Text Retrieval with External Information

Silvano Amato, Emilio Di Meglio, Maria Guerra

Dipartimento di Matematica e Statistica, Università "Federico II"
Via Cintia, Monte Sant'Angelo
80126 – Naples – Italy
{silamato, edimegli, mguerra}@unina.it

Abstract

Aim of this paper is to propose a statistical strategy based on the Partial Least Squares Regression (PLSR) in order to exploit external information in Text Mining processes. We focus on Text Retrieval, a procedure aimed at finding interesting information in large textual collections. In order to exploit external information, often available in collections, we assume a dependence structure between two sets of textual variables. By means of PLSR a latent structure taking into account word similarities and external information is obtained. Projection of documents and queries on the obtained latent structure allows effective document retrieval. The suggested strategy is applied to data which consist of two sets of variables: words of journal article abstracts and their respective keywords. The investigated approach enables more effective documents retrieval, as it takes into account the external information given by the keywords.

Keywords: PLS Regression, text mining, text retrieval, LSI.

1. Introduction

Text Mining can be considered as a process aiming at uncovering interesting information in large, non structured textual *corpora*. Text Retrieval is an important task of Text Mining; aim of this technique is to search for a specific piece of information of a specific topic in large textual repositories according to characteristic aspects of their content.

In text repositories, in addition to documents, we usually find information that is discarded by commonly used Text Mining strategies. This external information, if properly considered, could drastically improve performances. For example, in a text repository containing article abstracts, there is information such as keywords, authors or sources that could be very useful either for retrieval either for other Text Mining tasks. Text Mining tasks, in fact, are based on evaluating similarities among documents; these similarities are based on the meanings carried by the documents, and, external information can improve meanings extraction from documents.

In this paper we deal with document vectors arranged, following the *bag of word* encoding, in a lexical table, $\mathbf{F}^*$; a generic element f^*_{ij} of $\mathbf{F}^*$ is the relative frequency of occurrence of word j in document i. The external information is instead encoded as an indicator matrix $\mathbf{Y}$, that cross-tabulates the same documents of matrix $\mathbf{F}^*$ with the words representing the external information. Matrices $\mathbf{F}^*$ and $\mathbf{Y}$ are shown in figure 1. Usually lexical tables are highly dimensional and sparse, therefore a common step in retrieval problems is dimensionality reduction; this is usually performed by means of Singular Value Decomposition (SVD) based methods applied only to matrix $\mathbf{F}^*$.

Here we suggest a strategy for the analysis of lexical tables that takes into account external information. In this framework, non symmetrical relationships between the two groups of variables

$$\mathbf{F}^* = \begin{array}{c} \textbf{Abstract words} \\ \begin{bmatrix} f^*_{11} & f^*_{12} & \cdots & \cdots & f^*_{1p} \\ f^*_{21} & & & & f^*_{2p} \\ & & f^*_{ij} & & \\ \vdots & & & & \vdots \\ f^*_{n1} & & \cdots & & f^*_{np} \end{bmatrix} \end{array} \qquad \mathbf{Y} = \begin{array}{c} \textbf{Keywords} \\ \begin{bmatrix} 1 & 0 & \cdots & 1 \\ & & 1 & \\ 0 & & \vdots & 0 \\ \vdots & & 0 & \vdots \\ & \cdots & & \\ 1 & & & 1 \end{bmatrix} \end{array}$$

Figure 1. Matrices $\mathbf{F}^*$ *and* $\mathbf{Y}$.

is supposed.

The first group of variables (*independent*), arranged in matrix $\mathbf{F}^*$ is constituted by the words observed in texts; the second group (*dependent*), $\mathbf{Y}$, can be formed by the vocabulary of the external information.

The matrix $\mathbf{F}^*$ is weighted by marginal distributions of rows and columns so to consider chi-square metric. This is done in order to cope with two problems related to word frequency data, namely: different length of documents and triviality of high frequency words (Di Meglio, 2003). In fact, a word that appears only one time in a very short document cannot be considered less important than a word appearing 100 times in a 300 pages book. In most cases, furthermore, rare words are more informative than high frequency ones.

