Information Access using the Guide of User Requirements

Maristella Agosti

Department of Information Engineering – University of Padua
Via Gradenigo, 6/b – 35131 Padua – Italy
agosti@dei.unipd.it

Abstract. This study presents an interpretation of the evolution of the events in the information retrieval area. Focusing mainly on the last twenty years, the study 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 can be considered the guide of the researcher, prompting him to conceive and invent solutions of real use for the user himself.

Key words: information retrieval, information access, information retrieval process, user information need, information retrieval model, user's requirements, system-oriented information retrieval

1 Introduction

The term *information retrieval* identifies the activities that a person – the *user* – has to conduct to choose, from a collection of documents, those that can be of interest to him to satisfy a specific and contingent information need. It follows that the aim of the area of information retrieval is to help and support the user in choosing, among the available documents, those that, with higher probability are more suitable to satisfy his information need. Figure 1 sketches the situation: the user has the possibility of choosing the documents of interest to him from an available collection, but he needs to have a tool that can help him in choosing the subset of documents which are of of interest for him without needing to invest a lot of time in inspecting all the documents of the collection.

Figure 1 also shows the three main actors and aspects that information retrieval needs to address:

- user,
- collection of documents,
- retrieval, which means a function or model used in retrieving and accessing information.

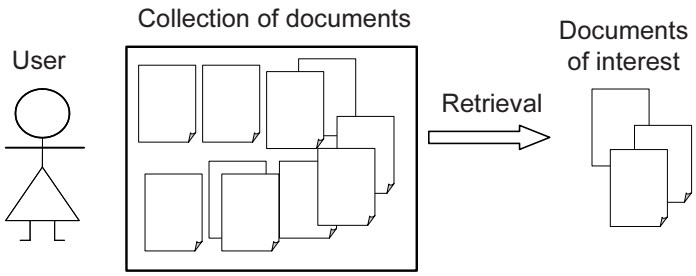

Fig. 1: Information retrieval aim

Where the user is the central actor of the situation, the collection of documents is the source from which information can be extracted in the hope it is of interest to the user, and retrieval is the function that transforms a user information need into a set of documents that are supposed to satisfy the user information need. The retrieval process begins with the user analysing his information need and trying to transform it in such a way that it becomes applicable to the collection of documents through a function that is able to produce as output the subset of documents that mostly satisfy the user information need.

When the collection of documents reaches a size that makes a manual inspection of the documents prohibitive, the construction and management of the collection together with the application of the retrieval function are managed in an automatic way through an information retrieval system. This means that an information retrieval system models and implements the retrieval process from the user input – i.e. it often takes the shape of a user *query* – to the production of an output that is constituted by a subset of collection documents most likely of interest to the user.

The approach to the modeling of the information retrieval process has dramatically changed over the years, mostly in a positive and evolutionary way, with the final aim of passing from an information retrieval approach towards an information access one where the real user is the focus of interest. This paper critically analyzes the evolution of the modeling of the information retrieval process, mainly in the last twenty years, relating the general analysis to the analysis of activities and experiences that have been conducted in the context of the Information Management Systems (IMS) Research Group of the Department of Information Engineering of the University of Padua, which started its activities twenty years ago. The approach used in this analysis is coherent with the one that has been named *system-oriented information retrieval* by Ingwersen and Järvelin in [24]. The term "system-oriented information retrieval" approach is intended to mean the approach of designing and developing models and systems of information retrieval that concentrate on the side of a *system* which has to be conceived and designed to support the

user in the access to the information. This approach is a sort of *system-side* approach and is in contrast to a *user-oriented* one.

The main focus of paper is on Sect. 2, which presents an interpretation of the evolution of the events in the information retrieval area using a system-oriented approach; attention is mostly focused on the last twenty years. Section 3 draws some final conclusions.

