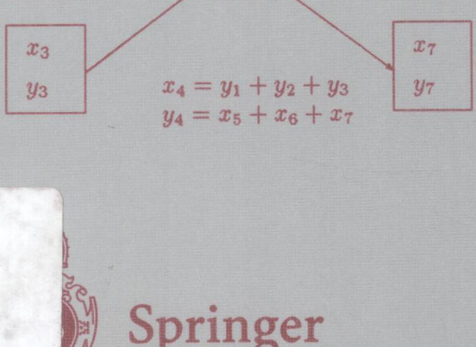
Lecture Notes in Computer Science 1980

Maristella Agosti Fabio Crestani
Gabriella Pasi (Eds.)

# Lectures on Information Retrieval

## Third European Summer-School, ESSIR 2000
Varenna, Italy, September 2000
Revised Lectures



$$x_4 = y_1 + y_2 + y_3$$
$$y_4 = x_5 + x_6 + x_7$$

Springer

Maristella Agosti   Fabio Crestani
Gabriella Pasi (Eds.)

# Lectures on Information Retrieval

Third European Summer-School, ESSIR 2000
Varenna, Italy, September 11-15, 2000
Revised Lectures

Springer

Volume Editors

Maristella Agosti
Universitá di Padova, Dipartimento di Elettronica e Informatica
Via Ognissanti, 72, 35131 Padova
E-mail: agosti@dei.unipd.it

Fabio Crestani
University of Strathclyde, Department of Computer Science
Glasgow G1 1XH, Scotland, UK
E-mail: fabioc@cs.strath.ac.uk

Gabriella Pasi
ITIM, Consiglio Nazionale delle Ricerche
Via Ampere, 56, 20131 Milano, Italy
E-mail: gabriella.pasi@itim.mi.cnr.it

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Lectures on information retrieval : third European summerschool ;
revised lectures / ESSIR 2000, Varenna, Italy, September 11 - 15,
2000. Maristella Agosti ... (ed.). - Berlin ; Heidelberg ; New York ;
Barcelona ; Hong Kong ; London ; Milan ; Paris ; Singapore ; Tokyo :
Springer, 2001
(Lecture notes in computer science ; Vol. 1980)
ISBN 3-540-41933-0

CR Subject Classification (1998): H.3, H.4, H.5, C.2.4, I.2,1

ISSN 0302-9743
ISBN 3-540-41933-0 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

Typesetting: Camera-ready by author, data conversion by Christian Grosche, Hamburg
Printed on acid-free paper SPIN 10781284 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science 1980
Edited by G. Goos, J. Hartmanis and J. van Leeuwen

Springer
*Berlin*
*Heidelberg*
*New York*
*Barcelona*
*Hong Kong*
*London*
*Milan*
*Paris*
*Singapore*
*Tokyo*

# Preface

*Information retrieval* (IR) is concerned with the effective and efficient retrieval of information based on its semantic content. The central problem in IR is the quest to find the set of relevant documents, among a large collection, containing the information sought, thereby satisfying a user's information need usually expressed in a natural language query. Documents may be objects or items in any medium: text, image, audio, or indeed a mixture of all three.

This book contains the proceedings of the *Third European Summer School in Information Retrieval (ESSIR 2000)*, held on 11–15 September 2000, in Villa Monastero, Varenna, Italy.

The event was jointly organised by the Institute of Multimedia Technologies of the CNR (National Council of Research) based in Milan (Italy), the Department of Electronics and Computer Science of the University of Padova (Italy), and the Department of Computer Science of the University of Strathclyde, Glasgow (UK). Administrative support was provided by Milano Ricerche, a consortium of industries, research institutions and the University of Milano, whose purpose is to provide administrative and technical support for the research and development activities of its members.

This third edition of the European Summer School in Information Retrieval is part of the ESSIR series which began in 1990. The first was organised by Maristella Agosti of the University of Padova and was held in Bressanone (Italy) in 1990. The second ESSIR was organised by Keith van Rijsbergen of the University of Glasgow (UK) and held in Glasgow in 1995, in the context of the IR Festival.

At the time of the first ESSIR, the Internet did not exist, so there is no website available for this event, but from its second edition a web presentation has been made available: the URL for ESSIR'95 is: `http://www.dcs.gla.ac.uk/essir/`, and the URL for ESSIR 2000 is: `http://www.itim.mi.cnr.it/Eventi/essir2000/index.htm`. These websites contain useful material. In particular, the ESSIR 2000 website contains copies of the material distributed at the school (presentation, notes, etc.).

