Analysing part-of-speech for Portuguese Text Classification

Teresa Gonalves[1], Cassiana Silva[2], Paulo Quaresma[1], and Renata Vieira[2]

[1] Dep. Informática, Universidade de Évora, 7000 Évora, Portugal
tcg,pq@di.uevora.pt
[2] Unisinos, CEP 93.022-000 So Leopoldo, RS, Brasil
cassiana,renata@exatas.unisinos.br

Abstract. This paper proposes and evaluates the use of linguistic information in the pre-processing phase of text classification. We present several experiments evaluating the selection of terms based on different measures and linguistic knowledge. To build the classifier we used Support Vector Machines (SVM), which are known to produce good results on text classification tasks.
Our proposals were applied to two different datasets written in the Portuguese language: articles from a Brazilian newspaper (Folha de So Paulo) and juridical documents from the Portuguese Attorney General's Office. The results show the relevance of part-of-speech information for the pre-processing phase of text classification allowing for a strong reduction of the number of features needed in the text classification.

1 Introduction

Machine learning techniques are applied to document collections aiming at extracting patterns that may be useful to organise or retrieve information from large collections. Tasks related to this area are text classification, clustering, summarisation, and information extraction. One of the first steps in text mining tasks is the pre-processing of the documents, as they need to be represented in a more structured way to be fed to machine learning algorithms. In this step, words are extracted from the documents and, usually, a subset of words (stop words) is not considered, because their role is related to the structural organisation of the sentences and does not have discriminating power over different classes. This shallow and practical approach is known as bag-of-words. Usually, to reduce semantically related terms to the same root, a lemmatiser is applied.

Finding more elaborated models is still a great research challenge in the field; natural language processing increases the complexity of the problem and these tasks, to be useful, require efficient systems. Our proposal considers that there is still lack of knowledge about how to bring natural language and traditionally known techniques of data mining tasks together for efficient text mining. Therefore, here we make an analysis of different word categories (nouns, adjectives, proper names, verbs) for text mining, and perform a set of experiments of text classification over Brazilian and European Portuguese data. Our goal is to investigate the use of linguistic knowledge in text mining.

As classifier we used Support Vector Machines (SVM), which are known to be good text classifiers [8]. Other learning algorithms have been also applied such as decision trees [16], linear discriminant analysis and logistic regression [13], and naïve Bayes algorithm [10].

A method for incorporating natural language processing into existing text classification procedures is presented in [1] and a study of document representations based on natural language processing in four different corpora and two languages (English and Italian) is reported in [11]. Although they strongly claim against the union of NLP and text mining their experiments present just a few combinations of linguistic information. We believe that there is still much space for research in this area, and in this paper we show some interesting results of text classification regarding the simple linguistic knowledge of word categories.

In [6], SVM performance is compared with other Machine Learning algorithms and in [7] a thorough study on some preprocessing techniques (feature reduction, feature subset selection and term weighting) is made over European Portuguese and English datasets. The impact of using linguistic information on the preprocessing phase is reported in [15] over a Brazilian dataset.

This paper is organised as follows: in Section 2, a description of the used techniques and datasets is presented while Sections 3 and 4 describe the experiments. Conclusions and future work are pointed out in Sections 5 and 6.

2 Methods and Materials

In this section we describe the Support Vector Machines paradigm, the natural language tools applied for pre-processing the documents, the datasets studied and, at the end, the experimental setup is explained.

2.1 Support Vector Machines

Support Vector Machines (SVM) is a learning algorithm introduced by Vapnik and coworkers [4], which was motivated by the theoretical results from the statistical learning theory. It joins a kernel technique with the structural risk minimisation framework. A *kernel technique* comprises two parts: a module that performs a mapping into a suitable feature space and a learning algorithm designed to discover linear patterns in that space.

The *kernel function*, that implicitly performs the mapping, depends on the specific type and domain knowledge of the data source. The *learning algorithm* is general purpose and robust; it's also efficient, since the amount of computational resources required is polynomial with the size and number of data items, even when the dimension of the embedding space grows exponentially [14]. Key aspects of the approach can be highlighted as follows (ilustrated in Figure 1):

- Data items are embedded into a vector space called the feature space.

- Linear relations are discovered among images of data items in feature space.
- Algorithm is implemented in a way that the coordinates of the embedded points are not needed; only their pairwise inner products.
- Pairwise inner products can be computed efficiently directly from the original data using the kernel function.