2 The Evolution of System-Oriented Information Retrieval

Early Days

In the early days of computer science, the common approach to information retrieval was to consider the specific type of documents constituting the collection and to manage and design the system and applications around it. The attention of the system designer was concentrated on the type of documents, mostly because the available technology limited the possibilities of representation and management of different types of documents in a single system; the only documents which could be considered were only textual and written in a single language. In addition to this, the form, that is the external appearance of the documents, was taken into account, so a system was designed to describe or to manage one specific type of textual document at a time; these included, for example, catalogue, bibliographic, and full-text documents. The consequence was that it was more efficient and effective to concentrate on one single type of document at a time and to build a system able to manage abstracts of documents [27], bibliographic documents [13], or full-text documents [64]. Sometimes the specific subject area of the documents, together with their external form, was also considered an aspect to take into consideration when designing a system; therefore, a system was designed for textual documents that were materialized in a specific form and that were all related to a specific subject area. Two significative examples are the medical area, where the Medical Literature Analysis and Retrieval System (MEDLARS) was built with the prime purpose of producing the *Index Medicus* and other recurring bibliographies [26][1], and the legal area, where the LEXIS system was designed to provide a service specifically devoted to the manipulation of legal information [64, p. 46].

1977-1986: The Last Decade of Centralized Systems

The attention of the system designer was prevalently focused on the textual collection of documents more than the user and the specific function or model to implement in the system. Still in the decade from 1977 to 1986 systems

[1] Reprinted in [34, pp. 223–246].

were only able to manage single type of document, and the systems were named in accordance with the specific type of documents that collections were designed to manage in a specialized way. But by the end of that decade the relational database management systems started to be generally available and it became possible to manage, in a single application, both structured and unstructured – textual – documents. At the same time, the application of distributed systems started to be considered, thus opening the way to the study of distributed systems also in the area of the management of unstructured information [6]. The possibility of introducing special-purpose hardware was also under study worldwide; in fact, a sector of research and development was flourishing with the aim of producing special-purpose hardware devices to be introduced in conventional computer systems to improve system capabilities for the management of numerical and non-numerical applications; in [1] the introduction of special-purpose hardware in the range of information management systems is examined and discussed.

As well, it was in those years that a new generation of library automation systems started to be designed with the purpose of enabling the retrieval of a combination of different types of data: structured catalogue data together with unstructured data representing the contents of the catalogued documents. This new type of library automation systems also supports the interactive retrieval of information through textual data – this interaction is made possible through the component known as *Online Public Access Catalogue (OPAC)* [23].

1987-1996: Towards a User-Oriented Decentralized Environment

By the end of the previous decade, probably in part due to the rapid introduction and evolution of personal computer systems and telecommunications, the attention of information management system researchers started to move from the collection of documents towards the user and the retrieval model, with the focus on the invention of models better able to support user-system interaction. So, at the beginning of that decade, researchers took new directions, trying to support the users with new types of retrieval models or with a combination of different models. Some significant examples of the attempt to open new directions towards new types of retrieval models are the work of van Rijsbergen, reported in [36] and in [35], where a model based on non-classical logic was proposed, the work of Croft [17] which proposed new approaches to intelligent information retrieval, and a collaborative work of many researchers [15] where distributed expert-systems were dealt with. A new type of system able to carry out an effective combination of multiple searches based on different retrieval models is proposed by Croft and Thompson [19], and the method of combining different information retrieval approaches has proved to be really effective over the years and through different ways of implementing solutions as reported in [18].

It is in this scientific context that the Information Management Systems Research Group was formed at the University of Padua in 1987. The group was joined by scientists formerly working with the School of Engineering, the School of Statistics and the School of Psychology. The background of the members of the group was in information retrieval, database management, and library automation systems – these last can be considered the "ancestors" of present-day *digital library systems*. The background topics of interest can be appreciated looking at the left-side of Fig. 2, which represents the range of the research interests and the topics dealt with by the group starting from 1987 up to now.