The aim of ESSIR 2000 was to give participants a grounding in the core subjects of IR, including methods and techniques for designing and developing IR systems, web search engines, and tools for information storing and querying in digital libraries. To achieve these aims, the program of ESSIR 2000 was organised into a series of lectures divided into foundations and advanced parts as reported in the next section. The lecturers were leading European researchers (with only one non-European exception), their course subjects strongly reflecting the research work for which they are all well known.

ESSIR 2000 was intended for researchers starting out in IR, for industrialists who wish to know more about this increasingly important topic and for people working on topics related to the management of information on the Internet. This book, distributed at the school in draft form to incorporate in the final version useful participants' comments, contains 12 chapters written by the school's lecturers, providing surveys of the state of the art of IR and related areas.

## Book Structure

The ESSIR 2000 programme of lectures and this book are divided into in two parts: one part on the foundations of IR and related areas (e.g. digital libraries), and one on advanced topics.

The part on foundations contains seven papers/chapters. In Chap. 1, Keith van Rijsbergen introduces some underlying concepts and ideas essential for understanding IR research and techniques. He also highlights some related hot areas of research, emphasising the role of IR in each. In Chap. 2, Norbert Fuhr presents the main mathematical models of IR. This paper provides the theoretical basis for representing the informative content of documents and for estimating the relevance of a document to a query. In Chap. 3, Páraic Sheridan and Carol Peters detail the issues and proposed solutions for multilingual information access in digital archives. Chapter 4, by Stephen Robertson, addresses the topic of evaluation, a very important aspect of IR. In Chap. 5 and 6, Alan Smeaton and John Eakins address issues and techniques related to indexing, browsing and searching multimedia information (audio, image, or digital video). Finally, in Chap. 7 Ingeborg Solvberg covers the basics and the challenges of digital libraries.

The part on advanced topics contains five papers/chapters. In Chap. 8, Peter Ingwersen concentrates on user issues and the usability of interactive IR. Chap. 9, by Fabio Crestani and Mounia Lalmas addresses the use of logic and uncertainty theories in IR. Closely related is Chap. 10, by Gabriella Pasi and Gloria Bordogna, which presents the area of research that aims at modelling the vagueness and imprecision involved in the IR process. In Chap. 11, Maristella Agosti and Massimo Melucci address the use of IR techniques on the Web for searching and browsing. Finally, in Chap. 12, Yves Chiaramella addresses the issues related to indexing and retrieval of structured documents.

## Acknowledgements

The editors would like to thank all the participants of ESSIR 2000 for making the event a success. ESSIR 2000 was a success not just for the quality of the lectures, the authority of the lecturers, and the beautiful surroundings, it was a success because it was informal and interactive. For the best part of a week, more than 60 participants and 12 lecturers exchanged ideas and inspirations on where IR is at and where it should go. Many attendants (not just school participants, but some of the lecturers too) returned home with renewed encouragement and motivation.

We thank the sponsoring and supporting institutions for making it possible, financially, to hold the event. Also, we thank the Local Organising Committee,

the student volunteers and the personnel of Villa Monastero (Rino Venturini) for their invaluable help.

A special thanks to all the lecturers for their contributions, encouragement, and support. The quality of this book is mostly due to their work.

Finally, we would like to thank the Board of the Special Interest Network on Information Retrieval of the Council of European Professional Informatics Societies (CEPIS-IR), which includes Keith van Rijsbergen, Norbert Fuhr and Alan Smeaton, for their scientific support and invaluable advice on the school content and program.

September 2000

Maristella Agosti
Fabio Crestani
Gabriella Pasi

# Organisation and Support

## Scientific Program and Organising Committee

ESSIR 2000 was jointly organised by:

- Maristella Agosti, Department of Electronics and Computer Science, University of Padova, Padova, Italy;
- Fabio Crestani, Department of Computer Science, University of Strathclyde, Glasgow, UK;
- Gabriella Pasi, Institute of Multimedia Technologies, National Council of Research (CNR), Milan, Italy.

## Local Organising Committee

ESSIR 2000 was locally organised by the Institute of Multimedia Technologies of CNR in Milan, Italy. In particular by: Gabriella Pasi, Gloria Bordogna, Paola Carrara, Alba L'Astorina, Luciana Onorato and Bruna Zonta.

