A Text Mining Strategy based on Local Contexts of Words

Simona Balbi, Emilio Di Meglio

Dip. di Matematica e Statistica – Università "Federico II" di Napoli

sb@unina.it, edimegli@unina.it

Abstract

Aim of the paper is to propose a Text Mining strategy based on statistical tools, which make more efficient the extraction of information buried in massive quantities of documents. Usually, in Text Mining procedures (such as in textual data analyses) we deal with a *corpus* consisting of a set of documents. In order to build the data structure to be processed, each document is encoded in a *document vector*, according to the *bag-of-words model*, which associates words and their frequencies for the given document. Documents are considered as a whole. The proposed mining strategy identifies *interesting* sentences in the *corpus* we deal with, where to concentrate the knowledge extraction. The sentence interest will depend on the researcher's objective. The proposed procedure is useful when we are interested in local contexts for words. Prior information, i.e. expert knowledge, is included, as an input for the procedure, but differently to content analysis, the key-word system is automatically built. The strategy can be applied in any case we can introduce information for partitioning documents in lower order grammatical units (e.g. sentences, but also paragraphs, etc.). The mining procedure consists in two steps: first of all the Text Categorisation, i.e. the recognition of the *interesting* sentences, by means of a statistical segmentation procedure, and then the knowledge extraction from the identified sub-texts. The procedure first step produces association rules useful in filtering e-mail, chat, or Web access, too. The paper aims at contributing to the day-by-day wider literature on Text Mining, devoted to go beyond the "bag-of-words" model of structuring the data set in document vectors, enhancing the role of a statistical perspective. An application on Italian on-line job offers ends the paper, showing the effectiveness of the proposal.[1]

Keywords: bag-of-words, association rules, segmentation, text categorisation.

1. Introduction

It is well-known that while data mining deals with numerical data arranged in structured data bases, Text Mining deals with unstructured documents, written in natural language. There are not negligible consequences in dealing with unstructured materials. In numerical data bases, the elementary units are the field contents in records. In Text Mining, things are not so simple. Not trivially, from a statistical viewpoint, it is not definitely clear how to structure the data set to be analysed. First of all we need to answer the questions on which are the statistical units, and which are the variables.

In the field of linguistic statistics and textual data analysis, we can find interesting applications working on letters (vowels, consonants), words (e. g. graphical forms, lemmas, textual forms), groups of words (e.g. repeated segments, quasi-repeated segments), sentences (i.e. grammatical self-contained speech units with a capital letter at the beginning and a full stop at

[1] The paper has been financed by the Italian Ministero dell'Università e della Ricerca scientifica, in the project OUTCOMES (Occupation as a University Target and Careers of Outgoing graduates Maximising their use of Educational Skills), PRIN 2002

the end), and so on. However, the complexity of the choice has some remarkable implications. We can structure the knowledge extraction process moving from data of different levels and orders, so as to make more efficient and effective our results. The practice of representing documents as bag-of-words (i. e. encoding documents in vectors, associating words and their frequencies, and processing documents as a whole) can be overcome, aiming at enhancing the context in which a word has been used.

Here we propose a methodology, based on Text Categorisation, which exploits the structural organization of documents in sentences. This strategy consists of two step: in the first step, statistical units are sentences described by words, while in the second step units are words, characterised by their frequencies in each sentence. In the first step, Text Categorisation is undertaken to discriminate interesting sentences from uninteresting ones (having as an input expert knowledge). This approach, given an informative need, allows to eliminate all the information not useful to satisfy this specific need. Once we have reduced the *corpus* only to the sentences related to our problem, we can apply a proper statistical method for knowledge extraction. In this way, computational speed ups are obtained; documents and terms similarities (fundamental measures for any Text Mining application) are targeted on the particular informative need. This strategy attempts to go beyond the bag-of-words model. In fact, being the bag-of-word built on sentences, relations among words are, in some way, taken into account.

