mensions into studies of information production process improvements, such as the study of the relationship between IQ process measures and IQ outcome measures in [15]. Similarly, IP life cycles should be integrated into information production process models. IQ governance studies are needed that include regular outcome measures to be used to direct efforts at process analysis and improvement. Governance studies are also needed that will develop methods for costing and justifying investments in IQ improvements. The purpose of these research efforts is to develop theories, methods, tools, and recommendations that will help organizations actively manage their information and thus ensure the delivery of high-quality information to information consumers.

Cross-References

- ▸ Information Quality and Decision-Making
- ▸ Information Quality Assessment
- ▸ Information Quality Policy and Strategy

Recommended Reading

1. Proceedings of the International Conference on Information Quality (ICIQ). (1996 and yearly since then), available at: http://mitiq.mit.edu and at http://mitiq.mit.edu/ICIQ.
2. Madnick SE, Wang RY, Lee YW, Zhu H. Overview and framework for data and information quality research. ACM J Data Inf Qual. 2009;1(1, Article 2):1–22.
3. Wang RY, Lee YW, Pipino LL, Strong DM. Manage your information as a product. Sloan Manag Rev. 1998;39(4):95–105.
4. Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. J Manag Inf Syst. 1996;12(4):5–34.
5. Strong DM, Lee YW, Wang RY. Data quality in context. Commun ACM. 1997;40(5):103–10.
6. Lee YW, Strong DM, Kahn BK, Wang RY. AIMQ: a methodology for information quality assessment. Inf Manag. 2002;40(2):133–46.
7. Pipino LL, Lee YW, Wang RY. Data quality assessment. Commun ACM. 2002;45(4):38–46.
8. Ballou DP, Wang RY, Pazer H, Tayi GK. Modeling information manufacturing systems to determine information product quality. Manag Sci. 1998;44(4):462–84.
9. Kahn BK, Strong DM, Wang RY. Information quality benchmarks: product and service performance. Commun ACM. 2002;45(4):184–92.
10. Wang RY. A product perspective on total data quality management. Commun ACM. 1998;41(2):58–65.
11. Shankaranarayan G, Ziad M, Wang RY. Managing data quality in dynamic decision environment: an information product approach. J Database Manag. 2003;14(4):14–32.
12. Cao L, Zhu H. Normal accidents: data quality problems in ERP-enabled manufacturing. ACM J Data Inf Qual. 2013;4(3, Article 11):1–26.
13. Lee YW, Pipino L, Funk J, Wang RY. Journey to data quality. Cambridge, MA: MIT Press; 2006.
14. Davidson B, Lee YW, Wang RY. Developing data production maps: meeting patient discharge data submission requirements. Int J Healthc Technol Manag. 2004;6(2):223–40.
15. Lee YW, Strong DM. Knowing-why about data processes and data quality. J Manag Inf Syst. 2004;20(3):13–39.

Information Retrieval

Giambattista Amati
Fondazione Ugo Bordoni, Rome, Italy

Synonyms

Document retrieval; Text retrieval

Definition

Information retrieval (IR) deals with the construction of automatic systems that allow users to inquire about textual data of any kind through natural language queries. The retrieved information from IR systems may vary from a ranked list of relevant textual items of any kind, such as full documents or their excerpts, or can be distilled into more elaborated forms, such as document summaries or answers to questions. Information retrieval is an empirical science that studies representation, storage, and access to information and covers a large number of interdisciplinary topics of theoretical

computer science including information theory, machine learning, coding theory, probability theory, programming theory, computational semantics, natural language processing, logics, and algebra. From a practical perspective, research on IR includes data representation; storage and retrieval, such as indexing, data encoding, and text compression; document and term classification and clustering; systems architecture; distributed systems; and document-query matching functions (IR models) and user-oriented studies and aspects of behavioral science, such as data visualization, browsing, user interfaces, system evaluation, user relevance feedback, and automatic query expansion. With the advent and diffusion of the Web and the dramatic increase of public available information sources, IR research also focuses on efficiency in terms of query response time and storage space of indexes in a distributed setting, as well as on personalized search where results can be filtered and adapted to user's area of interest taking into account, for example, time, geographical knowledge, and user's historical data.

