Maristella Agosti (Ed.)

Information Access through Search Engines and Digital Libraries

INRE

Preface

The Information Management Systems (IMS) Research Group was formed in the Department of Information Engineering (formerly Department of Electronics and Computer Science) of the University of Padua, Italy, in 1987 when the department was established. The group activities are concerned with the design, modeling and implementation of advanced information retrieval tools – such as search engines – and digital library systems.

The main aims of the IMS research group are:

- to contribute to the advancement of basic and applied research in the area of the management of digital information in its diverse and multiform materializations, by participating in national and European research projects, together with serving on editorial boards of international journals and program committees of international and European conferences and initiatives,
- to provide a good environment to facilitate the acquisition of knowledge of information management foundations to graduate and undergraduate students, and actively organize workshops and summers schools devoted to the transfer of competence to young researchers and particularly PhD students, and
- to make possible the transfer of results and expertise to industry and public organizations of the region surrounding Padua, but also at the national and European level.

The IMS research on Information Retrieval addresses theoretical methods and experimental approaches for the efficient and effective gathering, indexing, organization and retrieval of all and only the multimedia data that are relevant to users' information needs from large document collections. Specific research efforts have been directed to solve crucial aspects of retrieval of information from the Web and to design and implement effective search engines.

The IMS research on Digital Libraries addresses methods, systems, and tools to build and make available effective and distributed digital library

management systems to end users. What makes Digital Libraries different from the Web is the rigorous and sophisticated organisation of heterogeneous multimedia data, and the need to support distributed, fast and easy access to those data.

Through its long-standing tradition in information retrieval, and more recently in digital libraries, the group has gained a strong reputation at the national and European level, and it has good relationships with many outstanding researchers worldwide.

The papers in this book report the original research results built on the past work of the group which open up new directions and new areas of possible fruitful cooperation in the context of new research projects.

In the initial paper, *Information Access using the Guide of User Requirements*, Agosti presents an interpretation of the evolution of events in the information retrieval area. Focusing mainly on the last twenty years, the paper pays particular attention to the system which needs to be envisaged and designed to support the end user in accessing relevant and interesting information. The end user is considered the guide of the researcher, since he prompts the researcher to conceive and invent solutions of real use for the user himself.

Crestani's paper, *From Linking Text to Linking Crimes: Information Retrieval, But Not As You Know It*, proves that information retrieval techniques that have been used for a long time to identify links between textual items for the automatic construction of hypertexts and electronic books are proving of great value in different application areas. Crestani presents an approach to automatic linking of textual items that is used to prioritise criminal suspects in a police investigation. Crimes are linked to each other and to suspects in a conceptual model that closely resembles the one used to design hypertexts. A free-text description of an unsolved crime is compared to previous offence descriptions where the offender is known. By linking the descriptions, inferences about likely suspects can be made. Language Modeling is adapted to produce a Bayesian model which assigns a probability to each suspect. The model presented in this paper could be easily extended to take account of additional crime and suspect linking data, such as geographical location of crimes or suspect social networks. This would enable large networks of investigative information automatically constructed from police archives to be browsed.

Melucci's paper, *Modeling Retrieval and Navigation in Context*, addresses the topic of *context* in information retrieval; in fact current information retrieval systems are designed and implemented to retrieve all and only the documents relevant to the information need expressed by the user without considering the context in which the user is when asking for information. But what is relevant to one user in one place at one time may no longer be relevant to another user, in another place or at another time. This means that an information retrieval system should be context-aware. In practice classical systems, such as search engines, are unaware of such a highly dynamic search environment and contextual features are not captured at indexing-time, nor

are they exploited at retrieval-time. Melucci's paper describes a model for navigation and search in context, that is, the navigation and search which adapts the retrieval results according to what the end user does during interaction. As hypertext is the main information retrieval tool for navigating information spaces, so the paper illustrates how the model can be applied when automatic links are built for document collections or electronic books, although the model is more general and can be applied in the future to different domains.

In the following paper, *Two Algorithms for Probabilistic Stemming*, Melucci and Orio face a central topic of information retrieval, that of the *stemming* which is performed to allow words, which are formulated with morphological variants, to group up with each other, indicating a similar meaning. Most stemming algorithms reduce word forms to an approximation of a common morphological root, called "stem", so that a relevant document can be retrieved even if the morphology of their own words is different from the one of the words of a given query. Two probabilistic stemming algorithms are presented and the main conclusion of the reported research is that a stemmer can be built for many European languages without much linguistic knowledge and with simple probabilistic models. This scientific conclusion has been confirmed by various experiments carried out using different standard test collections and it opens up interesting possibilities in a European and international setting where the collections of documents that an information retrieval tool has to manage are most of the time multilingual.