For showing the effectiveness of the proposal, two different typologies have been built, starting from the same collection of on line job offers: one on the original corpus and the other one on the reduced corpus obtained by discarding sentences not containing skill requirements. Comparing results gives an idea of the information gained, as consequence of this light contextualisation.

2. Text Categorisation

Text Categorisation (also named Text Classification) is a Text Mining task with a broad domain of applications, ranging from automatic document indexing to document filtering, metadata generation, word sense disambiguation, hierarchical organization of web documents and any application that require selective documents organization, such as limited Web access (e.g. children, etc.). Text Classification is therefore of great utility in business and Information and Communication Technology applications but is also useful to extract knowledge that constitutes the starting point for other Text Mining applications.

Text Classification can be defined as the task of assigning a Boolean value to each pair (d_i, c_i) of DxC where D is a set of documents and C is a set of pre-defined categories. A value True is assigned if document d_i is classified under category c_i and a value false is assigned if document d_i is not classified under category c_i . Formally, the task is to approximate an unknown target function φ: DxC → *{T,F}* that describes how documents should be classified with a function $\hat{\varphi}$: DxC → *{T,F}* called the *classifier* in such a way that $\hat{\varphi}$ "coincides as much as possible" with φ.

The degree of "coincidence" between the target function and the classifier determines the effectivity of the classification algorithm. The classifier is built on a training set and is validated on a test set. The general problem of statistical classification in textual domain can be summarized as follows. We have a training set of document vectors, each labeled with one class, called target value. In practice, each object of the training set is represented in the form (x_i , y_i) where x_i is the document vector and y_i the class label value. The goal is to learn a mapping or a function f(x) that is able to predict a class value y, given a document (Hand *et al.*, 2001).

At this point it is useful to define the concepts of *model class*, *score function* and *optimization strategy*. The *model class* is a parametric family of classifiers, that is a function f(x, ξ), where ξ is a parameter vector. Usually, in Text Classification problems very few is known on this function. The *score function* is the function that numerically expresses the preference for a model over another. It is typically a function of the difference between the predicted value *y** and the true value *y*, measured with a proper dissimilarity index. The *optimization strategy* is the strategy used for finding the best parameters and models within the model class. The aim of the learning algorithm is therefore to minimize the score function as a function of ξ.

Once the parameters have been chosen, that is once the classifier has been trained, it is necessary to estimate its performance on a test set. The test set should never include data used in the training step. The model could in fact result overfitted. In literature, several techniques have been proposed. In the following section, the main techniques for classifying texts will shortly be reviewed.

Techniques for Text Classification

In general terms, we have three main families of statistical techniques for Text Categorisation: Decision Trees, regression methods and neural networks.

A decision tree is a tree structure that visualizes the document categorisation process. The tree is composed by nodes, branches and leaves. By following the tree from the top node it is possible to classify a new document by recursively choosing the appropriate branches until a leaf is reached. The tree is built on a training set with an iterative algorithm. The training set consists in a set of labeled documents, i.e. documents for which the category is *a priori* known. At each step it is chosen the variable which splits the data into groups leading to the greatest improvement of the score function. The iteration stops when each leaf contains one data point or identical data points or if a given stop rule is met. In this way the maximum tree is obtained. However, this tree describes perfectly the training set but leads to overfitting. For this reason the tree is pruned. Pruning consists in eliminating some branches of the tree in order to have a simpler model and a better performance. The goal is to find a model complex enough to capture the structures existing in the data, but not so complex to overfit. Most used algorithms are CART and C4.5.

Regression methods aim at explaining or predicting a continuous variable, on the basis of explicative variables. Linear Regression is the simplest and most widely used regression model. In Text Classification domain the aim is to obtain a binary value that indicates membership to a certain class. Logistic Regression, proposed for dealing with explicative categorical variables, allows to calculate a real valued ranking for the class membership. The result is, in fact, a *categorisation status value (CSV)*, i.e. a number between 0 and 1 representing the evidence for class membership. The parameters of logistic regression can be estimated using maximum likelihood. Regression methods give measures for the importance of each explicative variable in determining the class membership. It suffers however of some substantial drawbacks. First, it requires considerably more observations than variables to obtain valid estimates of parameters. It is also computationally expensive and this makes it a not suitable methodology for huge data sets.

