Hierarchical Text Classification for Supporting Educational Programs

Qi Ju∗, Chiara Ravagni⋆†, Alessandro Moschitti∗, and Giampiero Vaschetto⋆

∗DISI, University of Trento, Italy
{qi,moschitti}@disi.unitn.it

⋆Centro Studi Erickson, Italy
{chiara.ravagni,giampiero.vaschetto}@erickson.it

†University of Nuremberg, Germany

Abstract. More than two decades have passed since the first design of the CONSTRUE system [2], a powerful rule-based model for the categorization of Reuters news. Nowadays, statistical approaches are well assessed and they allow for an easy design of text classification (TC) systems. Additionally, the Web has emphasized the need of approaches for digesting large amount of textual information and making it more easily accessible, e.g., thorough hierarchical taxonomies like *Dmoz* or *Yahoo! categories*. Surprisingly, automated approaches have not proved yet to be indispensable for such categorization processes. This suggests that the role of TC might be different from simply routing documents to different topical categories.

In this paper, we provide evidence of the promising use of TC as a support for an interesting and high level human activity in the educational context. The latter refers to the selection and definition of educational programs tailored on specific needs of pupils, who sometime require particular attention and actions to solve their learning problems. TC in this context is exploited to automatically extract several aspects and properties from *learning objects*, i.e., didactic material, in terms of semantic labels. These can be used to organized the different pieces of material in specific didactic program, which can address specific deficiencies of pupils. The TC experiments, carried out with state-of-the-art algorithms and a small set of training data, show that automatic classifiers can easily derive labels like, *didactic context*, *school matter*, *pupil difficulties* and *educative solution type.*

Keywords: hierarchical text classification, information management applications, e-learning

1 Introduction

The last two decades have seen an impressive development of methods for automated text categorization (TC) [7]. This has been mainly due to the combination of two important factors: (i) the exponential development of the Web, requiring for effective methods of information access and management; and (ii) the enhancement in theory and practice of machine learning methods, which constitute the bases of TC.

Despite the success of the TC research, it is still not clear if such technology should be devoted to the design of topical categorization systems as very famous Web hierarchical categorization systems are currently manually maintained, e.g., *Dmoz* or *Yahoo! categories.* On the other hand, TC also regards the association of semantic labels that go beyond the simple routing of information to the most appropriate user feeds. Indeed, this kind of task inevitably suffers from errors in Recall and/or in Precision. Different would be the approach and results, if the outcome of the TC system were cooperatively used as a tool to organize the information in different and creative ways. In this respect, TC would be seen as a tool similarly to search engines, rather than an end-to-end system forced to demonstrate a very high accuracy.

In this paper, we report on our experience with the e-Value project, whose aims are the reorganization or combination of educational materials in different pedagogical contexts. The Erickson Research Centre has been cataloging a large set of published educational materials in smaller units, according to the SCORM (2004) standards, Shareable Content Object Reference Model[1]. These documents are used for the creation of novel and specific didactic product as follows: (i) school classes are evaluated about target cognitive processes; (ii) processes in which pupils have difficulties are detected and recorded in a huge database (DB) of normative data along with the results of its elaboration; (iii) The Decision Support System (DSS) chooses the proper didactic material for the class according to the DB content.

The above steps require: (a) to identify cognitive processes involved in pupils' learning; (b) to divide the didactic materials in smaller parts (learning objects); and (c) classify such objects according to their bibliographic characteristics and to the cognitive processes involved, which depends on the user context (e.g., age, class, special situations). An automatic classifier can be used for easing and speeding up the last step. It can provide a rough classification, which can constitute the starting point for the work of expert catalogers.

The use of the classifier would reduce the cataloging costs, both in terms of time and human resources. Indeed, any educational material, being part of a book, article or best practice, needs to be read and evaluated by experts, before being assigned to the proper categories; this process takes a huge amount of time. As an alternative model, the classifier can perform a first approximate categorization and after, the experts can refine it. The clear advantage is that materials pertaining to a certain subject can be directly assigned to its experts (working in that field), thus improving the accuracy of classification and avoiding the burden to exchange materials among the different experts.

However, the above scenario could be realized only if the adopted multi-class classifier (MCC) performed accurate hierarchical categorization. Given the novelty of the intended taxonomy, it is not simple to predict if MCC can deploy the needed accuracy. For this purpose, we have:

- designed a new taxonomy that meets the organization needs of e-Value;
- defined an annotation procedure and produced an initial datasets of 122 documents, organized in 112 categories (of course the documents are repeated in the hierarchy); and
- implemented an MCC, which exploits state-of-the-art TC models such as, Support Vector Machines, structured in binary flat categorizers.