Naturally, at the beginning of its activities the group was tempted to continue to address the same research areas that were previously dealt with by its components. However, also partly due to the flourishing scientific context of those years, the group made use of the possibility of synergically applying the competence of the members, and immediately started to propose new ideas and, in particular, a new approach to information retrieval based on the linking of documents and of descriptive objects based on the hypertext paradigm [2]. The study of hypertext methods was at the basis of a new sort of *network-based* and *associative information retrieval* entitled EXPLICIT. Associative information retrieval methods are those retrieval methods which have been

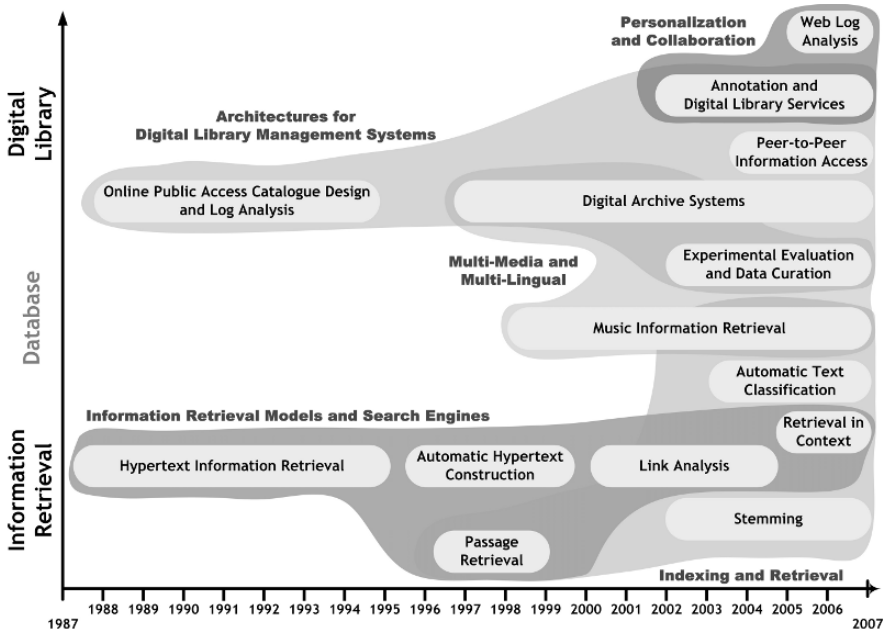

Fig. 2: The range of the research interests of the IMS Research Group in its twenty years of existence

proposed and experimented since the early days of information retrieval [20, 30]. They seek to expand query formulation by adding to an initial query some new terms related to the terms of the initial query, and similarly expand the retrieved document set using terms related to the already used terms; [32] makes use of associative information retrieval methods and shows that the difficulty encountered in applying associative retrieval methods resides in the identification of related terms and documents which would improve retrieval operations, and this is one of the research topic that still remains to be fully solved.

The EXPLICIT model was based on a two-level architecture initially proposed in [10] and refined in [5]. The two-level architecture holds the two main parts of the informative resource managed by an information retrieval tool: on the one hand the collection of content objects (e.g. a single collection of documents, different collections of different types of digital content objects), and on the other the term structure, which is a schema of concepts that can be composed of either one single indexing structure or some cooperative content representation structures such as those depicted in Fig. 3, in a sort of a "semantic network". The system manages this network to retrieve information of use for the final user, but also to present the representation of the contents of the collections to the user, who can use them for browsing and becoming acquainted with the information richness of the managed collections.

