Text Classification

A course to be delivered at WebBar'07
August 27–31, 2007 – Varenna, Italy

Fabrizio Sebastiani
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
Via G. Moruzzi, 1 – 56124 Pisa, Italy
E-mail: fabrizio.sebastiani@isti.cnr.it
Phone: +39 050 315 2892
Fax: +39 050 315 3464

Basic info

The level of the tutorial is **introductory/intermediate**. No prerequisites are needed, apart from a generic knowledge of IR fundamentals; basic concepts of machine learning will be explicitly introduced as they are needed.

The duration of the tutorial is **half-day**.

About the instructor

Since 2000 Fabrizio Sebastiani's main research interests have been at the intersection among information retrieval, machine learning, and human language technology, with particular emphasis on text classification, text clustering, and applications of text classification such as lexicon learning, sentiment classification, and survey coding. On these and other topics he has published several articles in international journals, conferences, and edited collections. He has guest-edited a special issue on automated text classification of the Journal of Intelligent Information Systems [JS02], and has been the Chairman of the ACM SIGIR 2002 Workshop on Operational Text Classification Systems. He has been the Area Chair for "Machine Learning for IR, Text Data Mining, Clustering, Text Categorization" at SIGIR'03, SIGIR'04, SIGIR'05, SIGIR'07. On text classification he has given several tutorials at international conferences and courses at summer schools, among which the IJCAI, COLING, ECDL conferences, and the ESSLLI and ESSIR summer schools. His review article "Machine learning in automated document categorisation" [Seb02] is, as of today, the most quoted article in *ACM Computing Surveys* since 2000, totalling more than 1000 quotations according to Google Scholar.

Fabrizio Sebastiani is the Editor-in-Chief (with Jamie Callan) of the newly launched journal *Foundations and Trends in Information Retrieval* (Now Publishers). He is currently a member of the Editorial Boards of the *Journal of the American Society for Information Science and Technology*, *Information Retrieval*, *Information Processing and Management*, and *ACM Transactions on Information Systems* journals; he has also been a reviewer for more than 30 different international journals. In 2003 he was program chairman of the 25th European Conference on Information Retrieval (ECIR-03). Since July 2003 to June 2007, he has been the Vice-Chairman of ACM SIGIR. He is a Program co-Chairman of the upcoming ACM SIGIR 2008 conference.

Some of his recent publications related to text classification are listed at the end of this document.

Text Classification

Tutorial Proposal for SIGIR'06

1 Theme of the tutorial

Text classification (also known as text categorization) is the task of automatically sorting a set of documents into categories from a predefined set. This task has several applications, including automated indexing of scientific articles according to predefined thesauri of technical terms, filing patents into patent directories, selective dissemination of information to information consumers, automated population of hierarchical catalogues of Web resources, spam filtering, identification of document genre, authorship attribution, survey coding, and even automated essay grading. Automated text classification is attractive because it frees organizations from the need of manually organizing document bases, which can be too expensive, or simply not feasible given the time constraints of the application or the number of documents involved. The accuracy of modern text classification systems rivals that of trained human professionals, thanks to a combination of information retrieval (IR) technology and machine learning (ML) technology. This tutorial will outline the fundamental traits of the technologies involved, of the applications that can feasibly be tackled through text classification, and of the tools and resources that are available to the researcher and developer wishing to take up these technologies for deploying real-world applications.

2 Objective of the tutorial

The objective of this tutorial is to make the attendees aware of the concepts and techniques for automatically or semi-automatically classifying documents into a set of topical categories, and to review the most recent trends and techniques. The course is addressed at students, researchers, and practitioners active interested in the application of quantitative techniques for automatically dealing with large corpora of texts.

3 Detailed contents of the tutorial

This section details the contents of the course, including approximate timing information.

1. A definition of the text classification (TC) task [20 min]
   (a) Single-label vs. multi-label TC
   (b) "Hard" vs. "soft" TC
2. Applications of TC [30 min]
   (a) Automatic indexing for Boolean information retrieval
   (b) Spam filtering
   (c) Focused crawling
   (d) Authorship attribution and genre classification
   (e) Other applications
3. The machine learning approach to TC [10 min]
   (a) Training set and test set
   (b) The architecture of a TC system
4. Indexing and dimensionality reduction [40 min]
   (a) Dimensionality reduction

   (b) Term selection
   (c) Term extraction

5. Methods for the inductive construction of a classifier [60 min]
   (a) Probabilistic models
   (b) Regression models
   (c) Decision tree classifiers
   (d) Inductive rule learning (in Disjunctive Normal Form)
   (e) On-line and batch models for linear classifiers
   (f) Example-based classifiers
   (g) Kernel methods and support vector machines
   (h) Boosting methods
   (i) Semi-supervised methods

6. Evaluating TC algorithms [30 min]
   (a) Precision and recall
   (b) Combinations of precision and recall
   (c) Other measures of TC effectiveness