## Sponsoring Institutions

The main sponsoring and supporting organisation was the Special Interest Network on Information Retrieval of the Council of European Professional Informatics Societies (CEPIS-IR). CEPIS-IR provided a running grant, which made it possible to award a number of bursaries to support young students and researchers to attend the school. CEPIS-IR also provided invaluable advice on the school program.
The other sponsors were:

- Arnoldo Mondadori Editore, Verona, Italy;
- Microsoft Italia, Milan, Italy;
- Oracle Italia, Milan, Italy;
- Sharp Laboratories of Europe, Oxford, UK;
- 3D Informatica, San Lazzaro di Savena (Bologna), Italy.

## Supporting Institutions

ESSIR 2000 benefited from the support of the following organisations:

- CEPIS-IR (Special Interest Network on Information Retrieval of the Council of European Professional Informatics Societies);
- AEI (Gruppo Specialistico Tecnologie e Applicazioni Informatiche);
- EUREL (Convention of National Societies of Electrical Engineers of Europe).

# Lecture Notes in Computer Science

For information about Vols. 1–1944
please contact your bookseller or Springer-Verlag

Vol. 1945: W. Grieskamp, T. Santen, B. Stoddart (Eds.), Integrated Formal Methods. Proceedings, 2000. X, 441 pages. 2000.

Vol. 1946: P. Palanque, F. Paternò (Eds.), Interactive Systems. Proceedings, 2000. X, 251 pages. 2001.

Vol. 1947: T. Sørevik, F. Manne, R. Moe, A.H. Gebremedhin (Eds.), Applied Parallel Computing. Proceedings, 2000. XII, 400 pages. 2001.

Vol. 1948: T. Tan, Y. Shi, W. Gao (Eds.), Advances in Multimodal Interfaces – ICMI 2000. Proceedings, 2000. XVI, 678 pages. 2000.

Vol. 1949: R. Connor, A. Mendelzon (Eds.), Research Issues in Structured and Semistructured Database Programming. Proceedings, 1999. XII, 325 pages. 2000.

Vol. 1950: D. van Melkebeek, Randomness and Completeness in Computational Complexity. XV, 196 pages. 2000.

Vol. 1951: F. van der Linden (Ed.), Software Architectures for Product Families. Proceedings, 2000. VIII, 255 pages. 2000.

Vol. 1952: M.C. Monard, J. Simão Sichman (Eds.), Advances in Artificial Intelligence. Proceedings, 2000. XV, 498 pages. 2000. (Subseries LNAI).

Vol. 1953: G. Borgefors, I. Nyström, G. Sanniti di Baja (Eds.), Discrete Geometry for Computer Imagery. Proceedings, 2000. XI, 544 pages. 2000.

Vol. 1954: W.A. Hunt, Jr., S.D. Johnson (Eds.), Formal Methods in Computer-Aided Design. Proceedings, 2000. XI, 539 pages. 2000.

Vol. 1955: M. Parigot, A. Voronkov (Eds.), Logic for Programming and Automated Reasoning. Proceedings, 2000. XIII, 487 pages. 2000. (Subseries LNAI).

Vol. 1956: T. Coquand, P. Dybjer, B. Nordström, J. Smith (Eds.), Types for Proofs and Programs. Proceedings, 1999. VII, 195 pages. 2000.

Vol. 1957: P. Ciancarini, M. Wooldridge (Eds.), Agent-Oriented Software Engineering. Proceedings, 2000. X, 323 pages. 2001.

Vol. 1960: A. Ambler, S.B. Calo, G. Kar (Eds.), Services Management in Intelligent Networks. Proceedings, 2000. X, 259 pages. 2000.

Vol. 1961: J. He, M. Sato (Eds.), Advances in Computing Science – ASIAN 2000. Proceedings, 2000. X, 299 pages. 2000.

Vol. 1963: V. Hlaváč, K.G. Jeffery, J. Wiedermann (Eds.), SOFSEM 2000: Theory and Practice of Informatics. Proceedings, 2000. XI, 460 pages. 2000.

Vol. 1964: J. Malenfant, S. Moisan, A. Moreira (Eds.), Object-Oriented Technology. Proceedings, 2000. XI, 309 pages. 2000.

Vol. 1965: Ç. K. Koç, C. Paar (Eds.), Cryptographic Hardware and Embedded Systems – CHES 2000. Proceedings, 2000. XI, 355 pages. 2000.