Let $\mathbf{D}_r^{-1}$ and $\mathbf{D}_c^{-1}$ be the diagonal matrices of the marginal row and column distribution, the weighted matrix $\mathbf{F}$ is:

$$\mathbf{F} = \mathbf{D}_r^{-1}\mathbf{F}^*\mathbf{D}_c^{-1} \tag{1}$$

The modeling of the non symmetrical relationship between $\mathbf{F}$ and $\mathbf{Y}$ and dimensionality reduction are jointly achieved by means of Partial Least Squares Regression (PLSR) (Wold *et al.*, 1984) described in section 2.

The PLSR Retrieval strategy, here proposed, performs retrieval in the subspace spanned by the columns of matrix $\mathbf{Y}$, in which the units (documents) of matrix $\mathbf{F}$ are projected.

The subspace generated by the PLSR components can also be effectively used for documents clustering. In this subspace, in fact, similarities among documents are based on semantic similarities and external information. In this work, indeed, we concentrate on Text Retrieval.

2. PLS Regression

PLS Regression has been introduced by Svante Wold *et al.* (1984) in order to model the non symmetrical relationship between two groups of variables with aim of maximizing predictive power of the model and to cope with multicollinearity among variables. From the latter perspective, PLS is an alternative to Principal Components Regression (PCR) (Massy, 1965); literature (e.g. Næs and Martens, 1985; de Jong, 1993) shows that PLS leads to more parsimonious model.

PLS Regression can be interpreted as a penalized Canonical Correlation Analysis (CCA), with a PCA in the $\mathbf{F}$ space and a PCA in the $\mathbf{Y}$ providing the penalties. Namely, Höskuldsson (Höskuldsson, 1988) shows that in the first step, PLS algorithm performs SVD of $\mathbf{F}'\mathbf{Y}$ (if $\mathbf{F}$ is the regressor matrix and $\mathbf{Y}$ is response the one).

Let r be the rank of $\mathbf{F}$; we define $\mathbf{F}_0 = \mathbf{F}$ and $\mathbf{Y}_0 = \mathbf{Y}$. First pair of PLS components $\mathbf{t}_1 = \mathbf{F}_0\mathbf{w}_1$ and $\mathbf{u}_1 = \mathbf{Y}_0\mathbf{c}_1$ are such that

$$\begin{cases} \max\limits_{\mathbf{w}_1,\mathbf{c}_1} \text{cov}(\mathbf{F}_0\mathbf{w}_1, \mathbf{Y}_0\mathbf{c}_1) \\ \mathbf{w}_1'\mathbf{w}_1 = 1 \\ \mathbf{c}_1'\mathbf{c}_1 = 1 \end{cases}$$

where cov$(\cdot,\cdot)$ stands for covariance. We now denote residual matrix $\mathbf{F}_1 = \mathbf{F}_0 - \mathbf{t}_1\mathbf{p}_1$ and $\mathbf{Y}_1 = \mathbf{Y} - \mathbf{t}_1\mathbf{c}_1$, where $\mathbf{p}_1 = \mathbf{F}_0'\mathbf{t}_1/\mathbf{t}'\mathbf{t}_1$ and $\mathbf{c}_1 = \mathbf{Y}'\mathbf{t}_1/\mathbf{t}'\mathbf{t}_1$. Second pair of PLS components $\mathbf{t}_2 = \mathbf{F}_1\mathbf{w}_2$ and $\mathbf{u}_2 = \mathbf{Y}_1\mathbf{c}_2$ are such that

$$\begin{cases} \max\limits_{\mathbf{w}_2,\mathbf{c}_2} \text{cov}(\mathbf{F}_1\mathbf{w}_2, \mathbf{Y}_1\mathbf{c}_2) \\ \mathbf{w}_2'\mathbf{w}_2 = 1 \\ \mathbf{c}_2'\mathbf{c}_2 = 1 \\ \mathbf{w}_1\mathbf{F}_0'\mathbf{F}_1\mathbf{w}_2 = 0 \end{cases}$$