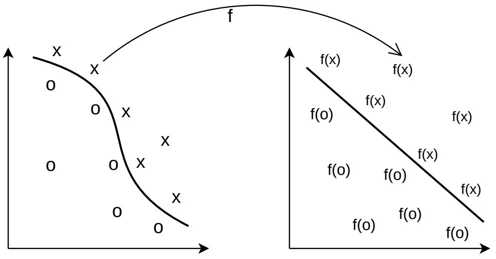

Fig. 1. Kernel function: Data's nonlinear pattern transformed into linear feature space.

The *structural risk minimisation* (SRM) framework creates a model with a minimised VC (Vapnik-Chervonenkis) dimension. This developed theory[17] shows that when a model's VC dimension is low, the expected probability of error is also low, which means good performance on unseen data.

2.2 Natural language processing tools

We applied a Portuguese stop-list (set of non-relevant words such as articles, pronouns, adverbs and prepositions) and POLARIS, a lexical database [9], to generate the lemma for each Portuguese word.

The POS tags were obtained through the syntactic analysis performed by PALAVRAS [2] parser, which was developed in the context of the VISL project (Visual Interactive Syntax Learning – `http://www.visl.sdu.dk/`) in the Institute of Language and Communication of the University of Southern Denmark. Possible morpho-syntactic tags are:

- adjective (*adj*),
- adverb (*adv*),
- article (*det*),
- conjunction (*conj*),
- interjection (*in*),
- noun (*n*),
- numeral (*num*),
- preposition (*prp*),
- pronoun (*pron*),
- proper noun (*prop*) and
- verb (*v*).

Portuguese is a morphological rich language: while nouns and adjectives have 4 forms (two *genders* – masculine and feminine and two *numbers* – singular and plural), a regular verb has 66 different forms (two *numbers*, three *persons* – 1st, 2nd and 3rd and five *modes* – indicative, conjunctive, conditional, imperative and infinitive, each with different number of *tenses* ranging from 1 to 5).

PALAVRAS parser is robust enough to always produce an output even for incomplete or incorrect sentences (which might be the case for the type of documents used in text mining tasks). It has a comparatively low percentage of errors (less than 1% for word class and 3-4% for surface syntax)[3].

It's output is the syntactic analysis of each phrase and the POS tag associated with each word. For example, the morphological tagging of the phrase 'O Manuel ofereceu um livro ao seu pai./Manuel gave a book to his father.' is:

```
o [o] <artd> <dem> DET M S
Manuel [Manuel] PROP M S
ofereceu [oferecer] V PS 3S IND VFIN
um [um] <quant> <arti> DET M S
livro [livro] N M S
a [a] <prp>
o [o] <artd> <dem> DET M S
seu [seu] <pron-det> <poss> M S
pai [pai] N M S
```

2.3 Dataset Description

As already mentioned, we performed the experiments over two datasets: FSP, a Brazilian Portuguese dataset of newspaper articles from "Folha de São Paulo" and PAGOD – Portuguese Attorney General's Office Decisions, an European Portuguese dataset of juridical documents.

FSP dataset. FSP is a subset of the NILC corpus (Núcleo Inter-institucional de Linguística Computacional – http://www.nilc.icmc.usp.br/nilc/) containing 855 documents from the year of 1994.

These documents are related to five newspaper sections, each one having 171 documents: informatics, property, sports, politics and tourism. Since each document belongs to one of the five possible classes, we have a multi-class problem. From all documents, there are 19522 distinct words, and, on average, 215 running words (tokens) and 124 unique words (types) per document.

PAGOD dataset. On the other hand, PAGOD has 8151 juridical documents, represents the decisions of the Portuguese Attorney General's Office since 1940 and delivers 96 MBytes. All documents were manually classified by juridical experts into a set of categories belonging to a taxonomy of legal concepts with around 6000 terms. Each document is classified into multiple categories so, we have a multi-label classification task. Normally, it is solved by splitting into a set of binary classification tasks and considering each one independently.