Vol. 1966: S. Bhalla (Ed.), Databases in Networked Information Systems. Proceedings, 2000. VIII, 247 pages. 2000.

Vol. 1967: S. Arikawa, S. Morishita (Eds.), Discovery Science. Proceedings, 2000. XII, 332 pages. 2000. (Subseries LNAI).

Vol. 1968: H. Arimura, S. Jain, A. Sharma (Eds.), Algorithmic Learning Theory. Proceedings, 2000. XI, 335 pages. 2000. (Subseries LNAI).

Vol. 1969: D.T. Lee, S.-H. Teng (Eds.), Algorithms and Computation. Proceedings, 2000. XIV, 578 pages. 2000.

Vol. 1970: M. Valero, V.K. Prasanna, S. Vajapeyam (Eds.), High Performance Computing – HiPC 2000. Proceedings, 2000. XVIII, 568 pages. 2000.

Vol. 1971: R. Buyya, M. Baker (Eds.), Grid Computing – GRID 2000. Proceedings, 2000. XIV, 229 pages. 2000.

Vol. 1972: A. Omicini, R. Tolksdorf, F. Zambonelli (Eds.), Engineering Societies in the Agents World. Proceedings, 2000. IX, 143 pages. 2000. (Subseries LNAI).

Vol. 1973: J. Van den Bussche, V. Vianu (Eds.), Database Theory – ICDT 2001. Proceedings, 2001. X, 451 pages. 2001.

Vol. 1974: S. Kapoor, S. Prasad (Eds.), FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science. Proceedings, 2000. XIII, 532 pages. 2000.

Vol. 1975: J. Pieprzyk, E. Okamoto, J. Seberry (Eds.), Information Security. Proceedings, 2000. X, 323 pages. 2000.

Vol. 1976: T. Okamoto (Ed.), Advances in Cryptology – ASIACRYPT 2000. Proceedings, 2000. XII, 630 pages. 2000.

Vol. 1977: B. Roy, E. Okamoto (Eds.), Progress in Cryptology – INDOCRYPT 2000. Proceedings, 2000. X, 295 pages. 2000.

Vol. 1978: B. Schneier (Ed.), Fast Software Encryption. Proceedings, 2000. VIII, 315 pages. 2001.

Vol. 1979: S. Moss, P. Davidsson (Eds.), Multi-Agent-Based Simulation. Proceedings, 2000. VIII, 267 pages. 2001. (Subseries LNAI).

Vol. 1808: S. Pande, D.P. Agrawal (Eds.), Compiler Optimizations for Scalable Parallel Systems. XXVII, 779 pages. 2001.

Vol. 1980: M. Agosti, F. Crestani, G. Pasi (Eds.), Lectures on Information Retrieval. Proceedings, 2000. XI, 311 pages. 2001.

Vol. 1983: K.S. Leung, L.-W. Chan, H. Meng (Eds.), Intelligent Data Engineering and Automated Learning – IDEAL 2000. Proceedings, 2000. XVI, 573 pages. 2000.

Vol. 1984: J. Marks (Ed.), Graph Drawing. Proceedings, 2001. XII, 419 pages. 2001.

Vol. 1985: J. Davidson, S.L. Min (Eds.), Languages, Compilers, and Tools for Embedded Systems. Proceedings, 2000. VIII, 221 pages. 2001.

Vol. 1987: K.-L. Tan, M.J. Franklin, J. C.-S. Lui (Eds.), Mobile Data Management. Proceedings, 2001. XIII, 289 pages. 2001.

Vol. 1988: L. Vulkov, J. Waśniewski, P. Yalamov (Eds.), Numerical Analysis and Its Applications. Proceedings, 2000. XIII, 782 pages. 2001.

Vol. 1989: M. Ajmone Marsan, A. Bianco (Eds.), Quality of Service in Multiservice IP Networks. Proceedings, 2001. XII, 440 pages. 2001.

Vol. 1990: I.V. Ramakrishnan (Ed.), Practical Aspects of Declarative Languages. Proceedings, 2001. VIII, 353 pages. 2001.

Vol. 1991: F. Dignum, C. Sierra (Eds.), Agent Mediated Electronic Commerce. VIII, 241 pages. 2001. (Subseries LNAI).

Vol. 1992: K. Kim (Ed.), Public Key Cryptography. Proceedings, 2001. XI, 423 pages. 2001.

