Centre National de Recherche Technologique
Text Indexing Seminar
Rennes, France – April 3, 2002

Automated Text Categorization: Tools, Techniques and Applications

Fabrizio Sebastiani

Istituto di Elaborazione dell'Informazione
Consiglio Nazionale delle Ricerche
56124 Pisa, Italy
E-mail: `fabrizio@iei.pi.cnr.it`
WWW: `http://faure.iei.pi.cnr.it/~fabrizio/`

When it was proclaimed that the Library contained all books, the first impression was one of extravagant happiness. All men felt themselves to be the masters of an intact and secret treasure. There was no personal or world problem whose eloquent solution did not exist in some hexagon.

(. . .)

As was natural, this inordinate hope was followed by an excessive depression. The certitude that some shelf in some hexagon held precious books and that these precious books were inaccessible, seemed almost intolerable.

[Jorge Luis Borges, *The Library of Babel*, 1941]

Tackling Information Overload by Text Categorization

Nowadays, there are two main paradigms for tackling information overload :

1. build high-quality tools for searching an unstructured document base, such as the Web. This is the "standard" answer from text search
2. build high-quality tools for structuring the document base, e.g. into a digital library. This is the answer from automated text categorization (ATC)

It is the latter approach that we will concentrate on in this talk.

Overview of this talk

1. A definition of the ATC task
2. Learning and evaluating text classifiers
3. Applications of ATC
4. Applications of ATC at IEI-CNR
   (1) Automated indexing of scientific articles under hierarchical classification schemes
   (2) Personalized information delivery to digital library users
   (3) Automated construction of thematic lexicons
   (4) Automated survey coding
5. Conclusion

1. A definition of the ATC task

ATC is the task of approximating the unknown target function $\Psi : \mathcal{D} \times \mathcal{C} \rightarrow \{T, F\}$ by means of a function $\Phi : \mathcal{D} \times \mathcal{C} \rightarrow \{T, F\}$ called the classifier , such that Ψ and Φ “coincide as much as possible”. Here

- $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$ is a fixed set of pre-defined categories ;
- $\mathcal{D}$ is a domain of documents.

Depending on the application, classification may be

- single-label : exactly one category must be assigned to each document. A special case is when $|\mathcal{C}| = 2$ (the binary case).
- multi-label : any number of categories can be assigned to each document.