Historical Background

Information retrieval has its origins in the 1950s to support the librarians' activities of indexing and accessing textual collections. Hans Peter Luhn is one of the pioneers of IR [5] with his studies on automatic indexing that concerns the assignment of significant keywords to documents. In the early stage of IR, due to the limitations of computer capabilities, document retrieval was limited to satisfy the Boolean exact match between the query terms and document surrogates (title, subject headings, or abstract). Luhn thought that the automation of indexing and abstracting was less prone to errors than human indexers. Before Luhn's original ideas, Shannon, the founder of the mathematical theory of communication, was the first to consider text generation as a "sequence of words" and processed as a discrete information source by a stochastic process (a discrete Markov process) [10]. Following Shannon's idea, Mandelbrot derived theoretically the empirical model by Eustop on word distribution and further studied by Zipf [14]. Mandelbrot showed that text generation is finding the least costly method of coding as obtained in the classical problem in communication theory [6]. The law of Estoup-Zipf-Mandelbrot establishes that the logarithm of rank of words by decreasing frequency is in linear relation with this frequency. An important milestone in the history of information retrieval is the introduction of evaluation measures for IR by Cleverdon. In 1953, Cleverdon led a project of Librarian of Cranfield College of Aeronautics to assess retrieval with a "document source question" method that consists in collecting questions and their relevance judgments from individuals; the Cranfield test collection is still publicly available for experimentation in IR [3]. In 1960, Maron and Kuhns used the term "probabilistic indexing" in the theory of IR to model the concept of relevance explicated in terms of the theory of probability [7]. Nevertheless, it was the vector space model, a model not based on probability theory, that influenced most of the research for some decades. The SMART project, dedicated to the realization of one of the first IR systems, the SMART system, was initiated at Harvard University in 1961 by Gerard Salton, but actually most of the research was conducted at the Cornell University [9]. The SMART system used the cosine of the index vectors derived from documents and search requests to obtain for each document a coefficient of similarity with each search request. After Maron and Kuhns, a probabilistic model of relevance was introduced by Stephen Robertson and Karen Spärck Jones [8] that has led to the development of the OKAPI system. The probabilistic model is based on Spärck Jones' observation that the significance of a term in a document is due to its rareness in the collection and it can be measured by the logarithm of the inverse of the relative frequency of the term in the collection (document frequency) [11]. In the early 1970s, the "cluster hypothesis" was stated by Jardine and van Rijsbergen [4]. According to the cluster hypothesis, both efficiency and

effectiveness of search could benefit from clustering similar documents because such clusters are easily retrieved by the same search requests, while Salton thought that clustering would have reduced effectiveness in favor of a better response time.