The paper by Di Nunzio, *Automated Text Categorization: The Two-Dimensional Probabilistic Model*, presents the *Two-Dimensional Probabilistic Model (2DPM)*, which is a retrieval model able to represent documents on a two-dimensional space. The model has proved to be a valid visualization tool for understanding the relationships between categories of textual documents, and for helping users to visually audit the classifier and identify suspicious training data. In addition, the model has the advantage of needing neither a reduction of the vocabulary of terms to reduce the complexity of the problem nor smoothing of probabilities to avoid zero probabilities during calculations. The paper shows that it is possible to address the modeling in information retrieval from a probabilistic point of view, using an approach that can be a little obscure to the end user, since it is an approach that operates server-side and makes the system, which is implemented making use of it, like a black box to the user; nevertheless, it can be used as a tool for better presenting the retrieved documents to the final user. The graphical representation of the groups of documents that are retrieved by the system can be further applied to the representation of documents that are given as output to users of information retrieval and digital library systems supporting them with graphical interfaces that give directions and help in the selection of documents of interest from among the many retrieved by the system.

The next paper, *Analysis of Web Link Analysis Algorithms: The Mathematics of Ranking* by Pretto, addresses relevant aspects of the algorithms

that have been designed to enhance the performance of Web search engines by exploiting the topological structure of the digraph associated with the Web. Link analysis algorithms are now also used in many other fields, sometimes far removed from that of Web searching. In many of their applications, their *ranking* capabilities are of prime importance; from here the need arises to perform a mathematical analysis of these algorithms from the perspective of the *rank* they induce on the nodes of the digraphs on which they work. Pretto's paper investigates the main theoretical results for the questions that arise when ranking is under investigation and some novel extensions are presented.

In *Digital Annotations: a Formal Model and its Applications* Ferro focuses on the rich and elusive concept of *annotation*. Even though the concept of annotation is familiar to us, it turns out to be particularly elusive when it comes to being explicitly and formally defined, mainly because it is a far more complex and multifaceted concept than one might imagine at a first glance. There are different viewpoints about annotation, which are often considered as separated, and this situation prevents us from exploiting synergies and complementarities among the different approaches, and makes it more difficult to determine what the differences between annotations and other concepts are, and what the advantages or disadvantages of using annotations are, even when they seem so similar to other concepts. The paper discusses different perspectives and presents a formal model that formalizes the main concepts concerning annotations and defines the relationships between annotations and annotated information resources. In addition, this formal model constitutes the necessary tool that can be used to design and implement search algorithms that can make use of annotations to enhance retrieval capabilities from multimedia distributed collections of digital content.

In *Music Indexing and Retrieval for Multimedia Digital Libraries* Orio addresses the topic of multimedia retrieval based on content with a focus on *music information retrieval*. Starting from the consideration that users of multimedia digital libraries have different levels of knowledge and expertise – and this is particularly true for music, where the level of music education may vary remarkably among users, who may range from casual listeners to performers and composers – untrained users may not be able to use bibliographic values or take advantage of metadata when searching for music, the consequence is that the access to music digital libraries should really be content-based. The main idea underlying content-based access and retrieval is that a document can be described by a set of features that are directly computed from its content. This approach is the basis for most of the methodologies for information retrieval, where the content of a textual document is automatically processed and used for indexing and retrieval. Even if multimedia data require specific methodologies for content extraction, the core information retrieval techniques developed for text may be extended to other media. The paper presents a novel methodology based on approximate indexing of music documents. The basic idea is to merge the positive effects of document indexing in terms of efficiency and scalability, with the positive effects of approximate matching in

terms of robustness to local mismatches. The methodology has been tested with encouraging results for future developments and applications.

The paper *A Statistical and Graphical Methodology for Comparing Bilingual to Monolingual Cross-Language Information Retrieval* by Crivellari Di Nunzio and Ferro is a collaborative study of cross-lingual *Information Retrieval Systems (IRSs)* and on a deep analysis of performance comparison between systems which perform monolingual tasks, i.e. querying and finding documents in one language, with respect to those which perform bilingual tasks, i.e. querying in one language and finding documents in another language. The study aims at improving the way of comparing bilingual and monolingual retrieval and strives to provide better methods and tools for assessing the performances. Another aspect of the work is that it can help the organizers of an evaluation forum during the topic generation process; in particular, the study of the hardness of a topic can be carried out with the goal of refining those topics which have been misinterpreted by systems. The authors propose a twofold methodology which exploits both thorough statistical analyses and graphical tools: the former will provide *MultiLingual Information Access (MLIA)* researchers with quantitative and more sophisticated analysis techniques, while the latter will allow for a more qualitative comparison and an easier presentation of the results. Concrete examples about how the proposed methodology can be applied by studying the monolingual and bilingual tasks of the *Cross-Language Evaluation Forum (CLEF)* 2005 and 2006 campaigns are provided.

As the reader can appreciate in reading the book, all the papers contribute towards the development of the area of *information access* to digital contents and give insights to future possible enhancements of information management systems.

I would like to thank Alexander Cormack, whose revisions have undoubtedly added to the quality of the work presented in this volume.

Finally, I would like to acknowledge the faculty and students that over the years have cooperated with the IMS group and that have collaborated in its achievements.

Padua, August 2007 *Maristella Agosti*