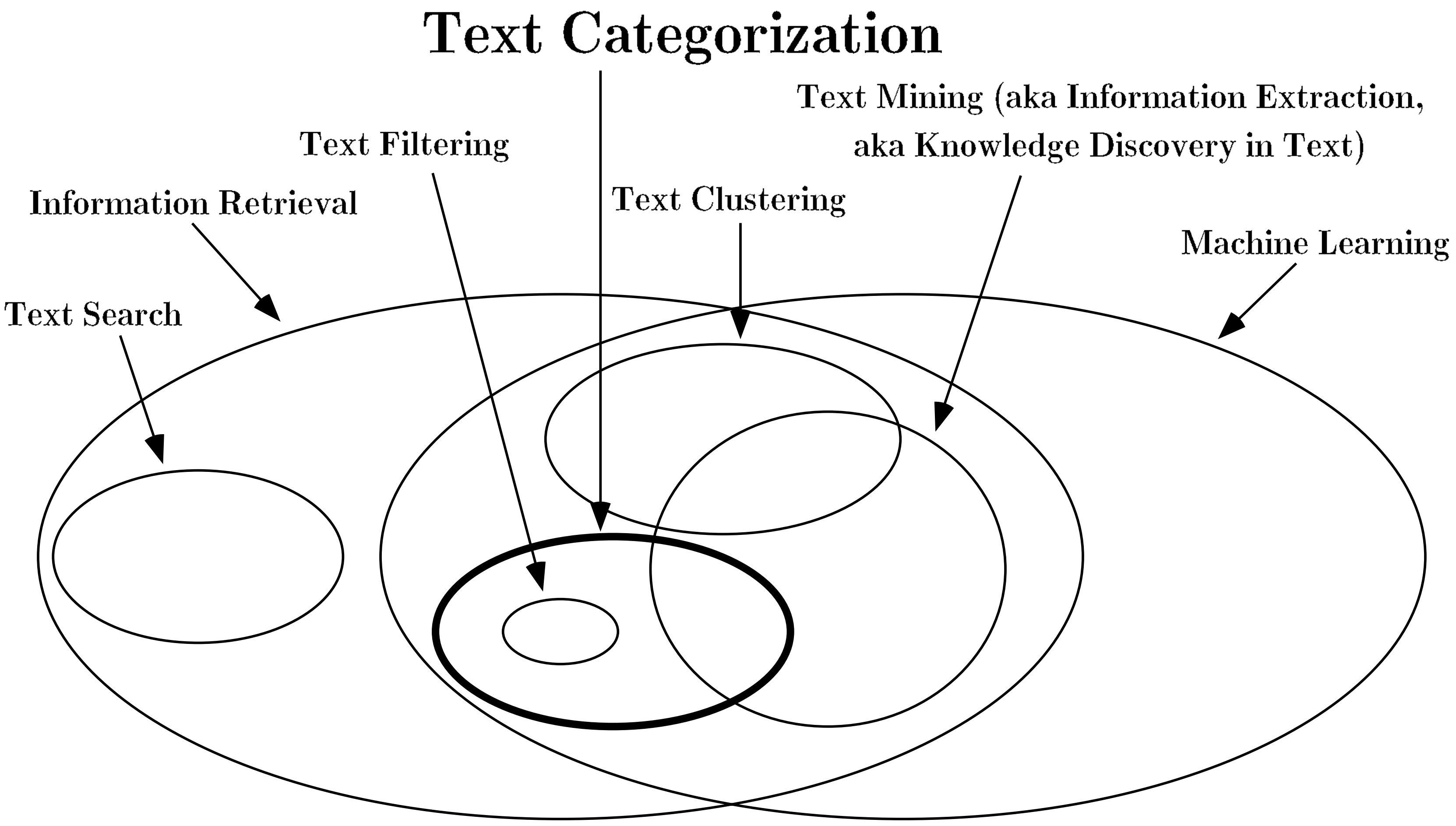
In general it is the case that:

- the categories are just symbolic labels. No additional knowledge of their “meaning” is available to help in building the classifier; in particular, the “text” constituting the label is not significant;

- the attribution of documents to categories should be realized on the basis of the *content* of the documents, and not on the basis of *metadata* that may be available from an external source.

Given that the content of a document is a *subjective* notion, this means that the fundamental notion of ATC, that of membership of a document in a category, cannot be decided deterministically.

In an operational environment the two assumptions above may not be verified, and one uses whatever source of knowledge is available.



Text Categorization

Text Mining (aka Information Extraction,
aka Knowledge Discovery in Text)

Text Filtering

Text Clustering

Information Retrieval

Machine Learning

Text Search

2. Learning and evaluating text classifiers

A binary classifier for class *c*, i.e. one that decides between *c* and (**not** *c*)

- is built automatically, by machine learning techniques, from a "training" set of documents preclassified under *c* and (**not** *c*)
- is tested by comparing its decisions with the human decisions encoded in a "test" set of documents preclassified under *c* and (**not** *c*)

Learning techniques often used in ATC are

- "legacy" techniques: decision trees, decision rules, probabilistic (Bayes) classifiers, neural networks, etc.
- "emerging" techniques: boosting, support vector machines.

Usually, a document is represented as a (sparse) vector of weights, where the length of the vector is the number of *terms* that occur at least once in the training set.

Weights may be binary (indicating presence or absence of the term in the document), or non-binary (indicating how much the term contributes to the semantics of the document). In this latter case, weighting functions from IR (such as $tf * idf$) are used.

Previous to weighting, stop word removal and stemming are often performed. Dimensionality reduction (aka feature selection) may also be performed in order to improve the efficiency of the system.