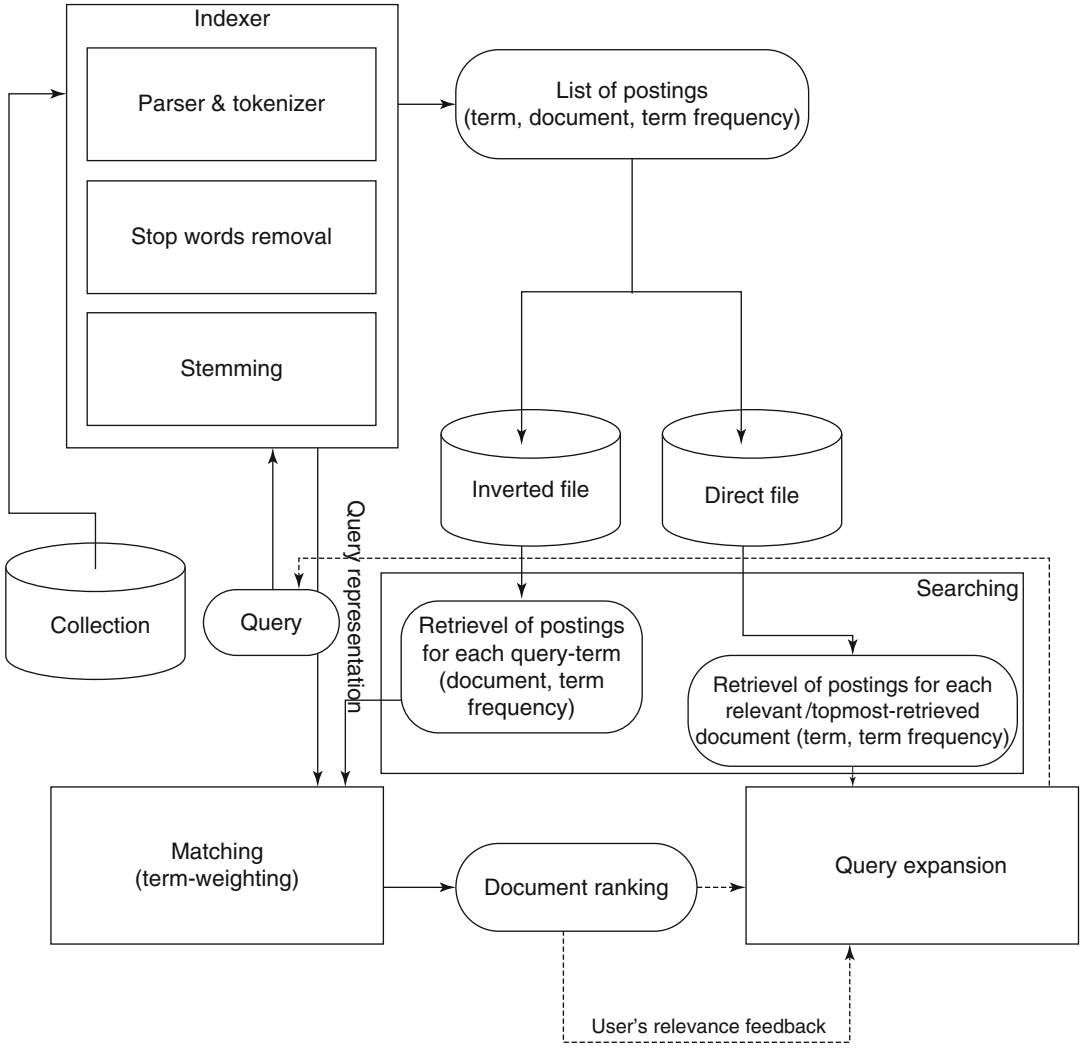
The first main international conference on IR, ACM SIGIR, was held in 1978 [1]. The first commercial search engines appeared in the 1980s, while the 1990s saw the birth of the first Web search engines. Since the beginning of the TREC (Text REtrieval Conference) series in 1992, a conference organized by the US government's National Institute of Standards and Technology and dedicated to large-scale evaluation of text retrieval methodologies, very large test collections of full-text documents and standards for retrieval evaluation are available. Evaluation is an important issue in IR; therefore TREC has a significant impact in research because it provides an objective evaluation of fresh techniques and approaches and promotes new specialized retrieval tasks, as well as the transfer of emerging new ideas into commercial systems and Web search engines.

Foundations

Information retrieval systems have four main components: query representation and document indexing, term and document retrieval, query-document matching, and relevance feedback processing (see Fig. 1). Typically, a user submits a query made up of a few words and phrases or several sentences. The internal representation of the original query and the text extracted from a document are both processed by the same tokenizer: a *parser* extracts the words taking into account specific properties of the language, then the most frequent words and other words that have a functional role in the text (*stop words*) are removed, and finally the constituents of the lexicon of the system (*tokens*) are created by removing linguistic suffixes or prefixes (*stemmer*); their multiplicity of occurrence in the text is recorded. The text of each document is indexed into two files: the *inverted* and the *direct* files. Both files contain the set of *pointers* of the collection, also known as *postings*, that is, the matrix containing all possible term-document pairs of the collection. With the inverted file, one can access the set of pointers (e.g., term, document, frequency) and thus the set of documents containing a given term together with some extra information about these pointers, that is, the frequency of that term in these documents and possibly the positions occupied by that given term in each document. With the direct file, one can instead access the set pointers (e.g., document, term, frequency), that is, the set of terms contained in a given document together with the frequency of each term in that document. Positions of terms within documents are used to restrict search with the use of proximity operators. For example, one can submit the query "information retrieval" wishing to find and rank by relevance all and only all documents containing the word "information" followed by the word "retrieval." While *information AND retrieval* is the query to find and rank all and only all relevant documents containing both the words "information" and "retrieval" irrespective of the positions occupied by these words in the text. Frequencies are instead essential to obtain a ranked list of documents.