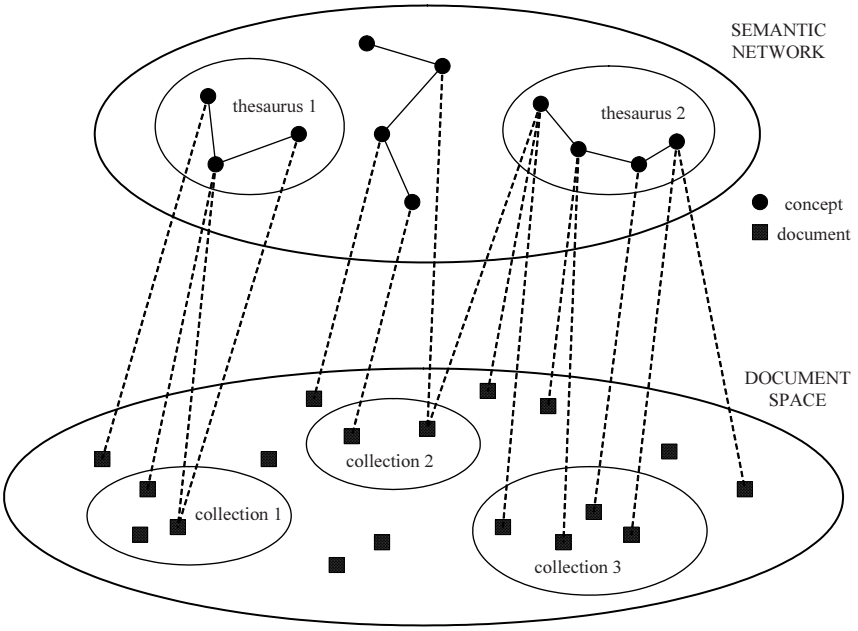

Fig. 3: The EXPLICIT architecture and model

The model was presented in [4] where the prototype developed to validate the model in a personal computing environment was also introduced; in the prototype the schema of concepts was made available to the end user as a frame of reference in the query formulation process. As shown in Fig. 3, the EXPLICIT model supports the concurrent use of different schemas of concepts to satisfy the information needs of different categories of users; the schemas of objects can be a term structure derived from the automatic indexing process, together with a classification system and/or a thesaurus.

In those years other hypertext network-based models were introduced, opening an area of information retrieval that was named *hypertext information retrieval*. Most of hypertext information retrieval models represent, as one layer, the document network and, as another, a concept network. Links relate concept nodes and document nodes together. The users, who make use of systems based on those models, can browse both the concept network and the document network, and move back and forth between the two informative structures. An important characteristic of hypertext information retrieval models is the capacity of integrating different information access approaches: browsing, navigation, and retrieval. The book by Agosti and Smeaton [14] is a reference for the different models that have been proposed in the area.

Having as a basis hypertext information retrieval models, a new area started to emerge, that of the *automatic construction* or *authoring* of hypertexts to be used for information retrieval purposes. Many proposals were suggested as reported in Crestani's chapter of this book, and the IMS research group has contributed to those results.

Together with the work on the automatic hypertext construction, the group was also continuing its activities in library automation, participating in the design and development of the first academic OPAC made available through Internet access in Italy [11].

1997-2006: The Overwhelming Amount of Digital Documents

During this decade it became clear that the most relevant aspect is the continuous growth of diverse collections of documents in digital form, so major efforts are tackling different aspects related to the growing of digital collections. The problem is dealt with through proposals of new retrieval models able to manage the Web collection of documents [12], large-scale evaluation efforts [21], and multimedia content retrieval.

In relation to models able to manage Web documents, it is worth noting that the different hypertext information retrieval models that have been proposed in the previous decade are precursors of Web retrieval techniques employing link information [24], such as the proposal by Marchiori, which suggested exploiting the hyperlink structure of the Web to try to enhance the performance of Web search engines in [28] and the PageRank algorithm proposed by Brin and Page [16]. Based on these new techniques different search engines were designed and made available to the general public.