For all documents, we found 68877 distinct words and, on average, 1608 tokens and 366 types per document. A preliminary evaluation showed that, from all potential categories only about 3000 terms were used and from all 8151 documents, only 6773 contained at least one word in all experiments. Table 1 presents the top ten categories (the most used ones) and the number of documents belonging to each one.

category (Portuguese)	category (English)	# docs
pensão por serviços excepcionais	excepcional services pension	906
deficiente das forças armadas	army injured	678
prisioneiro de guerra	war prisoner	401
estado da Índia	India state	395
militar	military	388
louvor	praise	366
funcionário público	public officer	365
aposentação	retirement	342
competência	competence	336
exemplar conduta moral e cívica	exemplary moral and civic behaviour	289

Table 1. PAGOD's top ten categories.

2.4 Experimental setup

Now we present the choices made in our study: the kind of kernel used, the representation of documents and the used measures of learners' performance.

The linear SVM was run using the WEKA [18] software package from Waikato University, with default parameters (complexity parameter equal to one and normalised training data) and performing a 10-fold cross-validation procedure.

To represent each document we chose the bag-of-words approach, a *vector space model* (VSM) representation: each document is represented by the words it contains, with their order and punctuation being ignored. From the bag-of-words we removed all words that contained digits.

Learner's performance was analysed through precision, recall and F_1 measures [12] of each category (obtained from contingency table of the classification – prediction *vs.* manual classification). For each one, we calculated the micro- and macro-averages and made significance tests regarding a 95% confidence level.

3 Baseline experiments

In this section, we first present the IR techniques used, the experiments made and the results obtained.

We considered three typical information retrieval preprocessing techniques: feature reduction/construction, feature subset selection and term weighting. For each technique, we considered several experiments as described below.

Feature Reduction/Construction. On trying to reduce/construct features we used some linguistic information: we applied a Portuguese stop-list and POLARIS to generate the lemma for each Portuguese word. We made three sets of experiments:

- rdt_1: consider all words of the original documents
- rdt_2: consider all words but the ones that belong to the stop-list (stop words)
- rdt_3: all words (except the stop words) are transformed into its lemma

Feature Subset Selection. For selecting the best features we used a filtering approach, keeping the ones with higher scores according to different functions:

- scr_1: *term frequency.* The score is the number of times the feature appears in the dataset; only the words occurring more frequently are retained;
- scr_2: *mutual information.* It evaluates the usefulness of an attribute by measuring the Mutual Information with respect to the class. Mutual Information is an Information Theory measure [5] that ranks the information received to decrease the uncertainty. The uncertainty is quantified through the Entropy measure.

For each filtering function, we tried different threshold values. This threshold is given by the number of times the word appears in all documents – thr_n means that all words appearing less than n times are eliminated. For each threshold we looked at the number of words retained and used it to select the features.

Term Weighting. Finally, for the term weighting experiments, we made two different experiments:

- wgt_1: uses $TF(w_i, d_j)$ normalised to unit length. $TF(w_i, d_j)$ is the number of times word w_i occurs in document d_j.
- wgt_2: *TFIDF representation.* It's $TF(w_i, d_j)$ multiplied by $log(N/DF(w_i))$, where N is the total number of documents and $DF(w_i)$ is the number of documents in which w_i occurs. The measure is normalised to unit length.

3.1 Experiments

For the `FSP` dataset, we performed experiments for all options of feature reduction/construction, scoring function and term weighting (rdt_1, rdt_2 and rdt_3; scr_1 and scr_2; wgt_1 and wgt_2) and tried the following threshold values: thr_1, thr_5, thr_{10}, thr_{20}, thr_{30},..., thr_{90}, totalling a number of 132 experiments.

For the `PAGOD` dataset, we performed experiments for rdt_2 and rdt_3 options of feature reduction/construction, scr_1 and scr_2 scoring functions and wgt_1 and wgt_2 term weighting schemes. We tried the threshold values thr_1, thr_{50}, thr_{100}, thr_{200},..., thr_{900} (88 experiments).

Table 2 presents the number of words ($\#words$) and per document averages of token (avg_{tok}) and type (avg_{typ}) for each feature reduction/construction setup and Table 3 shows the number of features obtained for each threshold value.

	FSP			PAGOD		
	$\#words$	avg_{tok}	avg_{typ}	$\#words$	avg_{tok}	avg_{typ}
rdt_1	19522	215	124	68877	1608	366
rdt_2	19352	134	100	68679	963	331
rdt_3	13317	128	91	42399	921	258

Table 2. Baseline experiments: number of words and averages for each dataset.