A Neural Network text classifier is a network of interconnected computing units where the input units represent the terms; the output units, the categories and the weights on the connections, the dependence relations among units. The classification process of a document is performed as follows: the term weights are loaded into the input units; the following units are activated on the basis of the input weights and the connections weights; unit activation is propagated forward in the network and the final classification decision result is determined by the value assumed by the output unit. Neural Network connections weights are learnt on a training set usually through a process called *backpropagation*. In backpropagation, the weights of a training document are loaded and if misclassification occurs the error is back-propagated so to adapt the parameters of the network to minimize this error. In practice, training a neural network consists in minimizing a score function in the parameter space, solving a non linear optimization problem. This is done by descending to a local minimum given a random starting point. The most used optimization techniques are *steepest descent* and *conjugate gradient*. The simplest type of neural network is the perceptron, which is a linear classifier. Non linear Neural Networks are instead networks with one or more hidden layers of units that represent, in Text Classification higher order relations between terms. This is one of the main advantages of using Neural Networks for Text Classification. Neural Networks are however computationally expensive if the network structure is too complex and cannot be as easily interpreted as classification trees and logistic regression.

This review does not claim to be exhaustive. Several methodologies have been proposed either in statistical literature either in Machine Learning literature, being Text Classification a problem studied in both fields. Among these we mention Support Vector Machines, linear discriminant analysis, Bayesian inference methods, maximum entropy modeling, genetic algorithms, association rules induction, bagging and boosting. For a complete overview see Sebastiani (2002).

3. A two-step strategy

In this paper we propose a two-step data analysis strategy, aiming at extracting knowledge from huge *corpora*, by both reducing the computational burden of the textual data analysis and accounting for the context in which words appear. In doing that we deal with some tasks and some solutions developed in Text Mining frame, first of all *Text Categorisation*.

As previously said, Text Categorisation consists in classifying text into one of several predefined categories. This task is performed by software called *automatic categorisers*. The software basically mimics the human process of evaluating the relevance of a document with respect to the topic of interest. A common way for doing that consists in introducing a keyword index, which defines logical rules, e.g. if a certain word occurs in a text, then the text will be identified in some sense and assigned to the related category. Words can be combined by logical operators (AND, OR, NOT), enriching the discriminative power of the rule. The procedure requires expert knowledge, i.e. the key words index, as input, or, alternatively, can be based on statistical methods for categorisation, as previously mentioned. Here we follow the latter approach, proposing statistical tools for reducing the quantity of external knowledge input and determining automatically the rules to apply in order to tag the texts. In the first step of the strategy, we apply the categorisation algorithm to discriminate part of documents (e.g. sentences), inherent to the objectives of the analysis. Therefore, only parts tagged as interesting will be consider in the following analysis step. The strategy can be useful in any case we can introduce information for partitioning documents in units of different levels. The synctatical structure presents in a document suggest to consider it as a complex unit, consisting of lower order units (sentences) and, hierarchically, much lower units (words). Moving at different levels will enable us to efficiently discover the knowledge we are looking for.

STEP 1

Aim: identifying sentences in the document, related to the topic of interest
Tools: statistical techniques for discrimination
Input: a training set and a test set, both consisting in sentences tagged by expert knowledge (0 = uninteresting; 1 = interesting)
Output: logical rules for identifying interesting sentences in the document to be analysed

STEP 2

Aim: eliminating uninteresting sentences in the document, by applying the logical rules identified in STEP 1;
Tools: advanced software or language for dealing with text (e.g. UltraEdit, Perl)
Input: the logical rules identified in STEP 1 and the document to be analysed
Output: a new document consisting of the sentences related to the topic of interest, to be analysed with the proper textual data analysis techniques, according to researcher's objectives