The preliminary experiments on the overall hierarchy of 112 nodes show promising results, ranging from a Micro-F1 of above 95% for the first level to about 70% on the whole hierarchy. This outcome is rather promising and enables future research in the use of TC for the efficient implementation of educational programs.

In the reminder of this paper, Section 2 describes the tackled task in more detail, Section 3 reports on our results and Section 4 derives the final conclusions.

2 Automatic Support to the e-Value project

The main objective of the e-Value project is to design, develop and test a multimedia platform (consisting of a set of web applications), which integrates the evaluation of various learning abilities and the application of didactic processes. These can benefit from automatic methods for classifying the didactic material used in such

[1] http://www.adlnet.gov/capabilities/scorm/scorm-2004-4th

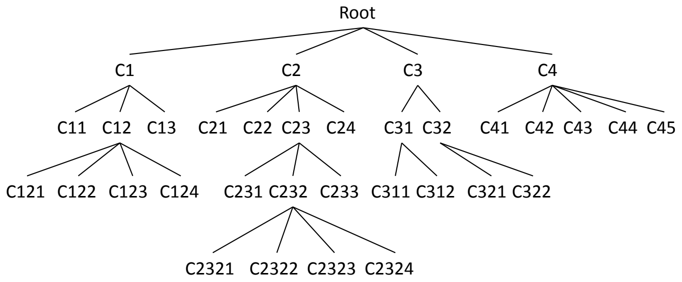


Fig. 1. Hierarchical Categorization Scheme of e-Value (only category with at least 1 training documents are present

processes. The next sections describe the problem in more detail and suggest how a TC system can be used in such context.

2.1 e-Value Framework

The framework includes different interconnected processes:

- standard evaluation procedures and dynamic assessment of learning abilities of pupils;
- collection of normative data, e.g., educational material and pupils' evaluations;
- continuous data flow, i.e., the related database is continuously updated and the normative data currently available is integrated and compared with the new arriving data; and
- qualitative and quantitative evaluation of the collected data.

The educational material is used for defining didactic products, which address specific action (intervention). It consists of books, CD-ROMs, collections of articles, etc. The e-Value project aims at both using independently and jointly the materials above.

Designing an intervention often requires the use of units taken from several books or CD-ROMs but including the entire sources is very ineffective, considering that only some small parts will be used. To enable more flexibility in the creation of training programs, the material collections are divided into basic training units, called learning objects, which can be reassembled in a flexible way. This requires to analyze the materials to be used in the interventions and selecting the portion involved in the target cognitive processes.

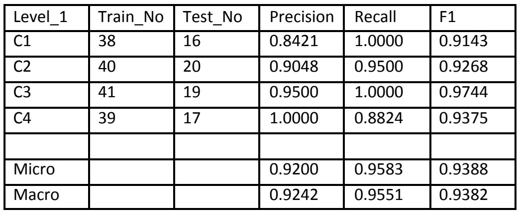
2.2 A framework use-case

A use of the framework is illustrated by the following example. In a school context some classes are evaluated with respect to targeted cognitive processes. The tests may reveal that some of the pupils have difficulties in certain processes. Thus, the test results are recorded (building a large database of normative data) along with some elaboration of them, i.e., basic data statistics. Then the DSS chooses the proper didactic material for the class by proposing different material to pupils requiring attention and quick intervention. For this purpose the educational team need to:

- identify every cognitive process that can be involved in learning. At the moment, this has been restricted to mathematics and reading-writing (with linguistic skills and metaphonetics);

