldots\ldots, c_{|C|}]$ is a set of predefined categories. A value of T assigned to (d_j, c_i) indicates a decision to file d_j (also called *positive example*) under c_i , and a value of F assigned to (d_j, c_i) indicates a decision not to file d_j (also called *negative example*) under c_i. In other words a function, able to approximate the relation between the set D of documents and the set C of categories, is needed by means of a classifier able to learn the main characteristic of each categories on the basis of the information contained in the documents.

We emphasize that the *c* categories are simply symbolic labels (authors' name, spam or not spam, different topics) and often no further information is available.

Regarding the relation between documents and categories two different formulations should be mentioned (Sebastiani, 2003):

- Given a document $d_j \in D$, all the categories under which it could be filed, are checked, (*document-pivoted categorization*-DPC);
- Given a category $c_i \in C$, all the documents that could be filed under it, are checked (*category-pivoted categorization*-CPC).

Actually the above distinction is more practical than theoretical, in any case it can be useful whenever C and D are not fully available from the beginning.

Now we want to refer two main approaches employed in the construction of a text classifier.

The first one is the *Knowledge Engineering*, born in '80, dealing with the manual construction of several decision rules, of type *if (some characteristic) then (category i-th)*. As the reader can simply infer this approach was too expensive either from a human or a time point of view. In fact those rules needed, not only to be thought about by an human expert of the relative topic, but also to be updated every time some constitutive elements change.

The ideal field for TC was found in *Machine learning* community that introduced the idea of a supervised classifier able to learn the key elements of each category on the basis of some preclassified documents. We underline that according to this formulation, the wish is not simply constructing a classifier, but mainly creating a builder of classifier automatizing the whole process.

Another important aspect of every application of text mining, not only text categorization, is the transformation of a textual document in a analyzable database. The idea is to represent a document d_j with a vector of weights $d_j = (w_{1j}, ..., w_{|T|j})$ according to which T is the dictionary, that is the set of terms present at least once in at least *k* documents, and *w* measures the importance of the term. With the concept of term the field literature is usual to refer to a single word from the moment that this kind of representation produces the best performance (Dumais et al. 1998).

On the other side w_i is commonly identified with the count of the i-th word in the document d_j, or with some variation of it like *tf-idf* (Salton & Buckley, 1988).

The data cleaning step is a key element of a text mining activity in order to avoid the presence of highly language dependant words. That latter objective is obtained by means of several activities such as conversion to lower case, managing of acronym, removing of white space special characters, punctuation, stop words, handling of synonymous and also word stemming.

A specific word frequency gives us information about the importance of it in the document analyzed, but it does not mean that the number they appear with, is proportional to the importance in the description of the above document. What is typically more important, it is the proportion between the frequencies of words to each other in the text.

Even after all the activities of data cleaning presented, the number of relevant words

is still very high (in the order of 10^3), thereby features selection methods must be applied (Forman, 2003).

MAIN THRUST

Once defined the preprocessing aspects of a text categorization analysis, it is necessary to focus on one or more classifiers that will assign every single document to the specific category according to the general rule specified above:

- The first and oldest algorithm employed in the field under analysis is the Rocchio method (Joachims, 1997) belonging to the linear classifier class and based on the idea of a category profile (document prototype). A classifier built with such methodology rewards the proximity of a document to the centroid of the training positive examples and its distance from the centroid of the negative examples. This kind of approach is quite simple and efficient however, because of the introduction of the linear division of the documents space, as all the linear classifiers do, it shows a low effectiveness giving rise to the classification of many documents under the wrong category.
- Another interesting approach is represented by the memory based reasoning methods (Masand et al., 1992) that, rather than constructing and explicit representation of the category (like Rocchio method does), rely on the categories assigned to training documents that result more similar to the test documents. Unlike the linear classifiers, this approach doesn't linearly divide the space and therefore we are not expected to have the same problems seen before. However the weak point of such formulation resides in the huge time of classification due to the fact that it doesn't exist a real phase of training and the whole computational part is left to classification time.
- Among the class of non numeric (or symbolic) classifier, we mention decision trees (Mitchell, 1996) which internal nodes represent terms, branches are tests on weights for each words and finally leaves are the categories. That kind of formulation classifies a document d_j by means of a recursive control of weights from the main root, trough internal nodes, to final leaves. A possible training method for a decision tree consists in a 'divide and conquer' strategy according to which first, all the training examples should present the same label, and if it is not so, the selection of a term t_k is needed on the basis of which the document set is divided. Documents presenting the same value for t_k are selected and each class is put in a separated sub-tree. That process is repeated recursively on each sub-tree until every single leaf contains documents belonging to the same category. The key phase of this formulation relies on the choice of the term based on indexes like information gain or entropy.
- According to another formulation, the probabilistic one, the main objective is to fit $P(c_i/d_j)$ that

Related Content

Data Mining Tool Selection

Christophe Giraud-Carrier (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 511-518).*

www.irma-international.org/chapter/data-mining-tool-selection/10868

Data Mining with Cubegrades

Amin A. Abdulghani (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 519-525).*

www.irma-international.org/chapter/data-mining-cubegrades/10869

Distance-Based Methods for Association Rule Mining

Vladimír Bartík (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 689-694).*

www.irma-international.org/chapter/distance-based-methods-association-rule/10895

DFM as a Conceptual Model for Data Warehouse

Matteo Golfarelli (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 638-645).*

www.irma-international.org/chapter/dfm-conceptual-model-data-warehouse/10888

On Association Rule Mining for the QSAR Problem

Luminita Dumitriu (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 83-86).*

www.irma-international.org/chapter/association-rule-mining-qsar-problem/10802