4. Looking for skills required in on-line Italian job offers

4.1. *Motivations*

A research net connecting eight Italian Universities is working on the correspondence between demand and offer of graduates on Italian labour market (*OUTCOMES: Occupation as a University Target and Careers of Outgoing graduates Maximising their use of Educational Skills*). Therefore it is important to analyse the skills required, and how they are required by the labour market (Balbi and Di Meglio, 2003). Here we propose to analyse on-line job offers, applying the proposed strategy. The first motivation is connected with some peculiarities of these texts. A job advertisement is usually composed by several sentences. Each sentence has a specific role, common in the most cases of job advertisements. Some sentences contain information on the required skills, others describe the professional profiles, others contain different information. In other terms, job advertisements have a light structure, and it is possible to identify which sentences are devoted to describe the required skills. Moreover, the vocabulary is not very wide, but some word connotations depend on the kind of sentences, where they are used. Therefore, a first analysis devoted to the mining of sentences has been performed.

4.2. *The first step: building logical rules for skill requirement sentences*

The document to by analysed is a set of job offers appeared on a specialised Italian site during the first 8 months of 2003. We have collected 2017 advertisements. We randomly choose a subset, in order to build the training set, consisting in 200 advertisements, composed by 670 sentences. After a manual tagging of the advertisements, we labelled 502 sentences with 0 (= *uninteresting*) and the remaining 168 with 1 (= *skill requirements*). Once the training set had been prepared, we went on building the logical rules for identifying skill requirement sentences in the rest of the document. For sake of simplicity, here we have chosen to ignore all the problems related to linguistics (e.g. lemmatisation, disambiguation, etc.), and we dealt with graphical forms (Lebart *et al.*, 1998). There are many statistical techniques devoted to classifying units (e.g. discriminant analysis, segmentation, and so on), and there are many mining algorithms developed for solving classification problems, by association rules. As suggested in a previous paper (Balbi and Summa, 2001), Symbolic Marking seems to efficiently perform in the case of Text Mining, being a segmentation method with a very high performance in the case of huge data sets. Additionally its results are easily to be expressed in terms of logical operators.

4.2.1. *Symbolic Marking*

Symbolic Marking (SM, Gettler-Summa, 1998) is a non-binary segmentation technique which aims at finding the association structures in a group G_i belonging to a typology naturally defined, or obtained by a previous classification analysis. Symbolic Marking takes into account logical relations, as conjunctions and disjunctions, between attributes describing the units in G_i. The result can be expressed in natural language as logical rules, connecting attributes with logical operators.

The procedure finds *marking cores* g_k's, i.e. nuclei of individuals identical with respect to some variables, called characterising variables, y_j. The y_j's give a symbolic description of $g_k \in G_i$:

$$g_k : [y_1 = a_1] \wedge [y_2 = a_2] \wedge \ldots \wedge [y_r = a_r]$$

By joining *n* marking cores g_k's (defined by logical AND's, $\wedge$) with the logical disjunctive operator OR ($\vee$), one obtains a (partial) description of the group G_i, in terms of the selected characteristic variables y_j's, optimal according to the principles stated by Gordon (1999): *i)* minimising the number of false negatives (individuals in G_i, but not corresponding to the description); *ii)* minimising the number of false positives (individuals described by the conjunction of the marking cores, but out of G_i); *iii)* each conjunction of categories has to be statistically meaningful with respect to the test chosen for evaluating the strength of the link between G_i and each core (different measures have been proposed and used in the commercial software for symbolic marking, e.g. in SPAD a *test-value* based on a hyper-geometric distribution).

Thus the description is given by:

$$G_i : g_1 \vee g_2 \vee \ldots g_n$$

Two measures are associated to each marking core (together with the *test value*):

DEBOR = % of g_k's individuals belonging to g_k, but not to G_i

REC = % of G_i individuals described by g_k

The value of REC(g_k) can be cumulated (RECCUM) over the different cores and used to decide the number of cores to be considered.