C1	**Categorizzazione contesto didattico**	C22	Metafonologia	C2435	Espressioni
C11	Scuola dell'infanzia	C221	Globale	C2436	Potenze
C12	Scuola primaria	C2211	Rima	C2437	Radici quadrate
C121	Primaria Classe I	C2212	Sillaba	C244	Calcolo-Numeri razionali
C122	Primaria Classe II	C222	Profonda	C245	Calcolo-Numeri relativi
C123	Primaria Classe III	C2221	Fonema	C246	Calcolo-Rapporti e proporzioni
C124	Primaria Classe IV	C23	Abilità linguistiche	C247	Calcolo-Calcolo letterale
C125	Primaria Classe V	C231	Lessico	C248	Problem-solving
C13	Scuola secondaria di 1 grado	C2311	Denominazione	C249	Capacità di orientarsi nello spazio
C2	**Categorizzazione materia**	C2312	Categorizzazione	C24_10	Costruire sistemi di riferimento convenzionali
C21	Letto-scrittura	C2313	Identificazione	C24_11	Geometria euclidea (piana)
C211	Prerequisiti	C2314	Definizione	C24_12	Misura di grandezze geometriche
C2111	Prerequisiti grafo motori	C2315	Polisemia	C24_13	Misura di grandezze fisiche
C2112	Prerequisiti visuo spaziali	C2316	Arricchimento lessicale	C24_14	Le trasformazioni geometriche
C2113	Teorie ingenue	C232	Morfo-sintassi	C25	Altro
C212	Decodifica	C2321	Concordanze	C3	**Categorizzazione situazione alunni**
C2121	Lettere	C2322	Struttura della frase	C31	BES
C2122	Sillabe	C2323	Analisi grammaticale	C311	Autismo
C2123	Parole	C2324	Analisi logica	C312	Udito
C2124	Non parole	C233	Narrazione	C313	Vista
C2125	Frasi-Brano	C2331	Comprensione racconto	C314	Psicomotricità
C213	Comprensione	C2332	Produzione racconto	C315	Sindrome di Down
C2131	Parole	C24	Matematica	C316	Altro
C2132	Frasi	C241	Numero	C32	DSA
C2133	Brano	C2411	Processi semantici	C321	Iperattività
C214	Compitazione	C2412	Conteggio	C322	Dislessia
C2141	Clettere	C2413	Processi pre-sintattici	C323	Disgrafia
C2142	Sillabe non ortografiche	C2414	Processi lessicali e sintattici	C324	Discalculia
C2143	Parole non ortografiche	C242	Calcolo - processi di base	C325	Combinazione di DSA diversi - altro
C2144	CNon parole	C2421	Segni delle operazioni	C326	Nessun DSA
C215	Ortografia	C2422	Fatti numerici	C4	**Categorizzazione tipo di intervento**
C2151	Oparole	C2423	Tabelline	C41	Potenziamento
C2152	Ofrasi	C2424	Calcolo a mente	C42	Recupero
C2153	Obrano	C243	Calcolo - numeri naturali	C43	Didattica insegnamento
C216	Stesura testo	C2431	Algoritmi di calcolo scritto	C44	Intervento logopedico
C2161	Pianificazione	C2432	Incolonnamento di numeri	C45	Intervento psicologico
C2162	Trascrizione	C2433	Multipli e divisori		
C2163	Revisione	C2434	Minimo Comune Multiplo e Massimo Comune Denominatore		

Table 1. Description of the different categories of the hierarchy in Figure 1.



Level_1	Train_No	Test_No	Precision	Recall	F1
C1	38	16	0.8421	1.0000	0.9143
C2	40	20	0.9048	0.9500	0.9268
C3	41	19	0.9500	1.0000	0.9744
C4	39	17	1.0000	0.8824	0.9375
Micro			0.9200	0.9583	0.9388
Macro			0.9242	0.9551	0.9382

Fig. 2. Performance for the first level

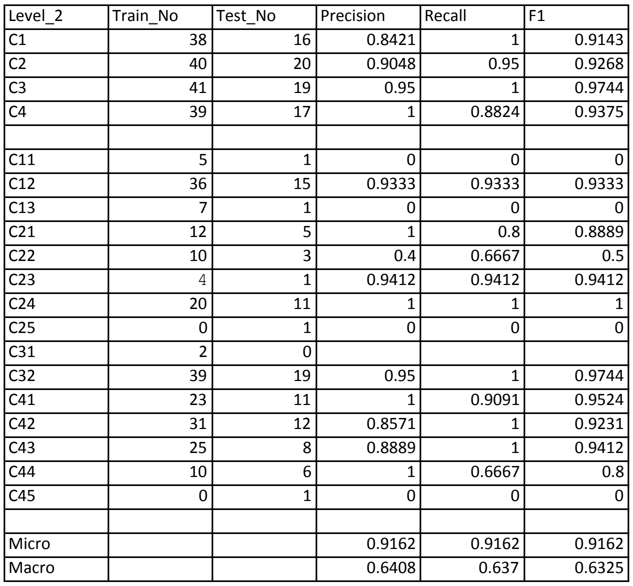
- divide the didactic materials in smaller parts (learning objects). This because the use of the entire books or CD-Rom would be unfeasible, considering that just a few exercises need to be applied. Thus the whole material has to be checked by experts to be subdivided in learning objects. The latter are then used to design the formative offer, in place of the entire material, obtaining a more personalized and individualized learning.
- Categorize the materials according to their bibliographic characteristics and, most importantly for the fruition of the materials, to features of the involved cognitive processes, e.g., the age, class and special situations of the target pupils etc.
- Porting the material from paper or optical media to an electronic format (pdf or swf) so that it can be reassembled online and offline.