FSP	thr_5	thr_{10}	thr_{20}	thr_{30}	thr_{40}	thr_{50}	thr_{60}	thr_{70}	thr_{80}	thr_{90}
	4420	2315	1153	745	529	397	334	265	222	199
PAGOD	thr_{50}	thr_{100}	thr_{200}	thr_{300}	thr_{400}	thr_{500}	thr_{600}	thr_{700}	thr_{800}	thr_{900}
	9477	6435	4236	3226	2577	2198	1897	1678	1514	1369

Table 3. Baseline experiments: number of features for each threshold value.

3.2 Results

Table 4 presents the minimum, maximum, average and standard deviation of all experiments.

	FSP						PAGOD					
	μP	μR	μF_1	MP	MR	MF_1	μP	μR	μF_1	MP	MR	MF_1
min	.863	.863	.863	.865	.863	.864	.497	.742	.606	.491	.687	.559
max	.982	.982	.982	.983	.982	.982	.878	.789	.816	.799	.741	.758
avg	.947	.947	.947	.947	.947	.947	.837	.768	.800	.768	.716	.734
stdev	.026	.026	.026	.026	.026	.026	.043	.013	.023	.034	.015	.022

Table 4. Baseline experiments summarising values.

FSP dataset. From all 132 FSP experiments, there were 19 'best' ones with no significant difference for all six performance measures (precision, recall and F_1 micro- and macro-averages). The distribution of these experiments on each setup was the following:

- for rdt_1, rdt_2 and rdt_3 there were 4, 6 and 9 'best' experiments,
- for scr_1 and scr_2 there were 0 and 19 'best',
- for wgt_1 and wgt_2 there were 9 and 10 'best' and finally,
- for thr_5, thr_{10}, thr_{20}, thr_{30} and thr_{40} there were 6, 6, 4 2 and 1 'best' values.

From these results, one can say that the most suited setup is lemmatisation along with mutual information scoring function and TFIDF weighting scheme. The thr_{40} threshold is the one with less features from the set of the 'best' ones.

PAGOD dataset. Table 5 presents, for the `PAGOD` dataset, the number of experiments with no significant difference with respect to the best one and the distribution of these experiments on each setup (for example, macro-F_1 have 16 best experiments: 7 belong to the rdt_2 setup and 9 to the rdt_3 one).

One can say that both rdt_2 and rdt_3 produce similar results and that term frequency scoring function along with term frequency weighting scheme is the setup with best results. The thr_{800} threshold is the biggest with good results.

	μP	μR	μF_1	MP	MR	MF_1
$best$	5	22	34	14	21	16
rdt_2	2	10	14	4	7	7
rdt_3	3	12	20	10	14	9
scr_1	2	21	19	2	20	14
scr_2	3	1	15	12	1	2
wgt_1	1	16	26	10	16	13
wgt_2	4	6	8	4	5	3
thr_1	0	2	0	0	2	0
thr_{50}	0	2	0	0	2	0
thr_{100}	0	2	0	0	2	0
thr_{200}	0	4	1	0	3	0
thr_{300}	0	4	2	1	4	3
thr_{400}	0	2	3	1	2	2
thr_{500}	0	2	6	1	2	3
thr_{600}	0	2	6	1	2	3
thr_{700}	0	1	5	3	1	2
thr_{800}	0	1	5	3	1	2
thr_{900}	5	0	6	4	0	1

Table 5. Baseline `PAGOD` experiments: number belonging to the set of best results.

Table 6 shows the precision, recall and F_1 for the best setups of both datasets; the values that belong to the set of best ones are bold faced. From these figures we can say that $rdt_3.scr_1.wgt_1.thr_{800}$ is the best setup for the `PAGOD` dataset since it has more significant best values than the other one.

	μP	μR	μF_1	MP	MR	MF_1
`FSP`.$rdt_3.scr_2.wgt_2.thr_{40}$	**.975**	**.975**	**.975**	**.976**	**.975**	**.975**
`PAGOD`.$rdt_2.scr_1.wgt_1.thr_{800}$	.846	.772	**.807**	.776	.720	.743
`PAGOD`.$rdt_3.scr_1.wgt_1.thr_{800}$	.846	**.782**	**.813**	.782	**.732**	**.753**

Table 6. Baseline experiments: precision, recall and F_1 micro- and macro-averages for the best setups.

4 POS tag experiments

This section presents the POS tag experiments made and the results obtained. From all possible parser tags (see Section 2.2), we just considered *n*, *prop*, *adj* and *v*. We tried all possible combinations of these tags.