Dealing with documents, we have a set of tagged documents described by its graphical forms, which play the role of the characterising variables, y_j.

4.2.2. *The extracted rules*

Applying the SPAD procedure for SM on the training set the following logical rules have been extracted.
Sentences containing skill requirements can be detected, by the rule:

conoscenza
OR (**esperienza** ***AND NOT*** **inquadramento**)
OR (**anni** ***AND NOT*** **responsabile** ***AND NOT*** **risorse**)
OR (**capacità** ***AND NOT*** **inquadramento**)

Sentences not containing skill requirements can be detected, by the rule:

***NOT* conoscenza *AND* (*NOT* anni) *AND* (*NOT* capacità)**

4.3. *The second step: discarding uninteresting sentences*

Once the logical rules have been identified, by means of macros built in UltraEdit software environment, only the sentences containing skill requirements have been selected.

A big reduction of the corpus size has been obtained, allowing a much faster computation in the subsequent analysis. From the original corpus containing 305,853 occurrences it has been extracted a sub-corpus (made of the original documents swept from the excluded sentences) containing 83,062 occurrences; circa 27% of the original corpus has therefore been selected.

4.4. *Knowledge extraction: a comparison of typologies*

The final aim of this analysis was to identify skill requirements in job offers. Once the text has been cleaned from all the redundant information, Text Mining techniques give results specifically tuned on the analysis scope. Here, a clustering procedure has been used in order to identify typologies of skills requirements. Typologies have been compared with the ones obtained on the original corpus.

Two lexical tables have been built: one on the original collection and the other one on the reduced collection. Following a common strategy in textual data analysis, preliminarily a Correspondence analysis has been performed on the two tables, and coordinates of documents on the first 10 factorial axes have been the input of the following clustering procedure.

By comparing results, the effectiveness of our proposal appears.

In the following table 1 the 5 classes identified in the full corpus are described. The optimal number of clusters is automatically identified by an index based on information loss.

Class 1 (74.9%)	**Class 2 (8.5%)**	**Class 3 (0.5%)**	**Class 4 (4.4%)**	**Class 5 (11.7%)**
Esperienza	Sede	Consulting	Master	Posti
Anni	Milano	Milano	Corso	Durata
Dati	Leader	Sede	Stage	Svolgimento
Curriculum	Cliente			
Capacità	Azienda			

Table 1. Typology on the full corpus

The clusters identified on the full collection are mainly described by information not useful for the analysis aim. Only the first cluster is partially described by words related to skills (*esperienza, capacità*). Classes 2 and 3 contain company descriptions and classes 4 and 5 descriptions of masters, internships and courses without a precise separation among these different offers.

Class 1 (85.5%	Class 2 (0.2%)	Class 3 (12.0%)	Class 4 (2.3%)
Esperienza	Esperto	Word	Dinamici
Anni	Relazioni	Excel	Vendita
Titolo	Industriali	Internet	Agenti
Disponibilità	Milano	Access	Obiettivi
Doti		Stage	Chimica
		Laurea	

Table 2. Typology on reduced corpus

The typology of reduced collection (Table 2) describes four skill groups: the first is a group mainly characterized by *esperienza*. This class describes the skills required to experienced workers; these skills are very heterogeneous even if they are all characterized by experience in some working field further described with other words in the job offer. Being this group also quite large a deeper investigation is needed. The other three describe more defined skill profiles, respectively: industrial relations experts, internship candidates and salesmen. In particular we see that internship candidates are required to have a university degree and to have basic computer knowledge. To salespersons is instead required to be dynamic and to work by objectives. This typology, differently from the first one reaches the analysis objectives describing the skills required. An overview of the skills can be obtained by a finer clustering.