In the last phase the application of an automatic classifier can provide significant benefits to the whole process as explained in the following section.

2.3 Classification Task

To meet the need of the e-Value project, we have defined a new taxonomy as well as the annotation procedure and initial datasets. Our hierarchical categorization scheme is shown in Figure 1, whose more descriptive labels are reported in Table 1. The materials have to be classified according to four macro-categories, and then divided into a structure of sub-categories of 4 levels. Each category is meaningful for a correct description of the materials, from both administrative perspective (e.g., in which educational context should be applied) and subject/cognitive process viewpoint (e.g. Mathematics – Number – Lexical and semantic processes instead of Mathematics – Basic processes of calculus – Numerical facts). The Macro-categories are: C1 – School and class (referring to the ages 5 – 14); C2 – Subject/cognitive process (referring to the subjects of mathematics, linguistics, phonetics, reading-writing abilities); C3 – Pupils' situation (for the cases of special needs or particular situations); and C4 – Type of material (or the normal didactic usage in the class, or for pupils with special situation or greater difficulties in the subject).

Such automatic classification could improve the manual categorization costs, in terms of both time and human resource. Each piece of educational material, being part of a book, article or best practice, needs to be read and evaluated by experts, before being assigned to the proper categories, and this process takes a huge amount of time. Therefore, the use of an automatic classifier could significantly reduce the time required to read and evaluate the materials. Of course, experts will need to read part of the material in any case to refine and validate the output of the classifier. However, the materials pertaining to a certain subject can be directly routed to the experts of such field, thus improving the categorization accuracy.



Level_2	Train_No	Test_No	Precision	Recall	F1
C1	38	16	0.8421	1	0.9143
C2	40	20	0.9048	0.95	0.9268
C3	41	19	0.95	1	0.9744
C4	39	17	1	0.8824	0.9375
C11	5	1	0	0	0
C12	36	15	0.9333	0.9333	0.9333
C13	7	1	0	0	0
C21	12	5	1	0.8	0.8889
C22	10	3	0.4	0.6667	0.5
C23	4	1	0.9412	0.9412	0.9412
C24	20	11	1	1	1
C25	0	1	0	0	0
C31	2	0			
C32	39	19	0.95	1	0.9744
C41	23	11	1	0.9091	0.9524
C42	31	12	0.8571	1	0.9231
C43	25	8	0.8889	1	0.9412
C44	10	6	1	0.6667	0.8
C45	0	1	0	0	0
Micro			0.9162	0.9162	0.9162
Macro			0.6408	0.637	0.6325