A different way of approaching the problem is to consider that the number and size of Web collections and databases storing unstructured textual data that are made available to the final user present high degrees of heterogeneity at the level of content and language. Such a situation can be dealt with by extracting and managing specific and relevant pieces of data from larger data objects. The study of this possibility is named *passage retrieval* [31], which is the identification and extraction of fragments from large or short heterogeneous full-text documents; passages are chunks of contiguous text belonging to a larger text, usually sentences or sequences of sentences identified by punctuation marks, such as full stops or semi-colons. A probabilistic method that models the document language to identify relevant passages has been proposed by Melucci and is presented in [29].

In this period large-scale evaluation campaigns for information retrieval were launched. *Text REtrieval Conference (TREC)* has been the precursor for large-scale efforts [22, 37], TREC has been followed up by Cross-Language Evaluation Forum mainly for European multilingual-efforts [7] and by *NII-NACSIS Test Collection for IR Systems (NTCIR)* mainly for Asian languages [25]. All these international evaluation campaigns have shown their validity over the decade, because the retrieval effectiveness of the retrieval systems has steadily grown over the years also in part due to the stimulus provided by those efforts.

Multimedia retrieval is the other relevant area that researchers started to face in a systematic way and for different media during the decade. The complexity of the management of collections of multimedia digital documents can be faced in particular for information retrieval purposes, but also from a general architectural point of view, that is, the area of digital libraries and digital library systems. The many European research groups participating in the DELOS Network of Excellence on Digital Libraries, in the Information Society Technologies (IST) Program of the European Commission[2], have greatly contributed to the development of the area, obtaining significant results. A European forum that has greatly contributed to the growth in attention and has given the research community a systematic appointment over the year has been *European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, which was inspired starting from 1996 by Costantino Thanos, who is also the coordinator of the DELOS Network of Excellence. The IMS research group has actively participated in the research regarding multimedia and digital libraries; most of the chapters of this book refer on pertinent research results.

From 2007 on

The new decade that starts this year is opening up new exciting challenges, in particular for multimedia search engines which are not yet available and for

[2] DELOS Web site at the URL: `http://www.delos.info/`

which much remains to be done to make them available. The IMS research group is participating in the *Search in Audio-visual content using Peer-to-peer Information Retrieval* (SAPIR) project[3] an EU IST FP6 research project, which will contribute to making multimedia search engines available.

Another area where the IMS group has actively operated, but which still has many topics which require the attention of researchers, is that of digital library systems both on the side of architectures and on the contents, so we can look forward to many exciting topics to be dealt with the next decade.

3 Conclusions

At this point the reader might ask himself why the title of this study is *Information Access using the Guide of User Requirements*. In fact, the user has been mentioned many times, when addressing the problem of the retrieval of information, but *user requirements* have never been explicitly dealt with.

The title derives from the consideration that the research activities of the IMS group have been mostly guided by the users that have cooperated over the years with the members of the group. The research proposals and solutions that have been conducted by the group have been derived from a common initial phase of requirements analysis and design that has always been conducted together with the end users of the specific system or tool that was under investigation. Just some of the examples are: the design of the EXPLICIT model together with experts of legal documents, the design of the DUO OPAC together with the librarians and the users of the University of Padua, the feasibility study of the *Digital Archive of Venetian Music (ADMV)* project aimed to build an effective digital library which users can fully access to retrieve bibliographic records, digitalized scores, and high quality sound of Venetian music which is presented in [3] and conducted together with experts of Italian national institutions, and more recently the *Imaginum Patavinae Scientiae Archivum (IPSA)* system together with national experts of illuminated manuscripts [8, 17].

This means that the guide of the IMS research group has always been the user, who must be the focus of the attention of the researchers who want to propose new information access solutions.