The direct file is used to perform other post-processing activities, such as *automatic query expansion*, or to *cluster* the retrieved documents into homogeneous or similar classes of documents or to present results in different output formats (e.g., into an XML form).

Both inverted and direct files are stored in a compressed form, and it is possible to achieve a very good rate of compression with respect to the size of the original textual data. For example, the inverted file of the TREC corpus WT10g containing 1,692,096 documents of 10 GB of text and 280,571,311 pointers can be compressed in about 385.5 MB, each pointer (term document, term frequency) requiring an average of 11.5 bits of information only. In general, the most frequent words are declared in a special list of words, the *stoplist*, and are not indexed because they require the storage of too many pointers (the size of the collection in the worst case). A com-



Information Retrieval, Fig. 1 A conceptual architecture for an information retrieval system

prehensive study of compression algorithms for text retrieval can be found in the book *Managing Gigabytes* [13].

The kernel of an information retrieval system is the query-document matching model, producing the document scores for the given query, also known as the *retrieval status value* (RSV) of the documents for that query. The matching model is the theoretical component responsible for the effectiveness and the quality of the retrieval.

Similarly to any other empirical science, information retrieval makes inferences by analysis of the empirical data consisting of three phases: *model specification*, *parameter estimation*, and *model evaluation*. However, unlike other empirical sciences, IR has two types of data: the postings, that is, the elements of the term-document matrix, and the user relevance data set that are in general provided by a set of document-query relevance assessment pairs. The relevance data set can be used to accomplish two different tasks: query reformulation for improving document retrieval and the evaluation of system performance.

Due to the existence of relevance feedback data, IR has two different kinds of models: the *query-language model* and the *term-document model*. The query models aim at providing the weights of the query terms irrespective of the observed document. The term-document model instead compares and weights the frequency of the term within a document with respect to its frequency in the collection.

For example, the query "What is a prime factor?", which is transformed into the query "prime factor" after stop-word removal and stemming, can be simply represented as the vector (prime = 1, factor = 1). After a first pass retrieval in the absence of user's feedback, one may deem the first retrieved documents as relevant and assume that the terms contained in this small portion of retrieved documents as a sample of the term population relative to the topic "prime factor." Then, the query-language model will assign new weights to the original query terms and add new terms to the query. For example, retrieving just the first three documents, the new query might be (prime = 1.3581, factor = 1.2327, integ = 0.2154, number = 0.1778, primal = 0.0941, ...). The term-document weights assigned by the term-document model will be resized according to query-term weights, that is, as the inner product of these two weight vectors.

A theory on evaluation of IR systems is mainly developed in van Rijsbergen's book [12]. The effectiveness of an IR system is evaluated by two standard measures: *recall* and *precision* that are defined as follows:

$$Recall = \frac{\mu(Rel \cap Ret)}{\mu(Rel)}$$

$$Precision = \frac{\mu(Rel \cap Ret)}{\mu(Ret)}$$

where $Ret = \{d|d \text{ is retrieved}\}$ and $Rel = \{d|d \text{ is relevant}\}$, and $\mu = |\cdot|$ is the counting measure. Then, $|Rel \cap Ret|$ is the number of relevant and retrieved documents, $|Ret|$ is the number of retrieved documents, and $|Rel|$ is the number of relevant documents.

Key Applications

Main applications of information retrieval concern the construction of search engines either for specific domains, like biomedicine and genomics, law, Web, and blogs, or adapted to particular types of document structure (e.g., XML or hypertext documents) search engines or dedicated to multimedia and digital libraries.