Vol. 1993: E. Zitzler, K. Deb, L. Thiele, C.A.Coello Coello, D. Corne (Eds.), Evolutionary Multi-Criterion Optimization. Proceedings, 2001. XIII, 712 pages. 2001.

Vol. 1995: M. Sloman, J. Lobo, E.C. Lupu (Eds.), Policies for Distributed Systems and Networks. Proceedings, 2001. X, 263 pages. 2001.

Vol. 1997: D. Suciu, G. Vossen (Eds.), The World Wide Web and Databases. Proceedings, 2000. XII, 275 pages. 2001.

Vol. 1998: R. Klette, S. Peleg, G. Sommer (Eds.), Robot Vision. Proceedings, 2001. IX, 285 pages. 2001.

Vol. 1999: W. Emmerich, S. Tai (Eds.), Engineering Distributed Objects. Proceedings, 2000. VIII, 271 pages. 2001.

Vol. 2000: R. Wilhelm (Ed.), Informatics: 10 Years Back, 10 Years Ahead. IX, 369 pages. 2001.

Vol. 2001: G.A. Agha, F. De Cindio, G. Rozenberg (Eds.), Concurrent Object-Oriented Programming and Petri Nets. VIII, 539 pages. 2001.

Vol. 2002: H. Comon, C. Marché, R. Treinen (Eds.), Constraints in Computational Logics. Proceedings, 1999. XII, 309 pages. 2001.

Vol. 2003: F. Dignum, U. Cortés (Eds.), Agent Mediated Electronic Commerce III. XII, 193 pages. 2001. (Subseries LNAI).

Vol. 2004: A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing. Proceedings, 2001. XII, 528 pages. 2001.

Vol. 2006: R. Dunke, A. Abran (Eds.), New Approaches in Software Measurement. Proceedings, 2000. VIII, 245 pages. 2001.

Vol. 2007: J.F. Roddick, K. Hornsby (Eds.), Temporal, Spatial, and Spatio-Temporal Data Mining. Proceedings, 2000. VII, 165 pages. 2001. (Subseries LNAI).

Vol. 2009: H. Federrath (Ed.), Designing Privacy Enhancing Technologies. Proceedings, 2000. X, 231 pages. 2001.

Vol. 2010: A. Ferreira, H. Reichel (Eds.), STACS 2001. Proceedings, 2001. XV, 576 pages. 2001.

Vol. 2011: M. Mohnen, P. Koopman (Eds.), Implementation of Functional Languages. Proceedings, 2000. VIII, 267 pages. 2001.

Vol. 2013: S. Singh, N. Murshed, W. Kropatsch (Eds.), Advances in Pattern Recognition – ICAPR 2001. Proceedings, 2001. XIV, 476 pages. 2001.

Vol. 2015: D. Won (Ed.), Information Security and Cryptology – ICISC 2000. Proceedings, 2000. X, 261 pages. 2001.

Vol. 2018: M. Pollefeys, L. Van Gool, A. Zisserman, A. Fitzgibbon (Eds.), 3D Structure from Images – SMILE 2000. Proceedings, 2000. X, 243 pages. 2001.

Vol. 2020: D. Naccache (Ed.), Topics in Cryptology – CT-RSA 2001. Proceedings, 2001. XII, 473 pages. 2001

Vol. 2021: J. N. Oliveira, P. Zave (Eds.), FME 2001: Formal Methods for Increasing Software Productivity. Proceedings, 2001. XIII, 629 pages. 2001.

Vol. 2022: A. Romanovsky, C. Dony, J. Lindskov Knudsen, A. Tripathi (Eds.), Advances in Exception Handling Techniques. XII, 289 pages. 2001

Vol. 2024: H. Kuchen, K. Ueda (Eds.), Functional and Logic Programming. Proceedings, 2001. X, 391 pages. 2001.

Vol. 2026: F. Müller (Ed.), High-Level Parallel Programming Models and Supportive Environments. Proceedings, 2001. IX, 137 pages. 2001.

Vol. 2027: R. Wilhelm (Ed.), Compiler Construction. Proceedings, 2001. XI, 371 pages. 2001.

Vol. 2028: D. Sands (Ed.), Programming Languages and Systems. Proceedings, 2001. XIII, 433 pages. 2001.

Vol. 2029: H. Hussmann (Ed.), Fundamental Approaches to Software Engineering. Proceedings, 2001. XIII, 349 pages. 2001.