References

1. Agosti, M.: Special Purpose Hardware and Effective Information Processing. Information Technology: Research and Development **3**(1), 3–14 (1984)
2. Agosti, M.: Is Hypertext a New Model of Information Retrieval? In: Proceedings of the 12th International Online Information Meeting, Vol. I, pp. 57–62. Learned Information, Oxford (1988)

[3] `http://www.sapir.eu/`

3. Agosti, M., Bombi, F., Melucci, M., Mian, G.A.: Towards a digital library for the Venetian music of the eighteenth century. In: M. Deegan, J. Anderson, H. Short (eds.) DRH98: Selected Papers from Digital Resources for the Humanities 1998, pp. 1–15. Office for Humanities Communication, Publication 12, London (2000)
4. Agosti, M., Colotti, R., Gradenigo, G.: A two-level hypertext retrieval model for legal data. In: E.A. Fox (ed.) Proc. 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1991), pp. 316–325. ACM Press, New York, USA, Chicago, USA (1991)
5. Agosti, M., Crestani, F., Gradenigo, G., Mattiello, P.: An approach for the conceptual modelling of IR auxiliary data. In: Ninth Annual IEEE International Conference on Computers and Communications, pp. 500–505. Scottsdale, Arizona (1990)
6. Agosti, M., Dalla Libera, F., Johnson, R.G.: The Use of Distributed Data Bases in Libraries. In: R.D. Parslow (ed.) BCS '81: Information Technology for the Eighties, pp. 559–570 (1981)
7. Agosti, M., Di Nunzio, G.M., Ferro, N., Peters, C.: CLEF: Ongoing Activities and Plans for the Future. In: N. Kando (ed.) Proc. 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, pp. 493–504. National Institute of Informatics, Tokyo, Japan (2007)
8. Agosti, M., Ferro, N., Orio, N.: Annotating Illuminated Manuscripts: an Effective Tool for Research and Education. In: M. Marlino, T. Sumner, F. Shipman (eds.) Proc. 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2005), pp. 121–130. ACM Press, New York, USA (2005)
9. Agosti, M., Ferro, N., Orio, N.: Graph-based Automatic Suggestion of Relationships among Images of Illuminated Manuscripts. In: H.M. Haddad, K.M. Liebrock, R. Chbeir, M.J. Palakal, S. Ossowski, K. Yetongnoon, R.L. Wainwright, C. Nicolle (eds.) Proc. 21st ACM Symposium on Applied Computing (SAC 2006), pp. 1063–1067. ACM Press, New York, USA (2006)
10. Agosti, M., Gradenigo, G., Mattiello, P.: The hypertext as an effective information retrieval tool for the final user. In: A.A. Martino (ed.) Pre-proceedings of the 3rd International Conference on Logics, Informatics and Law, Vol. I, pp. 1–19 (1989)
11. Agosti, M., Masotti, M.: Design of an OPAC Database to Permit Different Subject Searching Accesses in a Multi-disciplines Universities Library Catalogue Database. In: N.J. Belkin, P. Ingwersen, A. Mark Pejtersen, E.A. Fox (eds.) Proc. 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1992), pp. 245–255. ACM Press, New York, USA (1992)
12. Agosti, M., Melucci, M.: Information Retrieval on the Web. In: M. Agosti, F. Crestani, G. Pasi (eds.) Lectures on Information Retrieval: Third European Summer-School (ESSIR 2000), pp. 242–285. Springer, Berlin/Heidelberg (2001)
13. Agosti, M., Ronchi, M.E.: DOC-5 - The Bibliographic Information Retrieval System in CINECA Library Automation Project. In: Proceedings of ECODU-29 Conference, pp. 20–31. Berlin, Germany (1980)
14. Agosti, M., Smeaton, A.F. (eds.): Information Retrieval and Hypertext. Kluwer Academic Publishers, Boston, USA (1996)
15. Belkin, N.J., Borgman, C.L., Brooks, H.M., Bylander, T., Croft, W.B., Daniels, P.J., Deerwester, S.C., Fox, E.A., Ingwersen, P., Rada, R.: Distributed