Fig. 3. Performance for the second level

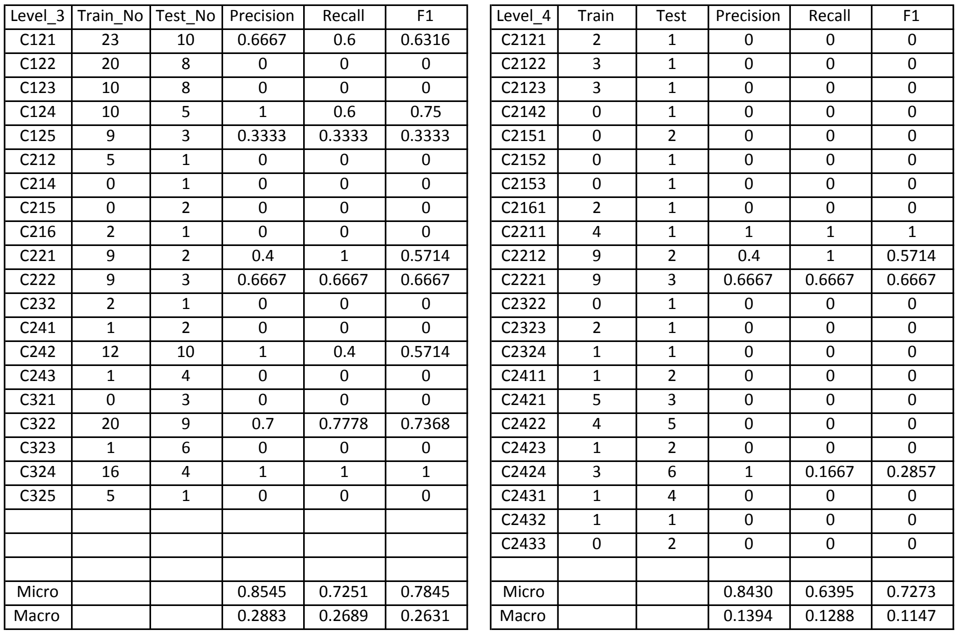
3 Experiments

The aim of our evaluation is to demonstrate that state-of-the-art TC methods can be applied to learn hierarchical classifiers for our e-Value taxonomy. This task is made complex by two different aspects: (i) in addition to topic labels such as, *Euclidean Geometry*, *Problem Solving* or *Geometric Transformation*, the taxonomy also contains semantic characterization such as *Story Development* or *Story Understanding*, whose characterization using simple terms seems harder; and (ii) given the novelty of the taxonomy, we could only produce a small dataset, which makes the learning of classification functions more difficult. To deal with and analyze such problems, we experimented with hierarchy subsets, defined according to the hierarchy's levels, ranging from 1 to 4 (the maximum depth of our hierarchy). The deeper the level, the more difficult TC is.

3.1 Setup

One major drawback of machine learning and thus of TC based on it is the need of training data, i.e., a set of documents manually classified into the referring taxonomy. This data is difficult to find and/or to produce as it requires human labor. Given the novelty of our taxonomy defined in Figure 1, no previous data was available. Thus, we set an annotation procedure (with only one annotator) of the didactic material available in the Erickson's database. We randomly selected 60 documents and we classified each of them according to all the 112 nodes of the taxonomy. This led to a dataset of 122 documents (repetitions are considered).

We randomly divided the above data in training and test set by taking care that for each document all its repetitions were all put either in the training or in the test set. The training data was used to learn the set of 112 binary classifiers, one for each category, following the one-vs-all schema. The output of the multi-class classifier is the merged set of the individual binary classifier decisions. Although simple, this is considered a state-of-the-art approach [5, 3]. We used default SVM parameters as the small training data prevented to apply any reasonable parameterization approach. We used a bag-of-term representation (string separated by space and punctuation) without applying any feature selection, stop list or lemmatization. Although, we are



Level_3	Train_No	Test_No	Precision	Recall	F1
C121	23	10	0.6667	0.6	0.6316
C122	20	8	0	0	0
C123	10	8	0	0	0
C124	10	5	1	0.6	0.75
C125	9	3	0.3333	0.3333	0.3333
C212	5	1	0	0	0
C214	0	1	0	0	0
C215	0	2	0	0	0
C216	2	1	0	0	0
C221	9	2	0.4	1	0.5714
C222	9	3	0.6667	0.6667	0.6667
C232	2	1	0	0	0
C241	1	2	0	0	0
C242	12	10	1	0.4	0.5714
C243	1	4	0	0	0
C321	0	3	0	0	0
C322	20	9	0.7	0.7778	0.7368
C323	1	6	0	0	0
C324	16	4	1	1	1
C325	5	1	0	0	0
Micro			0.8545	0.7251	0.7845
Macro			0.2883	0.2689	0.2631

(a) third level

Level_4	Train	Test	Precision	Recall	F1
C2121	2	1	0	0	0
C2122	3	1	0	0	0
C2123	3	1	0	0	0
C2142	0	1	0	0	0
C2151	0	2	0	0	0
C2152	0	1	0	0	0
C2153	0	1	0	0	0
C2161	2	1	0	0	0
C2211	4	1	1	1	1
C2212	9	2	0.4	1	0.5714
C2221	9	3	0.6667	0.6667	0.6667
C2322	0	1	0	0	0
C2323	2	1	0	0	0
C2324	1	1	0	0	0
C2411	1	2	0	0	0
C2421	5	3	0	0	0
C2422	4	5	0	0	0
C2423	1	2	0	0	0
C2424	3	6	1	0.1667	0.2857
C2431	1	4	0	0	0
C2432	1	1	0	0	0
C2433	0	2	0	0	0
Micro			0.8430	0.6395	0.7273
Macro			0.1394	0.1288	0.1147

(b) fourth level

Fig. 4. Performance for the third and fourth level. Categories with no document in the test set and the categories of upper levels are not reported.

confident that the latter may relevantly improves our models. We used the classical $log(TF) * IDF$ weighting scheme and normalized vectors.