7. Hierarchical TC and hypertext classification [40 min]
   (a) The hierarchical nature of the set of categories
   (b) Local selection of negative examples
   (c) Early pruning
   (d) Children-oriented feature selection
   (e) The presence of hypertextual pointers

8. Conclusion [10 min]
   (a) Current trends and future directions
   (b) Pointers to open-source and public-domain software
   (c) Pointers to evaluation campaigns and available test collections
   (d) Pointers to bibliographic references

References

[ALSZ06] Henri Avancini, Alberto Lavelli, Fabrizio Sebastiani, and Roberto Zanoli. Automatic expansion of domain-specific lexicons by term categorization. *ACM Transactions on Speech and Language Processing*, 3(1):1–30, 2006.

[ARS04] Henri Avancini, Andreas Rauber, and Fabrizio Sebastiani. Organizing digital libraries by automated text categorization. In *Proceedings of the International Conference on Digital Libraries (ICDL'04)*, pages 919–931, New Delhi, IN, 2004. Invited talk.

[CMS01] 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*, pages 78–102. Idea Group Publishing, Hershey, US, 2001.

[DS03] Franca Debole and Fabrizio Sebastiani. Supervised term weighting for automated text categorization. In *Proceedings of the 18th ACM Symposium on Applied Computing (SAC'03)*, pages 784–788, Melbourne, US, 2003. ACM Press, New York, US.

[DS05] Franca Debole and Fabrizio Sebastiani. An analysis of the relative hardness of Reuters-21578 subsets. *Journal of the American Society for Information Science and Technology*, 56(6):584–596, 2005.

[EFS06a] Andrea Esuli, Tiziano Fagni, and Fabrizio Sebastiani. MP-Boost: A multiple-pivot boosting algorithm and its application to text categorization. In *Proceedings of the 13th International Symposium on String Processing and Information Retrieval (SPIRE'06)*, pages 1–12, Glasgow, UK, 2006.

[EFS06b] Andrea Esuli, Tiziano Fagni, and Fabrizio Sebastiani. TreeBoost.MH: A boosting algorithm for multi-label hierarchical text categorization. In *Proceedings of the 13th International Symposium on String Processing and Information Retrieval (SPIRE'06)*, pages 13–24, Glasgow, UK, 2006.

[FS07] Tiziano Fagni and Fabrizio Sebastiani. On the selection of negative examples for hierarchical text categorization. In *Proceedings of the 3rd Language & Technology Conference (LTC'07)*, Poznań, PL, 2007. Forthcoming.

[GS03] Daniela Giorgetti and Fabrizio Sebastiani. Automating survey coding by multiclass text categorization techniques. *Journal of the American Society for Information Science and Technology*, 54(14):1269–1277, 2003.

[GSS00] 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, editors, *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'00)*, pages 59–68, Lisbon, PT, 2000. Published in the “Lecture Notes for Computer Science” series, number 1923, Springer Verlag, Heidelberg, DE.

[JS02] Thorsten Joachims and Fabrizio Sebastiani, editors. *Journal of Intelligent Information Systems.* (Kluwer Academic Publishers, Dordrecht, NL), 18(2). Special Issue on Automated Text Categorization, 2002.

[NSS03] Pio Nardiello, Fabrizio Sebastiani, and Alessandro Sperduti. Discretizing continuous attributes in AdaBoost for text categorization. In Fabrizio Sebastiani, editor, *Proceedings of the 25th European Conference on Information Retrieval (ECIR'03)*, pages 320–334, Pisa, IT, 2003.

[Seb02] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[Seb03] Fabrizio Sebastiani. Research in automated classification of texts: Trends and perspectives. In *Proceedings of the 4th International Colloquium on Library and Information Science (ICLIS'03)*, Salamanca, SP, 2003. Invited talk.

[Seb05a] Fabrizio Sebastiani. Text categorization. In Alessandro Zanasi, editor, *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, pages 109–129. WIT Press, Southampton, UK, 2005.

[Seb05b] Fabrizio Sebastiani. Text categorization. In Laura C. Rivero, Jorge H. Doorn, and Viviana E. Ferraggine, editors, *The Encyclopedia of Database Technologies and Applications*, pages 683–687. Idea Group Publishing, Hershey, US, 2005.

[Seb06] Fabrizio Sebastiani. Classification of text, automatic. In Keith Brown, editor, *The Encyclopedia of Language and Linguistics*, volume 2, pages 457–463. Elsevier Science Publishers, Amsterdam, NL, second edition, 2006.

[SSV00] Fabrizio Sebastiani, Alessandro Sperduti, and Nicola Valdambrini. An improved boosting algorithm and its application to automated text categorization. In Arvin Agah, Jamie Callan, and Elke Rundensteiner, editors, *Proceedings of the 9th ACM International Conference on Information and Knowledge Management (CIKM'00)*, pages 78–85, McLean, US, 2000.