Last constraint in the above system ensures $\mathbf{t}_1$ and $\mathbf{t}_2$ to be orthogonal to each other. A generic pair h of PLS component is given by solving

$$\begin{cases} \max\limits_{\mathbf{w}_h,\mathbf{c}_h} \text{cov}(\mathbf{F}_{h-1}\mathbf{w}_h, \mathbf{Y}_{h-1}\mathbf{c}_h) \\ \mathbf{w}_h'\mathbf{w}_h = 1 \\ \mathbf{c}_h'\mathbf{c}_h = 1 \\ \mathbf{w}_h'\mathbf{F}_{h-1}'\mathbf{F}_{j-1}\mathbf{w}_j = 0, \forall j < h \end{cases}$$

Columns of matrix $\mathbf{W}_{(h)}$, $w_1, \mathbf{w}_2, \ldots, w_h$, give PLS components $\mathbf{t}$ by means of residual matrix $\mathbf{F}_{h-1}$; in order to compute the $\mathbf{t}$'s by means of original matrix $\mathbf{F}$, we use a transformation of $\mathbf{W}_{(h)}$:

$$\widetilde{\mathbf{W}}_{(h)} = \mathbf{W}_{(h)}\left(\mathbf{P}_{(h)}'\mathbf{W}_{(h)}\right)^{-1}$$

At each step $h = 1, 2, \ldots, r$, PLS regression maximizes covariance between components of residuals matrices $\mathbf{F}_h$ and $\mathbf{Y}_h$; this corresponds to the inter–battery analysis by Tucker (Tenenhaus, 1998; Tucker, 1958) of matrices $\mathbf{F}_h$ and $\mathbf{Y}_h$; that is, because cov$(\mathbf{F}_{h-1}\mathbf{w}_h$, $\mathbf{Y}_{h-1}\mathbf{c}_h)$ can be expressed as the product of correlation between $\mathbf{F}_{h-1}\mathbf{w}_h$ and $\mathbf{Y}_{h-1}\mathbf{c}_h$ and $\sqrt{\text{var}(\mathbf{F}_{h-1}\mathbf{w}_h)\text{var}(\mathbf{Y}_{h-1}\mathbf{c}_h)}$. We can see PLS regression as a compromise between canonical correlation analysis (maximum correlation between $\mathbf{t}_h$ and $\mathbf{u}_h$) and OLS regression on principal component analysis (maximum variance of $\mathbf{t}_h$ and $\mathbf{u}_h$).

A fundamental issue in PLS Regression is selection of the number of components to include in the model by means of which we can tune estimates efficiency and distortion; most applied selection method is Cross Validation (e.g. Stone, 1990).

3. Text Retrieval

The classical problem in Text Retrieval is the search for a specific piece of information of a specific topic in large document repositories. In practice, using this methodology, an user should be able to retrieve the relevant documents given a certain natural language query.

A standard Text Retrieval method builds an index of documents and gives the user the possibility to perform searches in this index by formulating queries. Queries are usually formulated in natural language and express the concept the user wishes to retrieve. The system should then be able to compare the concept expressed in the query with all the documents, rank the documents in order of relevance and give back to the user the n most relevant ones. Text retrieval deals with retrieving contents, but contents can be expressed in many different ways using different words. A retrieval system should therefore be able to extrapolate concepts from documents and assess their similarities with the queries.

Text retrieval strategies

Three of the main strategies for retrieving documents from huge textual databases will now be presented: Boolean Model, Vector Space Model (VSM) and Latent Semantic Indexing (LSI). All these methods, except Boolean Model which is the most simple, are based on the *bag of words* documents encoding.

Boolean search is very similar to a search process in a classical database. In Boolean retrieval the system selects the documents that satisfy a logical expression of query terms using boolean operators such as AND, OR, NOT. This method is usually applied on textual repositories that have already been manually indexed with keywords; in this case queries are boolean expressions of keywords. Boolean model suffers from some serious drawbacks (Salton *et al.*, 1983): e.g., the number of retrieved documents depends on the frequency of the terms used in the query and on the used boolean operators.