Classification is a subjective task, and both human and automatic classifiers are error-prone. The effectiveness of a (human or automatic) classifier must be measured by a combination of

- recall : “How many documents truly belonging to the category have been deemed as such”?
- precision : “How many documents deemed to belong to the category truly belong to it”?

The effectiveness of automatically built classifiers now rivals that of human classifiers.

ATC research at IEI-CNR. Methods

We have investigated a number of methods and issues in ATC, including

- hypertext classification through "blurb indexing" [Attardi et al., THAI'99];
- novel "boosting" algorithms for learning text classifiers [Sebastiani et al., ACM CIKM'00];
- novel "feature selection" techniques for optimizing document representations for ATC [Galavotti et al., ECDL'00];
- "bigram indexing" for ATC [Caropreso et al., TD'01].

3. Applications of ATC

Historically, the most important applications of ATC are:

- automated document indexing with controlled vocabularies , i.e. classifying documents with categories (or "subject codes", or "descriptors", . . .) from e.g.
  - the Library of Congress Cataloging Scheme;
  - Dewey Decimal System.

  This is a form of *automated metadata generation*.

- personalized information delivery , i.e. routing a stream of documents to the interested users only by deciding, for User_i, whether the document belongs to User_i or (**not** User_i)

Other applications of ATC have been:

- filing patents under predefined patent categories
- filing "classified ads" into classes (e.g. deciding whether a given ad should be printed under Cars for Sale, or Real Estate, . . .)
- filing Web sites, or organizing search results, under Yahoo!-like hierarchical Web directories
- filtering unsuitable content (e.g. deciding between Pornography and (**not** Pornography)) or junk mail (Spam vs. (**not** Spam))
- detecting authorship of documents of disputed paternity (e.g. Shakespeare vs. (**not** Shakespeare))
- classifying images through the analysis of textual captions
- automatically identifying text genre (e.g. ForKids vs. (**not** ForKids))

ATC research at IEI-CNR. Applications

We have tackled a number of applications of ATC, including

- Automated indexing of scientific articles under hierarchical classification schemes
- Personalized information delivery to digital library users
- Automated construction of thematic lexicons
- Automated survey coding

Applications we intend to tackle in the near future are

- Automated authorship attribution
- Spoken text segmentation and categorization, with topic detection and tracking

4. ATC applications at IEI-CNR (1) Automated indexing of scientific articles under hierarchical classification schemes

CompCat is an internally funded project now starting at IEI-CNR and concerned with generating a classifier of scientific articles for building digital libraries in the computer science domain. Its main characteristics are:

- the categories are the ones from the ACM Classification Scheme (version of 1998)
- the training and test sets are years from 1998 onwards of the ACM Digital Library
- the classifier will be $\boxed{\text{interactive}}$, i.e. will suggest to the user a list of categories ranked according to their estimated appropriateness for the document

“Trivial” solution to the problem: for each (internal of leaf) category c, train a binary classifier that decides between c and (**not** c). This solution is problematic since it is

- inefficient : since a document may belong to multiple categories, this means invoking thousands of classifiers for each document
- ineffective : the information provided by the hierarchical structure of the category set is not exploited. There are thousands of categories, and there may thus be very few training documents in the “leaf” categories

In this work (joint with H. Avancini and A. Rauber) we exploit insights into hierarchical text categorization well-known from the literature:

- we bring to bear the hierarchical structure of the classifier by "soft pruning" (i.e. we consider only the k most promising "children" of a given node), thus enhancing efficiency;
- we make term statistics more robust by "shrinkage", thus enhancing effectiveness for categories at the lower levels in the hierarchy.

To these intuitions, we add a novel idea, i.e.

- discarding the naturally occurring hierarchy
- generating an artificial one by means of a hierarchical agglomerative clustering technique
- using this latter instead of the naturally occurring one for learning the final classifiers.

This apparently counterintuitive idea relies on the observation that naturally occurring hierarchical structures are a result of *human* conceptualization, while the ones generated by clustering are the result of *machine* (statistics-based) conceptualization. While the former are indeed advantageous for human understanding, a statistics-based algorithm such as shrinkage can profit more from the latter than the former.