For both datasets, we made experiments for the best baseline setup and three more obtained by reducing the number of features through new threshold values – thr_{40}, thr_{50}, thr_{60}, thr_{70} for FSP and thr_{800}, thr_{900}, thr_{1000}, thr_{1100} for PAGOD, totalling a number of 60 experiments for each dataset.

PAGODs' thresholds thr_{1000} and thr_{1100} have 1259 and 1160 features, respectively. Table 7 presents the per document averages of token (avg_{tok}) and type (avg_{typ}) for each POS tag (number and percent).

The proportion of verbs is similar in both datasets, but FSP has 2 times more percentage of proper nouns than PAGOD. This could be a reason for the different best baseline setups obtained in the previous section.

	FSP				PAGOD			
	# avg_{tok}	# avg_{typ}	% avg_{tok}	% avg_{typ}	# avg_{tok}	# avg_{typ}	% avg_{tok}	% avg_{typ}
adj	11	9	9.8%	10.8%	115	41	14.4%	16.4%
nn	52	37	46.4%	44.6%	423	110	52.8%	44.0%
prop	26	18	23.2%	21.7%	90	27	11.2%	10.8%
vrb	23	19	20.5%	22.9%	173	72	21.6%	28.8%

Table 7. POS experiments: averages (number and percent) of token and type.

4.1 Results

We compared all 60 experiments along with the best setup obtained from the baseline experiments.

FSP dataset. For all six measures, there were 6 'best' experiments with no significant difference in the thr_{40} and thr_{50} thresholds. They were:

- for thr_{40}: $rdt_3.scr_2.wgt_2$ (baseline experiment), nn+prop, nn+adj+prop and nn+adj+prop+vrb
- for thr_{50}: nn+prop+vrb and nn+adj+prop+vrb.

From these, we can say that although we could not enhance the classifier using POS tags for selecting features, it was possible to reduce their number with no reduction on performance if we use nouns, proper nouns and verbs or these along with adjectives.

Table 8 shows the values of precision, recall and F_1 for the baseline experiment (529 features) and the best combinations of tags for the highest threshold value thr_{50} (397 features). Once again, bold faced figures have no significant difference with the best one obtained.

	μP	μR	μF_1	MP	MR	MF_1
baseline	**.975**	**.975**	**.975**	**.976**	**.975**	**.975**
nn+prop+vrb	**.965**	**.965**	**.965**	**.965**	**.965**	**.965**
nn+adj+prop+vrb	**.967**	**.967**	**.967**	**.968**	**.967**	**.967**

Table 8. POS FSP experiments: precision, recall and F_1 micro- and macro-averages for baseline and best setups.

PAGOD dataset. Table 9 presents, the number of experiments with no significant difference with respect to the best one and the distribution for each combination of POS tags experiments. The combinations adj, prop, vrb, adj+vrb and prop+vrb had no experiment in the set of best ones.

	μP	μR	μF_1	MP	MR	MF_1
best	2	14	36	5	10	19
baseline	0	1	1	0	1	1
nn	0	0	4	1	0	0
nn+adj	0	0	4	1	0	0
nn+vrb	0	0	4	0	0	3
nn+prop	0	0	4	2	0	1
adj+prop	2	0	0	0	0	0
nn+adj+prop	0	4	4	1	2	4
nn+adj+vrb	0	2	4	0	1	2
nn+prop+vrb	0	3	4	0	2	4
adj+prop+vrb	0	0	3	0	0	0
nn+adj+prop+vrb	0	4	4	0	4	4

Table 9. POS PAGOD experiments: number belonging to the set of best results.

From the Table, we can say, again, that POS tags do not enhance the classifier but help feature selection by reducing the number of needed features. The best results were obtained using nouns combined with two of the other POS tags.

Table 10 shows the values of precision, recall and F_1 for the baseline experiment and those POS tags combinations for the highest threshold value (thr_{1100}) with results in the best set. We can say that nn+adj+prop and nn+adj+prop+vrb combinations are the best ones, since they have more best significant values.

5 Conclusions

This paper presents a series of experiments aiming at comparing our proposal of pre-processing techniques based on linguistic information with usual methods adopted for pre-processing in text classification. We find in the literature other alternative proposals for this pre-processing phase. Our approach differs from those since we propose single term selection based on different POS information.