Vector Space Model (VSM) is widely used in Text Retrieval mainly because it is easy to implement and is conceptually simple. Proximities among documents are in fact assumed to be similar to proximities in a multidimensional space and this makes possible to use statistical models for building retrieval systems. In VSM each document and each query are represented by vectors of term weights (Salton, 1989). In VSM Text Retrieval is performed by measuring the similarity between the query vector and the document vectors. This similarity can be measured in terms of Euclidean distance or in terms of angle between vectors. The documents that present a larger similarity or a smaller angle to the query are retrieved as relevant. Also, this scheme presents some drawbacks: it does not allow to consider synonymous and then dependencies.

Latent Semantic Indexing (LSI), introduced by Dumais *et al.* (Dumais *et al.*, 1988; Deerwester *et al.*, 1990) is a technique aiming at extracting the hidden semantic structures from texts by means of Singular Value Decomposition (SVD). The technique consists in projecting queries and documents in a space with "latent" semantic dimensions determined using SVD. The latent dimensions derive from the co-occurrence patterns among words. In this way a query and a document can have a high similarity even if they don't share any term (Manning, 2001) as long as the terms are semantically similar according to the LSI analysis. The assumption at the basis of LSI model is that there is an underlying latent semantic structure in word usage that is obscured by noise and by variability of word choice (Dumais *et al.*, 1988). SVD is used to capture the significant information and discard the noise.

LSI is based on Euclidean distance and this distance is considered not adequate for count data contained in lexical tables. Thus document vectors are not well represented in the latent space built by simply calculating SVD of matrix $\mathbf{F}^*$. This procedure, in fact, could be adequate for tables of continuous measurements; in a lexical table, instead it would give the same importance to a term appearing the same number of times in documents of extremely different lengths. In order to cope with these problems it has been proposed to use chi-square metric for text retrieval with LSI (Di Meglio, 2003).

4. Using external information in text retrieval

Text Retrieval is usually performed on the actual texts without considering any other information. A human user, when assessing the similarity between two documents or the relevance of a document uses also external information or metadata to support his/her decision. This information could be given by keywords, abstract, author, source and so on. Textual corpora, in fact, are formed by the actual documents and by other elements that can be explicit: titles, authors, keywords, classifiers, headers, or implicit: collocation, source, time and occasion of issue etc. This information, if properly considered, could drastically improve the performance of a retrieval system. Here we suggest a strategy that aims at exploiting external information available in the data.

PLS Regression is adopted to model the non symmetrical relation between the two groups of variables introduced in section 1. As described in section 2, the first step of PLS performs a SVD of $\mathbf{F}'\mathbf{Y}$. Being $\mathbf{Y}$ an indicator matrix, non zero elements of $\mathbf{F}'\mathbf{Y}$ represent co–presence of keywords and documents terms.

Proposed retrieval strategy is carried out in three steps:

- PLS Regression of $\mathbf{F}$ on $\mathbf{Y}$; computation of PLS components $(\mathbf{T})$ and selection of relevant components by means of the Q^2 index (Tenenhaus, 1998).
- The projection of documents $(\mathbf{t}_i)$ and query $(\mathbf{z}_q)$ vectors on the built subspace, are given by:

$$\mathbf{z}^q = \mathbf{v}^q \widetilde{\mathbf{W}}_{(h)} \tag{2}$$

$$\mathbf{t}^i = \mathbf{f}^i \widetilde{\mathbf{W}}_{(h)} \tag{3}$$

  $\mathbf{z}^q$ $(q = 1, 2, \ldots,$ number of queries), and $\mathbf{t}^i$ $(i = 1, 2, \ldots, n)$ are row vectors with h elements, where h is the selected number of components:

- Retrieval of relevant documents to query q, by means of the cosine of the angle between the projected query vector $\mathbf{z}^q$ and all projected documents vectors $\mathbf{t}^i$:

$$d_q^i = \frac{\mathbf{z}^{q\prime}\mathbf{t}^i}{||\mathbf{z}^q||\,||\mathbf{t}^i||} \tag{4}$$

  where $||\cdot||$ is the quadratic norm. For each query q, the distances d_q^i obtained from (4) are sorted in ascending order and, the first k documents are retrieved as the most relevant to query q (i.e. the closest), according to the number of documents required by the user.