ATC applications at IEI-CNR (2) Personalized information delivery to DL users

CYCLADES is a CEC-funded project (IST-2000-25456), started Feb 2001 and coordinated by IEI-CNR, aiming to provide a layer of services for users of a distributed DL of grey literature compliant with the Open Archives initiative (OAi) standard.

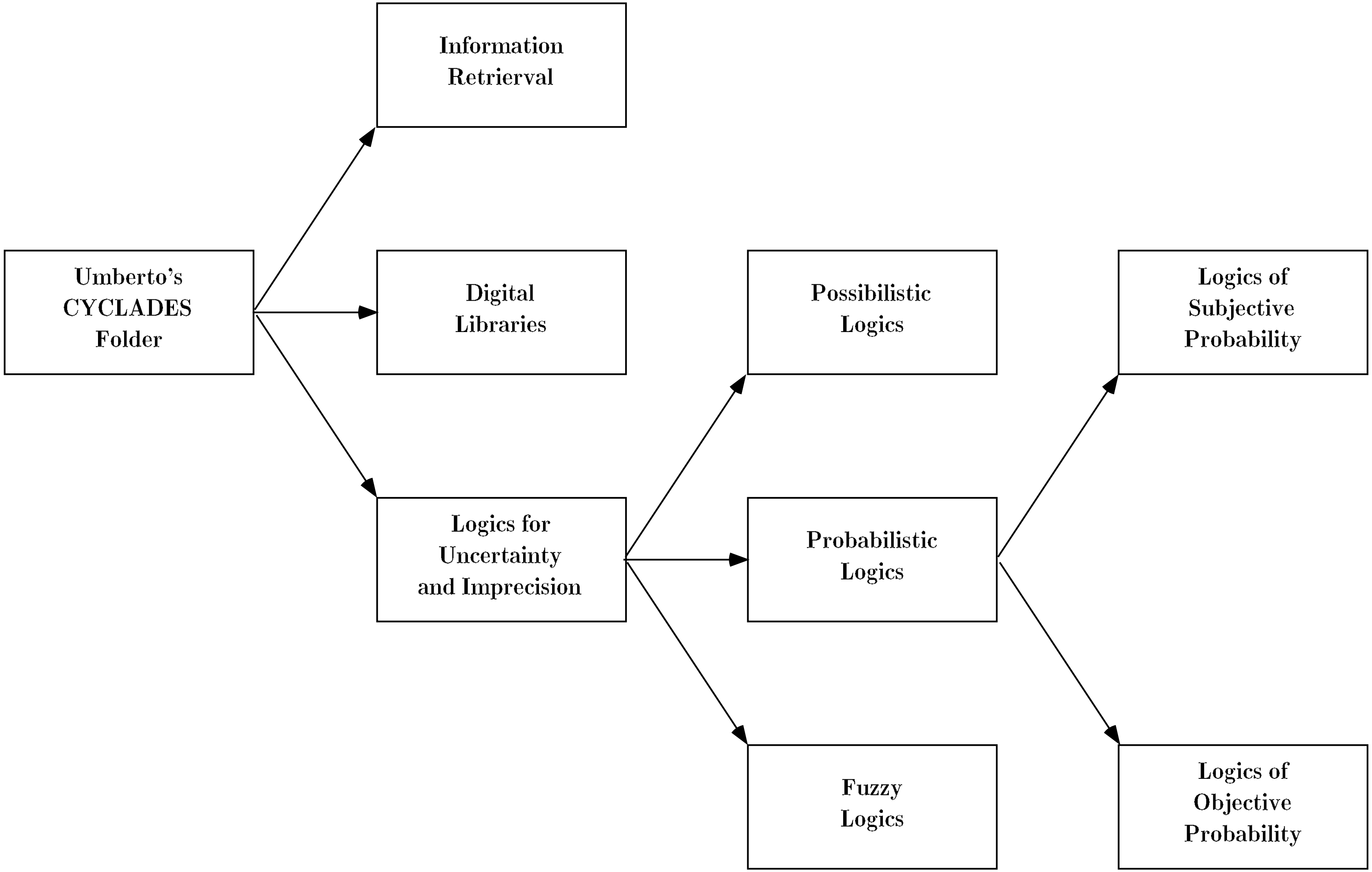
Key to CYCLADES is the Personalization Service (PS), responsible for allowing a user to interact with the system in a flexible and highly personalized way. We are building this (joint work with E. Renda and U. Straccia) by means of ATC techniques.