Expert-Based Information Systems: An Interdisciplinary Approach. Information Processing & Management **23**(5), 395–409 (1987)
16. Brin, S., Page, L.: The anatomy of a large scale hypertextual Web search engine. In: Proceedings of the World Wide Web Conference (1998)
17. Croft, W.B.: Approaches to Intelligent Information Retrieval. Information Processing & Management **23**(4), 249–254 (1987)
18. Croft, W.B.: Combining Approaches to Information Retrieval. In: W.B. Croft (ed.) Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval, pp. 1–36. Kluwer Academic Publishers, Norwell (MA), USA (2000)
19. Croft, W.B., Thompson, R.H.: I3R: a New Approach to the Design of Document Retrieval Systems. Journal of the American Society for Information Science **38**(6), 389–404 (1987)
20. Doyle, L.B.: Information Retrieval and Processing. Melville, Los Angeles (1975)
21. Harman, D.: Evaluation Issues in Information Retrieval, introductory paper to the Special Issue on: Evaluation Issues in Information Retrieval. Information Processing & Management **28**(4), 439–440 (1992)
22. Harman, D.: Overview of the First Text REtrieval Conference (TREC-1). In: D.K. Harman (ed.) The First Text REtrieval Conference (TREC-1). National Institute of Standards and Technology (NIST), Special Pubblication 500-207, Whasington, USA. `http://trec.nist.gov/pubs/trec1/papers/01.txt` [last visited 2007, March 23] (1992)
23. Hildreth, C.R.: Online public access catalog. In: M.E. Williams (ed.) Annual Review of Information Science and Technology (ARIST), Vol. 20, pp. 233–285 (1985)
24. Ingwersen, P., Järvelin, K.: The Turn. Springer, The Netherlands (2005)
25. Kishida, K., Chen, K.H., Lee, S., Kuriyama, D., Kando, N., Chen, H.H., Myaeng, S.H.: Overview of CLIR Task at the Fifth NTCIR Workshop. In: N. Kando, M. Takaku (eds.) Proc. of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access. `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/CLIR/NTCIR5-OV-CLIR-KishidaK.pdf` [last visited 2007, March 23] (2005)
26. Lancaster, F.W.: MEDLARS: Report on the evaluation of its operating efficiency. American Documentation **20**, 119–142 (1969)
27. Luhn, H.P.: The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development **2**(2), 159–165 (1958)
28. Marchiori, M.: The Quest for Correct Information on the Web: Hyper Search Engines. Computer Networks and ISDN Systems **29**(8–1), 1225–1235 (1997)
29. Melucci, M.: Passage Retrieval: A Probabilistic Technique. Information Processing & Management **34**(1), 43–68 (1998)
30. Salton, G.: Associative document retrieval techniques using bibliographic information. Journal of the ACM **10**, 440–457 (1963)
31. Salton, G., Allan, J., Buckley, C.: Approaches to Passage Retrieval in Full Text Information Systems. In: R. Korfhage, E. Rasmussen, P. Willett (eds.) Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993), pp. 49–58. ACM Press, New York, USA (1993)

32. Salton, G., Buckley, C.: On the use of spreading activation methods in automatic information retrieval. In: Y. Chiaramella (ed.) Proc. 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1988), pp. 147–159. ACM Press, New York, USA (1988)
33. Salton, G., McGill, M.J.: Introduction to modern information retrieval. McGraw-Hill, New York, NY, USA (1983)
34. Sparck Jones, K., Willett, P. (eds.): Readings in Information Retrieval. Morgan Kaufmann, San Francisco, CA, USA (1997)
35. van Rijsbergen, C.J.: A New Theoretical Framework for Information Retrieval. In: Proc. 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1986), pp. 194–200. ACM Press, New York, USA (1986)
36. van Rijsbergen, C.J.: A Non-Classical Logic for Information Retrieval. The Computer Journal **29**(6), 481–485 (1986)
37. Voorhees, E., Harman, D. (eds.): TREC: Experiment and Evaluation in Information Retrieval. The MIT Press, Cambridge, MA, USA (2005)