In the 9 class typology of Table 3 we get a complete overview of the different skills required by the job market. Classes 2,3,4 of the previous typology correspond to classes 7,8 and 9 of the new one, while class 1 is split in 6 new classes. The first describes a general profile for which knowledge, computer, foreign languages, dynamism and interpersonal skills are important. Class 2 describes jobs in the field of distribution for which are important degree, residence and interpersonal skills. Class 3 describes the graduates in economics for which age, experience and leadership are important. In Class 4 we find engineers profiles to which projecting and drawing skills are required. In class 5 there are profiles of professional salesman with motivation and experience. Class 6 is a residual group of advertisements offering computer courses to workers.

5. Conclusions

Here we have proposed a methodology, based on the principles of Text Classification, meant to extract interesting (in a specific domain) patterns of words. In this proposal there is a first attempt to go beyond the traditional bag-of-words as we encode sentences and not the whole documents. In this way, some sequential boundaries are built, as we are not interested in the general contexts in which words are used but in the local contexts, conveying the specific information of interest. This is a consequence of the fact that most of the times, documents are not completely unstructured; let's think about scientific articles (abstract, introduction, proposal, conclusion). This structural division can be more or less evident. Aim of this strategy is to exploit the structure embedded in the document and improve Text Mining task performances. Job offers are a very well suited example of semi-structured documents and we used an Italian on line job offers collection in order to show the effectiveness of our proposal, in clustering skill requirements.

Job advertisements are usually composed by the description of the job profile, the required skills, and further information (e.g. selection process, retribution, contacts). In the first step of our procedure we dealt with a tagged training set, which differently labeled skill requirement sentences. This first step can be viewed as a pre-processing step, aiming at identifying the words, and the association among words, which discriminate sentences carrying information about the selected topic. Its output are rules, validated on a test set. Then, rules are applied to the entire corpus, in order to discard not interesting parts. A reduced corpus is produced, input for further mining processes. The extracted subtexts in fact have less variability and contain less noise. This leads to better performances of visualization, retrieval and clustering techniques, as shown by the job offers.

Class 1 (25%)	**Class 2 (1,9%)**	**Class 3 (35,1%)**	**Class 4 (10,6%)**	**Class 5 (12,2%)**
Conoscenza	Diploma	Laureato	Meccanica	Professionisti
Office	Residenza	Economia	Disegno	Vendita
Lingue	Distribuzione	Esperienza	Tecnico	Laureati
Windows	Interpersonali	Giovane	Ingegnere	Motivazione
Interpersonali		Leadership	Automotive	Esperienza
Dinamismo		Responsabilità	Progettazione	Agenti

Class 6 (0,9%)	**Class 7 (0,2%)**	**Class 8 (11,9%)**	**Class 9 (2,2%)**
Corso	Esperto	Word	Dinamici
Windows	Relazioni	Excel	Vendita
Diplomati	Industriali	Internet	Agenti
Laureati	Milano	Stage	Obiettivi
		Access	Chimica

Table 3. Typology on reduced corpus with nine clusters

References

Balbi S. and Di Meglio E. (2003). Text Mining on-line job offers. *Bulletin of the International Statistical Institute, 54th session,* vol. (60/1): 65-67.

Balbi S. and Gettler-Summa M. (2001). Identifying Lexical Profiles by Symbolic Marking. In *Book of Short Papers CLADAG2001*, Palermo: 185-188.

Gettler-Summa M. (1998). *MGS in SODAS: Marking and Generalization by Symbolic Objects in the Symbolic Official Data Analysis Software*. Cahier 9935. Université Dauphine LISE CEREMADE.

Gordon A.D. (1999). *Classification*. Chapman & Hall CRC.

Hand D., Mannila H. and Smyth P. (2001). *Principles of Data Mining.* MIT Press.

Lebart L., Salem A. and Berry L. (1998). *Exploring Textual Data*. Kluwer Academic Publishers.

Manning C.D. and Schütze H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Sebastiani F. (2002) Machine Learning in automated Text Categorization. *ACM Computing surveys,* vol. (34/1): 1-47.