Personalization is viewed as a content-based notion ; i.e. the interaction with the user must take into account the user's interests.

A Cyclades user will obtain documents (in either "push" or "pull" modality) from the DL;

- each document will be classified into one or more of a hierarchically structured set of folders, each of which represents the user's subjective view of a topic which is of interest to her
- each user action (e.g. moving, removing, printing a document) will be be interpreted as "implicit feedback" on the meaning of the folders, and used to revise the classifiers.

An example personal folder hierarchy


Information Retrierval
Umberto's CYCLADES Folder
Digital Libraries
Possibilistic Logics
Logics of Subjective Probability
Logics for Uncertainty and Imprecision
Probabilistic Logics
Fuzzy Logics
Logics of Objective Probability

This application has meany features in common with the previous one (e.g. the category set is hierarchically structured; there are thousands of categories, and "leaf" categories may suffer from data sparseness), which means that some of the solutions adopted there will be exploited also here (e.g. "shrinkage", "soft pruning", ...).

In addition we intend

- to bring to bear implicit user feedback by adopting an "incremental" classifier (BalancedWinnow), thus enhancing effectiveness for data-sparse categories by exploiting the dynamic character of this application;
- to bring to bear not only document content, but also ratings given to the document by other users deemed "similar" to the current user. This yields a form of hybrid (i.e. ratings- and content-based) recommendation.

ATC applications at IEI-CNR
(3) Automated construction of thematic lexicons

This work (joint with A. Lavelli and B. Magnini) proposes a novel task, i.e. term categorization , for the automated construction of thematic lexicons.

The main idea is to apply ATC techniques in a dual way: while the purpose of ATC is that of classifying documents represented as vectors in a space of terms, the purpose of term categorization is that of classifying terms (into zero, one, or several categories belonging to a predefined set) represented as vectors in a space of documents.

Previous techniques for learning thematic lexicons need a set of preclassified documents. Our technique needs instead

- a small set of preclassified terms. For this, we have used WordNetDomains, an extension of WordNet in which each term has been labelled with one or more from a set of 238 among the most popular labels used in dictionaries for sense discrimination purposes (e.g. Zoology, Sport, Basketball)
- a corpus of unlabelled documents. For this we have used various subsets of the Reuters Corpus Volume I (RCVI).

Our technique extends the existing thematic lexicons by learning from the documents the associations between previously contained terms and new terms.

ATC applications at IEI-CNR
(4) Automated survey coding

This work (joint with D. Giorgetti and I. Prodanof) proposes that automated survey coding (ASC) be viewed as an ATC task. ASC

- is the task of analyzing the answers that a person has given to an open-ended questionnaire (e.g. a social survey) and "classifying" this person into one among a predefined set of categories, based on the answers.
- has many important applications, such as filing resumes into categories, classifying respondents to social surveys, etc.
- had previously been viewed in terms of similarity matching between the answer and a textual description of the category. This is unsuitable, since this description is usually not useful enough, and is often not available.

Example from a social survey conducted by the US National Opinion Research Center (NORC) :

Question: `Within the past month, think about the last time you felt really angry, irritated or annoyed. Could you describe in a couple of sentences what made you feel that way what the situation was`?

Applicable Categories :

- `ANGRYWRK`: `Situation involved work`;
- `ANGRYFAM`: `Situation involved family`;
- `ANGRYGVT`: `Situation involved government or government officials`;
- ...
- `OTHER: Situation did not fit the above categories`.

