Associazione Italiana per l'Intelligenza Artificiale

Atti del Workshop su
Proceedings of the Workshop on

Intelligenza Artificiale per i Beni Culturali e le Biblioteche Digitali

Artificial Intelligence for the Cultural Heritage and Digital Libraries

a cura di Luciana Bordoni
Giovanni Semeraro

Martedì, 25 Settembre 2001
Dipartimento di Informatica (CAMPUS)
UNIVERSITÁ DEGLI STUDI DI BARI

Organizing and Using Digital Libraries by Automated Text Categorization

Fabrizio Sebastiani
Istituto di Elaborazione dell'Informazione
Consiglio Nazionale delle Ricerche
56124 Pisa, Italy
E-mail: fabrizio@iei.pi.cnr.it

> 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]

In the short story *The Library of Babel*, Jorge Luis Borges envisions the world as a library of infinite size which contains *all* possible books, i.e. all possible sequences of symbols that can be built out of a given alphabet. Some of these sequences are totally random juxtapositions of symbols, with no morphological or syntactic well-formedness according to any known language; some others are morphologically and syntactically well-formed according to some known language, but are not meaningful; still some others are meaningful but untruthful; and some are both meaningful and truthful.

In such an all-encompassing, chaotic, unorganized library, humans are engaged in the endless quest for truth, which takes the form of the quest not of generically meaningful and truthful books, but of *the* book, the one that reveals eternal truth. Of course, if there is such an eternal truth, the Library contains such a book, since it contains them all: this justifies the "extravagant happiness" of humans, but the formidable character of this task also justifies their "depression".

Borges' death in 1985 sadly deprived him of the chance to know about the World Wide Web, definitely the most faithful enactment of his Library that has been realized to date, and probably one of the most faithful we can ever think of. Any Web user has surely experienced the "happiness" and "depression" Borges speaks of, in first realizing that the Web contains enormous amounts of useful information just a few clicks away, and in then realizing that without appropriate tools one might need to access huge quantities of irrelevant information before hitting on the relevant items.

The disciplines of *Information Retrieval* (IR) and *Automated Text Categorization* (ATC) have attacked the problem of *information overload* from two orthogonal perspectives. While IR strives to provide *better search tools* for seeking information in an unstructured collection of documents, the purpose of ATC is that of automatically providing *better structuring* of this collection so as to make search easier; it is on this latter discipline that this talk will concentrate.

ATC (see [6] for an introduction and review) is the discipline concerned with the construction of *automatic text classifiers*, i.e. programs capable of assigning to a document one or more among a set of predefined categories. Building these classifiers is itself done automatically, by means of a general inductive process that "learns" the characteristics of the categories from a set of preclassified documents.

ATC has a number of applications, some of them quite esoteric, including word sense disambiguation [2], Web page classification under hierarchical Internet directories [1], author identification for literary texts of unknown or disputed authorship [3], automated identification of text genre [4], and automated essay grading [5]. In this talk I will discuss two classes of applications, *automatic indexing with controlled vocabularies* and *personalized information delivery*, that are of direct concern to organizing and using digital libraries, respectively. I will discuss these two classes of applications by drawing from two projects ongoing at IEI-CNR.

The first project, codenamed COMPCAT, is an internally funded project concerned with building an interactive classifier of scientific articles in the computer science domain. Here, an author submits an article to the classifier, and the classifier suggests to the author a list of categories (drawn from the ACM Classification Scheme) ranked in order of estimated suitability to the article. The method by which classifiers are built takes advantage of the hierarchical structure of the ACM Classification Scheme.

The second project, codenamed CYCLADES, is a project funder by the CEC under the IST program that aims at building a layer of user services on top of an OAI-compliant distributed digital library. Here, text categorization is used for providing to the user personalized tools for information access: records obtained from the digital library (either at the user's or at the system's initiative) are classified into a hierarchy of topical folders according to the perceived meaning that the user attributes to these folders, where this meaning is automatically derived from either implicit or explicit user feedback.

References

[1] Susan T. Dumais and Hao Chen. Hierarchical classification of Web content. In Nicholas J. Belkin, Peter Ingwersen, and Mun-Kew Leong, editors, *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval*, pages 256–263, Athens, GR, 2000. ACM Press, New York, US.

[2] Gerard Escudero, Lluís Màrquez, and German Rigau. Boosting applied to word sense disambiguation. In Ramon López de Mántaras and Enric Plaza, editors, *Proceedings of ECML-00, 11th European Conference on Machine Learning*, pages 129–141, Barcelona, ES, 2000. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1810.

[3] Richard S. Forsyth. New directions in text categorization. In Alex Gammerman, editor, *Causal models and intelligent data management*, pages 151–185. Springer Verlag, Heidelberg, DE, 1999.

[4] Brett Kessler, Geoff Nunberg, and Hinrich Schütze. Automatic detection of text genre. In Philip R. Cohen and Wolfgang Wahlster, editors, *Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics*, pages 32–38, Madrid, ES, 1997. Morgan Kaufmann Publishers, San Francisco, US.

[5] Leah S. Larkey. Automatic essay grading using text categorization techniques. In W. Bruce Croft, Alistair Moffat, Cornelis J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 90–95, Melbourne, AU, 1998. ACM Press, New York, US.

[6] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 2002. Forthcoming.