5. Application to OHSUMED data

Proposed strategy has been applied to a dataset extracted by the OHSUMED collection (available from ftp://medir.ohsu.edu in the directory /pub/ohsumed) compiled by William Hersh this collections consists of medical journal abstracts. For each abstract author, keywords and journal title are provided.

We randomly selected a subset of 963 documents with respective keywords and used the latter as external information. 10 queries were built and relevance of documents to these queries has been assessed by three independent judges.

PLS has been run and first 36 components, yielding 33.3% of explained inertia, have been retained according to the Stone–Geisser Q^2 index.

The proposed method has been compared with LSI and LSI using chi-square metric (LSCI) on the matrix $\mathbf{X}$, and with LSI on matrix $[\mathbf{X}|\mathbf{Y}]$; the performances of the different methods has been measured by means of the Precision Index (Manning and Schütze, 2001). In order to make fair comparisons among the different methods we retained for each method applied a number of components such that achieve 33.3% of inertia. Table 1 shows the retained numbers of components.

	Components Number
PLS	36
LSI $\mathbf{X}$	81
LSCI	81
LSI $[\mathbf{X}\|\mathbf{Y}]$	100

Table 1. Number of components needed to achieve 33.3% of total inertia.

Precision Index measures the proportion of documents that the system got right and ranges between 0 and 1. Precision has been calculated, to take into account the ranking, at different *cutoff* points, that is, considering different numbers of retrieved documents. Usual cutoff points are 5, 10, 20 documents. We have used cutoff 1,2,. . . ,10. In figure 2 precision plot of the four used methods is shown. PLS retrieval globally outperforms the other methods. In fact PLS at cutoffs 1 to 8 performs better or at least equally than other methods. For cutoffs 9 and 10 LSI on $[\mathbf{X}|\mathbf{Y}]$ performs best, but this method performs poorly at lower cutoffs. In any case, PLS ranks the relevant documents always in the first positions.

In particular, after a deeper examination of queries and retrieved documents it appears that LSI and LSCI tend to give higher rank to documents in which the words used in the query have high occurrence. PLS retrieval strategy instead retrieves documents more similar to the query in terms of *meanings*, even if they don't share words with the query. This happens because also keywords are taken into account in the proposed retrieval strategy.

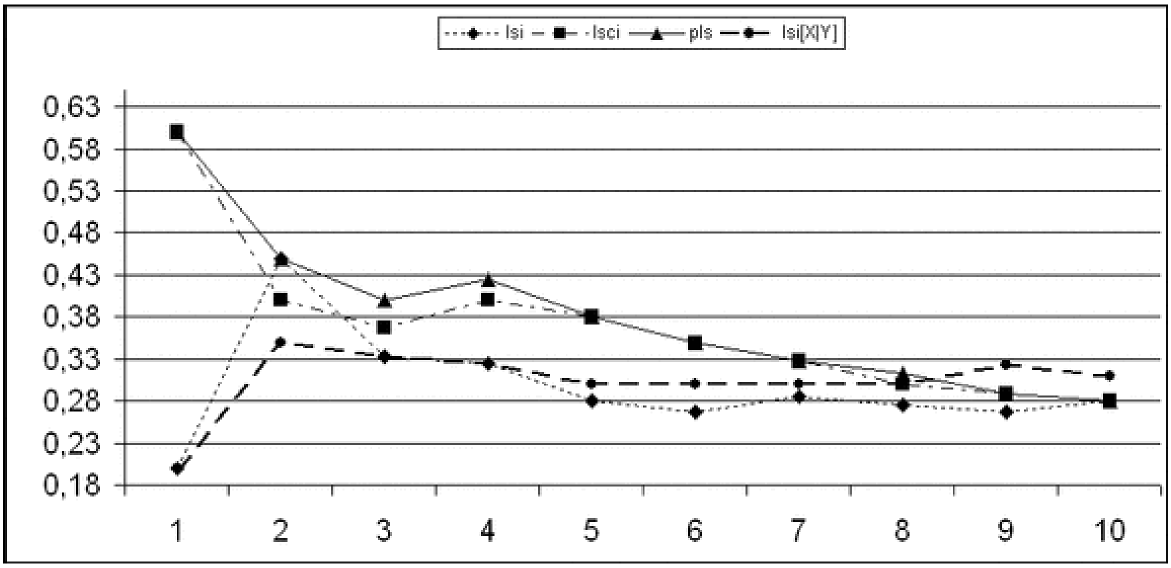

Figure 2. Precision Plot

6. Conclusions and perspectives

On the used data PLS retrieval outperforms both LSI and LSCI. This is due to the use of external information considered in the retrieval strategy.