	μP	μR	μF_1	MP	MR	MF_1
baseline	.846	**.782**	**.813**	.782	**.732**	**.753**
nn+adj+prop	.868	**.773**	**.818**	.791	**.720**	**.746**
nn+adj+vrb	.860	.765	**.810**	.783	.710	.738
nn+prop+vrb	.865	.770	**.815**	.788	.716	**.743**
nn+adj+prop+vrb	.860	**.776**	**.816**	.788	**.723**	**.749**

Table 10. POS PAGOD experiments: precision, recall and F_1 micro- and macro-averages for best setups.

From the results we were able to identify which setup is more suited for each dataset:

- for the newspaper articles, lemmatisation with mutual information scoring function and TFIDF weighting scheme, and
- for the juridical collection, lemmatisation with term frequency scoring function and normalised term frequency weighting scheme.

Selecting just some kind of tagged words allowed us to decrease the number of features (around 24%) without affecting learner's performance:

- for FSP, a decrease from 529 to 397 features was obtained using just the words tagged as noun, proper noun and verb.
- for PAGOD, a decrease from 1514 to 1160 was obtained using noun, adjective and proper noun tags.

As conclusion, the presented results support the claim that part-of-speech information can be, in fact, relevant in classification, allowing for a complexity reduction of the problem.

6 Future work

Regarding future work, we intend to perform further tests on different collections and languages. It will be important to evaluate if these results are binded to the Portuguese language and/or the kind of dataset domain.

Aiming to develop better classifiers, we intend to address the document representation problem by trying more powerful representations than the bag-of-words allowing to use word order and syntactical and/or semantical information in document representation. To achieve this goal we plan to use other kind of kernels such as the string kernel (see, for example, [14]).

References

1. A. Aizawa. Linguistic techniques to improve the performance of automatic text categorization. In *NLPRS*, pages 307–314, 2001.

2. E. Bick. *The Parsing System PALAVRAS – Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000.
3. E. Bick. A constraint grammar based question answering system for portuguese. In *Proceedings of the 11th Portuguese Conference of Artificial Intelligence – EPIA'03*, pages 414–418. LNAI Springer Verlag, 2003.
4. C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
5. Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecomunication. John Wiley and Sons, Inc, New York, 1991.
6. T. Gonalves and P. Quaresma. A preliminary approach to the multilabel classification problem of Portuguese juridical documents. In F. Moura-Pires and S. Abreu, editors, *11th Portuguese Conference on Artificial Intelligence, EPIA 2003*, LNAI 2902, pages 435–444, vora, Portugal, December 2003. Springer-Verlag.
7. T. Gonalves and P. Quaresma. Evaluating preprocessing techniques in a text classification problem. In *ENIA'05: Encontro Nacional de Inteligncia Artificial*, So Leopoldo, RS, Brasil, August 2005. (to appear).
8. T. Joachims. *Learning to Classify Text Using Support Vector Machines*. Kluwer academic Publishers, 2002.
9. J.G. Lopes, N.C. Marques, and V.J. Rocio. Polaris: POrtuguese lexicon acquisition and retrieval interactive system. In *The Practical Applications of Prolog*, page 665. Royal Society of Arts, 1994.
10. D. Mladenić and M. Grobelnik. Feature selection for unbalanced class distribution and naïve Bayes. In *Proceedings of the 16th International Conference on Machine Learning – ICML'99*, pages 258–267, 1999.
11. A. Moschitti and Roberto Basili. Complex linguistic features for text classification: A comprehensive study. In *ECIR*, volume 2997. Proceedings Editors: Sharon McDonald, John Tait, 2004.
12. G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
13. H. Schütze, D. Hull, and J. Pedersen. A comparison of classifiers and document representations for the routing problem. In *Proceedings of the 18th International Conference on Research and Developement in Information Retrieval – SIGIR'95*, pages 229–237, Seattle, WA, 1995.
14. J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
15. C.F. Silva, R. Vieira, F.S. Osorio, and P. Quaresma. Mining linguistically interpreted texts. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*, Geneva, Switzerland, August 2004.
16. R. Tong and L.A. Appelbaum. Machine learning for knowledge-based document routing. In Harman, editor, *Proceedings of the 2nd Text Retrieval Conference*, 1994.
17. V. Vapnik. *Statistical learning theory*. Wiley, NY, 1998.
18. I. Witten and E. Frank. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, 1999.