Conclusion

Nowadays ATC is considered

- essential for either tackling or supporting any application involving automated indexing with controlled vocabulary or automated metadata generation
- essential for tackling any application involving timely personalized information delivery

The success story of ATC is also going to encourage an extension of its methods and techniques to neighbouring fields of application. Techniques typical of ATC have already been extended successfully to the categorization of documents expressed in slightly different media; for instance:

- very noisy text resulting from optical character recognition.
- speech transcripts.

And for those of you interested in ATC ...

- "Machine Learning in Automated Text Categorization", *ACM Computing Surveys*, 34(1):1–47, 2002, at `http://faure.iei.pi.cnr.it/~fabrizio/Publications/ACMCS02.pdf`
- "Machine Learning in Automated Text Categorization", slides from the ECDL'01 (and probably ECDL'02) tutorial, at `http://faure.iei.pi.cnr.it/~fabrizio/Slides/ATCslides.pdf`
- On-line searchable bibliography on ATC (part of the *Collection of Computer Science Bibliographies*), with $\geq$ 350 entries, at `http://faure.iei.pi.cnr.it/~fabrizio/central.html#ATCbiblio`
- T. Joachims and F. Sebastiani, guest eds., *Journal of Intelligent Information Systems* 18(2), Special Issue on ATC, March-May 2002.
- 2nd Workshop on Operational Text Classification Systems, Tampere, FI, August 2002, at `http://faure.iei.pi.cnr.it/~fabrizio/otc2002.html`

References

1. Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.
2. Alberto Lavelli, Bernardo Magnini and Fabrizio Sebastiani. Building Thematic Lexical Resources by Bootstrapping and Machine Learning. *Proceedings of the LREC-02 Workshop on Linguistic Knowledge Acquisition and Representation*, Las Palmas, ES. Forthcoming.
3. Daniela Giorgetti, Irina Prodanof and Fabrizio Sebastiani. Mapping an Automated Survey Coding Task into a Probabilistic Text Categorization Framework. *Proceedings of PorTAL-02, International Conference on Natural Language Processing*, Faro, PT. Lecture Notes in Computer Science, Springer Verlag, Heidelberg, DE. Forthcoming.
4. Thorsten Joachims and Fabrizio Sebastiani (guest eds.), Special Issue on Automated Text Categorization, *Journal of Intelligent Information Systems*, 18(2):103–105, 2002.
5. Fabrizio Sebastiani. Organizing and using digital libraries by automated text categorization. In Luciana Bordoni and Giovanni Semeraro (eds.), *Proceedings of the AI*IA Workshop on Artificial Intelligence for Cultural Heritage and Digital Libraries*, Bari, IT, 2001, pp. 93–94.
6. Maria Fernanda Caropreso, Stan Matwin and Fabrizio Sebastiani. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In Amita G. Chin, editor, *Text Databases and Document Management: Theory and Practice*, Idea Group Publishing, Hershey, US, 2001, pp. 78–102.
7. Luigi Galavotti, Fabrizio Sebastiani, and Maria Simi. Experiments on the use of feature selection and negative evidence in automated text categorization. In José Borbinha and Thomas Baker (eds.), *Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries*, Lisbon, PT, 2000. Published in the "Lecture Notes for Computer Science" series, number 1923, Springer Verlag, Heidelberg, DE, pp. 59–68.

8. Fabrizio Sebastiani, Alessandro Sperduti and Nicola Valdambrini. An improved boosting algorithm and its application to automated text categorization. In *Proceedings of CIKM-00, 9th ACM International Conference on Information and Knowledge Management*, Washington, US, 2000, pp. 78–85. ACM Press, New York, US.

9. Giuseppe Attardi, Antonio Gullì, and Fabrizio Sebastiani. Automatic Web Page Categorization by Link and Context Analysis. In Chris Hutchison and Gaetano Lanzarone (eds.), *Proceedings of THAI'99, European Symposium on Telematics, Hypermedia and Artificial Intelligence*, Varese, IT, 1999, pp. 105–119.