Massive experiments need however to be performed in order to ascertain pros and cons of the proposed methodology, namely in order to describe the situations in which PLS retrieval is more appropriate than other methods. Further research is needed in order to understand the relationship between the distances computed in the projected documents and their distances in the original space.

Dimensionality reduction obtained by means of PLS can be easily exploited also for clustering problems. Typologies obtained in this manner use in fact external information to classify documents and can lead to more effective cluster identification.

The interpretation of components is also possible; this task can be carried out by means of the projection of words of both matrices $\mathbf{F}$ and $\mathbf{Y}$ on the same subspace. The scatter plot of these projection can be easily read by looking at Euclidean distances in the plane as well as at correlations between words.

Future research directions will include clustering with external information.

Acknowledgment

The authors wish to thanks the unknown referees for their useful suggestions.

References

Balbi S. and Giordano, G. (2000). Un'analisi dei dati testuali con informazioni esterne: le definizioni di "qualità". In *Proceedings of JADT 2000.*

Barker M. and Rayens W. (1999). Partial Least Squares for Discrimination. In *Nonlinear Models Conference* at UK.

Deerwester S., Dumais S., Furnas G., Landauer T. and Harshman R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, vol. (41/6): 391-407.

de Jong S. (1993). PLS fits closer than PCR. *Journal of Chemometrics*, vol. (7): 551-557.

Di Meglio E. (2003). Improving Text Retrieval with Latent Semantic Indexing using Correspondence Analysis. In *Atti del Convegno SIS.*

Dumais S.T., Furnas George W., Landauer Thomas K., Deerwester S. and Harshman R. (1998). Using latent semantic analysis to improve access to textual information. In *Proceedings of the Conference on Human Factors in Computing Systems CHI'88.*

Fisher R.A. (1936). The use of Multiple Measurement in taxonomic problems. *Ann. Eugen.*, vol. (7): 179-188.

Höskuldsson A. (1988). PLS Regression methods. *Journal of Chemometrics*, vol. (2): 211-228.

Manning C.D. and Schütze H. (2001). *Foundations of Statistical Natural Language Processing.* MIT Press.

Massy W.F. (1965). Principal Components Regression in exploratory statistical research. *Journal of the American Statistical Association*, vol. (60): 234-246.

Næs T. and Martens H. (1985). Comparison of prediction methods for multicollinear data. *Communication in Statistics — Simulation and Computation*, vol. (14/3): 545-576.

Salton G., Fox E. and Wu H. (1983). Extended boolean information retrieval. In *Communication of the ACM*, vol. (26/11): 1022-1036.

Salton G. (1989). *Automatic Text Processing.* Addison Wesley.

Stone M. and Brooks R.J. (1990). Continuum Regression: Cross–validated sequencially constructed prediction embarcing Ordinary least Squares, Pratial least Squares and Principal Component Regression. *J. R. Statist. Soc., B*, vol. (52/2): 237-269.

Stone M. (1974). Cross–validatory choice and assesment of statistical prediction. *J. R. Statist. Soc., B*: 237-269.

Tenenhaus M. (1978). *La Régression PLS, Théorie et Pratique*. Éditions Technip.

Tucker L.R. (1958). An inter-battery method of factor analysis. *Psychometrika*, vol. (23/2): 111-136.

Wold S., Rube R., Wold H. and Dunn W. (1984). The collinearity problem in linear regression.The PLS approach to generalized inverses. *SIAM Iorunal of Sci. Stat. Comput.*, vol. (5): 735-743.