Vol. 2030: F. Honsell, M. Miculan (Eds.), Foundations of Software Science and Computation Structures. Proceedings, 2001. XII, 413 pages. 2001.

Vol. 2031: T. Margaria, W. Yi (Eds.), Tools and Algorithms for the Construction and Analysis of Systems. Proceedings, 2001. XIV, 588 pages. 2001.

Vol. 2034: M.D. Di Benedetto, A. Sangiovanni-Vincentelli (Eds.), Hybrid Systems: Computation and Control. Proceedings, 2001. XIV, 516 pages. 2001.

Vol. 2035: D. Cheung, G.J. Williams, Q. Li (Eds.), Advances in Knowledge Discovery and Data Mining – PAKDD 2001. Proceedings, 2001. XVIII, 596 pages. 2001. (Subseries LNAI).

Vol. 2037: E.J.W. Boers et al. (Eds.), Applications of Evolutionary Computing. Proceedings, 2001. XIII, 516 pages. 2001.

Vol. 2038: J. Miller, M. Tomassini, P.L. Lanzi, C. Ryan, A.G.B. Tettamanzi, W.B. Langdon (Eds.), Genetic Programming. Proceedings, 2001. XI, 384 pages. 2001.

Vol. 2040: W. Kou, Y. Yesha, C.J. Tan (Eds.), Electronic Commerce Technologies. Proceedings, 2001. X, 187 pages. 2001.

# Contents

# Getting into Information Retrieval

C.J. "Keith" van Rijsbergen

Department of Computing Science, University of Glasgow
Glasgow G12 8QQ, Scotland
keith@dcs.gla.ac.uk

**Abstract** This is a general introduction to Information Retrieval concentrating on some specific topics. I will begin by setting the scene for IR research and introduce its extensive experimental evaluation methodology. I will highlight some of the related areas of research which are currently in fashion emphasising the role of IR in each. For each introductory topic I will illustrate its relevance to IR in the context of a multimedia and multi-lingual environment where appropriate. I will also try and relate these topics to the other papers contained in this volume. My main purpose will be to introduce some underlying concepts and ideas essential for the understanding of IR research and techniques.

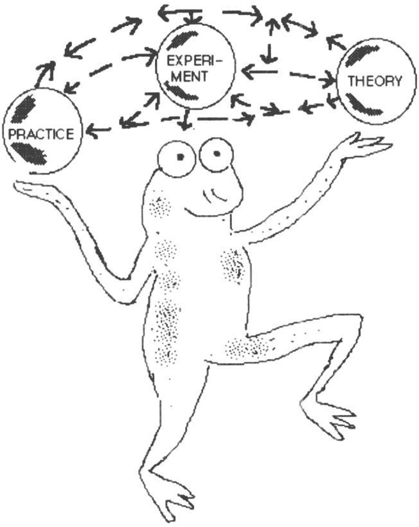
## 1 Introduction

As one who has been involved in information retrieval research since about 1969 it is wonderful to see how some of our work has been absorbed and adopted by a number of technologies. In particular it is fascinating to see how the development of the Web has spawned a number of exciting and unique IR research problems. In this paper I hope to touch on some of these but always from the perspective of an IR researcher who is looking to make connections between the IR research methodology and the interests of those focused on other technologies. For example, it is clear to us in the IR community that the web represents an emerging technology which encompasses much more than information retrieval, nevertheless, some of its important problems relate specifically to IR.

The history of IR is long and fraught [42]. For many years it was unclear whether it was a subject at all, then when it became a subject, it was claimed by both Information Science and Computer Science. Although in the early days during the 50's and 60's this difficulty was responsible for a number of frustrations, for example the unwillingness of librarians to accept hard experimental results, it now is also one of its strengths. We interact fruitfully, the information science community guarding us against technological, or system-based excesses, the computer science community representing a hard-nosed approach to experimental designs and being forced into taking user-interface issues seriously. A marriage made in heaven!

For years I have advocated the interplay of theory, practice, and experiment. My first serious attempt to talk about this was probably in a seminar presentation I gave in 1977 where I quoted the following from Freud:

M. Agosti, F. Crestani, and G. Pasi (Eds.): ESSIR 2000, LNCS 1980, pp. 1–20, 2000.



(During my 1977 talk, Robert Fairthorne[1], one the pioneers of IR was in the audience, and clearly taken with my three way balancing act drew the above cartoon.)