The performance is provided by means of Micro- and Macro-Average F1, evaluated from our test data over all 112 categories. Additionally, the F1s of the binary classifiers are reported. For measuring the performance of different hierarchical levels, only the nodes up to the target level are considered, e.g., for the first level, we only measure the Micro/Macro F1 of C1, C2, C3 and C4.

3.2 Results and Discussion

Table 2 reports the performance on the first level. We note that for each category there are about 40 documents for training. These seem to be enough as the accuracy of the individual categories as well as the overall Micro/Macro F1 is exceptionally high. This is not completely surprising as most documents are repeated in the above four categories.

Table 3 illustrates the results for the second level. We note that when the training documents are more than 20, very good results can be achieved. Low performance is shown for C11 and C13, which are trained with less than 7 documents. Additionally, they have only one test document, this means that their accuracy cannot really be estimated. The situation of C31 is even worse as it has no test documents. In this case, we do not report any accuracy in the related row. It should also be noted that, since we use one-vs-all schema, the accuracy of C1,..,C4 is the same as before. Thus, from now on, we will not report the accuracy of previously reported binary classifiers.

Table 4 shows the performance on levels 3 and 4. Again the few training documents available for the classifiers prevent to achieve a reasonable F1. There are some good cases such as C124 and C322 but also bad cases such as C122 and C123. The latter two refer to *Primaria Classe II* and *Primaria Classe III*, respectively, which have large overlap with the other classes, i.e., I, IV and V. For separating such categories, the simple bag-of-words may not be enough.

4 Conclusions

In this paper, we have described an interesting and new semantic classification problem in the context of the educational framework of the e-Value project. We have defined a new hierarchical taxonomy, which is promising for improving the production cycle of educational systems. To test the feasibility of the approach, we have also built a corpus annotated according to the above taxonomy. Such data was used for training an MCC based on SVMs. The results show that when there is a reasonable amount of training documents the classifiers can deploy remarkably high accuracy. On the other hand, the F1 of lower level categories is highly affected by data scarceness. Some categories would probably require the definition of more expressive features to better model their separation.

Possible solutions are also provided by previous work, which shows more advanced TC models, e.g., [6], in which global dependencies between hierarchical nodes are encoded in a gradient descendent learning approach. They experimented with Reuters Volume 1 (RCV1) [2] on a subhierarchy only containing 34 nodes. Other relevant work such as [4] and [1] uses a rather different datasets and a different idea of dependencies based on the feature distributions over the linked categories. Finally, [3] experiment with models similar to ours achieving state-of-the-art on RCV1.

Acknowledgements

The research described in this paper has been partially supported by the Italian Project e-Value (PAT) and by the European Community's Seventh Framework Programme (FP7/2007-2013) under the grants #231126: LivingKnowledge – Facts, Opinions and Bias in Time, #247758: EternalS – Trustworthy Eternal Systems via Evolving Software, Data and Knowledge, and #288024: LiMoSINe – Linguistically Motivated Semantic aggregation engiNes.

References

1. Dumais, S.T., Chen, H.: Hierarchical classification of web content. In: Belkin, N.J., Ingwersen, P., Leong, M.K. (eds.) Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval. pp. 256–263. ACM Press, New York, US, Athens, GR (2000), http://research.microsoft.com/~sdumais/sigir00.pdf
2. Hayes, P.J., Weinstein, S.P.: Construe/Tis: a system for content-based indexing of a database of news stories. In: Rappaport, A., Smith, R. (eds.) Proceedings of IAAI-90, 2nd Conference on Innovative Applications of Artificial Intelligence. pp. 49–66. AAAI Press, Menlo Park, US (1990)
3. Lewis, D.D., Yang, Y., Rose, T., Li, F.: Rcv1: A new benchmark collection for text categorization research. The Journal of Machine Learning Research (5), 361–397 (2004)
4. McCallum, A., Rosenfeld, R., Mitchell, T.M., Ng, A.Y.: Improving text classification by shrinkage in a hierarchy of classes. In: ICML. pp. 359–367 (1998)
5. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. J. Mach. Learn. Res. 5, 101–141 (December 2004), http://dl.acm.org/citation.cfm?id=1005332.1005336
6. Rousu, J., Saunders, C., Szedmak, S., Shawe-Taylor, J.: Kernel-based learning of hierarchical multilabel classification models. The Journal of Machine Learning Research (7), 1601–1626 (2006)
7. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)

[2] trec.nist.gov/data/reuters/reuters.html