> ..., I think ... that the great problems of the universe and of science have the first claim on our interest. But it is as a rule of very little use to form an express intention of devoting oneself to research into this or that great problem. One is then often at a loss to know the first step to take. It is more promising in scientific work to attack whatever is immediately before one and offers an opportunity for research. If one does so really thoroughly and without prejudice or preconception, and if one has luck, then since everything is related to everything, including small things to great, one may gain access even from such unpretentious work to a study of the great problems

I still largely agree with this slogan, or motto. Curiously I would claim that considerable progress in IR has been made precisely because IR researchers took seriously the solving of "whatever is immediately before" us. The theoretical models and breakthroughs largely arose out of detailed experimentation, and new models sometimes arose out of the failure of existing models to deliver the anticipated experimental performance. For example the failure of probabilistically based term dependence models to show improvement in effectiveness over simpler independence models has lead to a number of alternative approaches.

## 2 Some Meta-thoughts on IR

It seems to me that it is possible to characterise the IR viewpoint in a number of ways. To begin with no a priori assumptions are made about structure or

[1] Sadly Fairthorne has recently died, he was fondly known to some as the "frog-prince."

process, unless given by the raw data or some external constraints. This is most obvious when it comes to classifications; these are intended to reflect the inherent structure in the data and are not imposed. When it comes to features/attributes, relevance, or aboutness a categorical view is not always taken, that is, a document is not either relevant or not-relevant, a document is not either about $X$ or not about $X$, etc. Processes in IR are usually adaptive making them user-driven and context dependent, this is particular evident in relevance feedback. The semantics of objects are defined by the data, in other words, it is the distribution both within a document and across documents that give the "meaning" of terms. IR on the whole makes no claims about Knowledge we tend to work with notions of Information and as such consider the probability of propositions to be indefinitely revisable in the light of the weight of evidence. (This is an issue in the Bayesian context when $P(X) = 1$; as an exercise the reader might like to do a Bayesian revision of a proposition $X$ whose probability is one in the light of some new evidence.). Following from this we tend to work with contingent truths rather than necessary truth, and of course this effects the kind of logics we are interested in. Finally, a trend that has emerged in the last few years is that interactions with IR systems can be based on ostensive manipulation and definition, that is, systems react to what a user does, or points to, not only to what the user says or writes.

## 3 Practice, Experiments, and Theory

Let me say a little more about these three disparate activities in IR.

*Practice.* A huge amount of operational retrieval using the web takes place, and a lot of it is woeful. A major practical challenge for IR is to influence the design of search engines so that retrieval performance goes beyond what you get by just submitting a 2.4 word query. In electronic publishing, as pursued by the large publishers for example, much multimedia data is conveniently made available but unfortunately the search capabilities are mostly inadequate. Commerce seems have discovered the knowledge economy and so data mining and knowledge discovery are the flavour of the month. Of course there is a long history in IR using statistical techniques to model significance and dependence. If one thinks about the provision of materials for distance learning whether they be text, image or graphics, then once large repositories of such information becomes available a major issue will be its retrieval. Many of these issues, especially those concerned with standards, are now addressed in the context of Digital Libraries (see Sølvberg, this volume).

*Experiments.* There is a long and honourable tradition of experimental work in IR. Cyril Cleverdon one of the pioneers, together with Jack Mills and Michael Keen produced a series of reports, initially the Cranfield I (1960) study followed by a more substantial study in 1966 [10], "Factors determining the performance of indexing systems." These projects can claim to be responsible for founding the experimental approach that is now know as the "Cranfield Paradigm," it to this day continues in the extremely successful series of experiments known as TREC (see `http://trec.nist.gov`). To understand the difficulties associated with designing test collections for IR one may read the report by Sparck Jones and Van Rijsbergen [41]. Future experimenters are encouraged to examine the approach to experimentation in IR thoroughly. A classic summary of the IR approach can be found in the collection of papers edited by Sparck Jones, "Information retrieval experiment" in 1981.

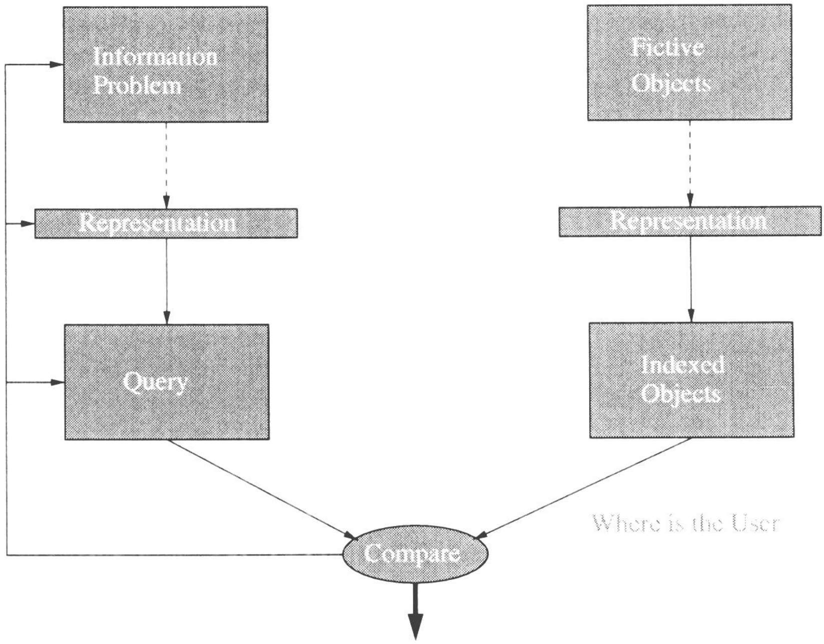
*Theory.* Much theory in IR has come about through "knob twiddling," this generally means adjusting a set of parameters for a given retrieval model and observing the effect on retrieval performance. Of course this can lead to mindless experimentation but it has also led to new variants of statistical models. Dissatisfaction with a given model, often because of poor retrieval, has led to proposals for new models embodying such disparate approaches as Bayesian Inference, Clustering, Non-classical Logic, Dempster Shafer Theory of Evidence, etc. Considerable theoretical work has also gone into the design of evaluation measures, that is, ways to mathematically, represent retrieval effectiveness, to average it, and to establish statistical significance. Two recent papers worth looking at demonstrating that the debate over effectiveness measures continues are [21] and [58]. Ever since the time of Cleverdon, Precision and Recall have been favoured. Unfortunately recall is not always readily available, think of retrieval from the Web, nor is precision always appropriate in dynamic task-oriented environments. To pursue this problem I recommend a look at [17].

## 4 IR System Architecture

Figure 2 shows a traditional view of an IR system. I believe that since the early seventies we have displayed it this way, and it has been used regularly ever since in papers on IR, just like I am doing now. It highlights one of the central concerns of IR, namely, Relevance Feedback. Of all the techniques invented to enhance retrieval effectiveness, relevance feedback is perhaps the most consistently successful. For a detailed overview see the survey article by [43]. A side-effect of this success has been to concentrate on system's development to improve and generalise relevance feedback perhaps to the detriment of actual user studies. It is common knowledge that many users do not understand relevance feedback and are not good at using it. Furthermore, when the initial input to the feedback cycle is poor, as it often is in Web searches, a sophisticated feedback mechanism is not much use: telling the IR system that all the retrieved documents are non-relevant is not helpful, the prior probability of such a retrieval is already very high.

## 5 The Twelve Dimensions of IR

I originally [50] designed the table in Table 1 as a way of comparing databases with information retrieval, however over time this comparison has become more generic. The differences between DB and IR have become less marked. I now



view this table as a way of focussing attention on a number of salient dimensions that span research in areas such as IR, databases, data-mining, knowledge discovery etc. It enables me to discuss IR research in a limited and constrained way without taking on the whole subject. In what follows I will address each one of these dimensions and describe where we are with research in that area. For a more recent discussion of this table in terms of data and document retrieval I recommend David Blair's book, Language and Representation in Information Retrieval, Elsevier, 1990.

## 6 Matching

Fundamental to any retrieval operation is the notion of matching. One can track progress in IR in terms of the increased sophistication of the matching function. Typically these functions are the consequence of a model of retrieval. For example the Boolean matching, and the Logical Uncertainty Principle (LUP) (see the paper by Crestani and Lalmas, this volume) both presuppose an elementary model and proof theory from formal logic. In the case of the LUP an assumption is made about how to measure partial entailment. There are four major IR models[2], vector-space, probabilistic, logical, and Bayesian net. Each has its corresponding matching function, for example the vector-space model predominantly uses the cosine correlation or one of its variants. Optimality criteria come

[2] A fifth model based on the ASK (Anomolous State of Knowledge) hypothesis does not really fit into this scheme, I will return to it in the section on models.