Future Directions

Notwithstanding the increase of computer processing power, most of IR applications deal with very large collections that cannot be in general processed by only one server. Both efficient implementation and distributed versions of IR systems are required. The design of efficient partitions of very large indexes that need to be distributed over a cluster of machines is an important research topic. Also, large community of users may wish to share their information and knowledge, but not their local indexes, and therefore merging local search results to obtain one effective global document ranking is a challenging research issue (*data fusion*). Most of the new technology for information retrieval is becoming more and more an asset of the most popular Web search engines, and therefore it is mainly the market that influences new academic research directions. Although social network analysis has achieved a quite mature technology based on variations of Markov chain models with stationary probability distributions, a new social phenomenon is now emerging in the Web, the *social tagging*; systems also provide collaborative tools to Internet users to store and search structured comments of web pages. Query search on movies, music, or books will be in the very next future conditioned, for example, by opinions or by recommendation. More generally, search will be more and more personalized and tailored to

one's own profile or to all other profiles similar to one's own. The major techniques to discover statistical relationships for hypertext documents and hyperlinks can be found in Chakrabarti's book [2].

Data Sets

Most of data sets and test collections can be found at the TREC (NIST) Web site http://trec.nist.gov/data.html. There are several types of test collections concerning different range of IR applications. To cite few of them, there are the ad hoc collections that are dedicated to the standard document retrieval based on a primitive notion of user's relevance, *Web* test collections that are dedicated to Web retrieval, and *Blog Track* and *Microblog Track* test collections that contain large collection of posts from the blogosphere and from Twitter's community, respectively.

URL to Code

There are several open source IR systems on the Web. In the following there is a list of the most popular and advanced academic IR research systems.

1. Indri and Lemur search engines developed by Carnegie Mellon University and the University of Massachusetts, Amherst, United States: http://www.lemurproject.org/
2. Terrier developed by University of Glasgow, United Kingdom: http://ir.dcs.gla.ac.uk/terrier/
3. Wumpus developed by University of Waterloo, Canada: http://www.wumpus-search.org/
4. Zettair by RMIT University of Melbourne in Australia: http://www.seg.rmit.edu.au/zettair/

Other IR systems are the open source Apache Lucene (http://lucene.apache.org/) and the SMART system (ftp://ftp.cs.cornell.edu/pub/smart/).

Cross-References

- Biomedical Scientific Textual Data Types and Processing
- Digital Libraries
- Advanced Information Retrieval Measures
- Information Retrieval Models
- Information Retrieval Operations
- Query Expansion Models
- Text Indexing Techniques

Recommended Reading

1. Annual international SIGIR conference, Proceedings of the ACM Special Interest Group on Information Retrieval Conference. http://www.sigir.org/
2. Chakrabarti S. Mining the Web: discovering knowledge from hypertext data. Amsterdam: Morgan-Kauffman; 2002.
3. Cleverdon C. The cranfield test on index language devices. ASLIB Proc. 1967;19(6):173–92.
4. Jardine N, van Rijsbergen CJ. The use of hierarchic clustering in information retrieval. Inf Storage Retr. 1971;7(5):217–40.
5. Luhn H. A statistical approach to mechanized encoding and searching of literary information. IBM J Res Dev. 1957;1(4):309–17.
6. Mandelbrot B. An informational theory of the statistical structure of language. In: Jackson W, editor. Communication theory, the second London symposium. Butterworth: London; 1953. p. 486–504.
7. Maron M.E. and Kuhns J.L. On relevance, probabilistic indexing and information retrieval. J ACM. 1960;7(3):216–44. http://doi.acm.org/10.1145/321033.321035.
8. Robertson SE, Sparck-Jones K. Relevance weighting of search terms. J Am Soc Inf Sci. 1976;27(3):129–46.
9. Salton G, Lesk ME. The SMART automatic document retrieval systems – an illustration. Commun ACM. 1965;8(6):391–8.
10. Shannon C. A mathematical theory of communication. Bell Syst Tech J. 1948;27(3):379–423 and 623–656.
11. Sparck JK. A statistical interpretation of term specificity and its application in retrieval. J Doc. 1972;28(1):11–21.
12. Van Rijsbergen C. Information retrieval. 2nd ed. London: Butterworths; 1979.
13. Witten IH, Moffat A, Bell TC. Managing gigabytes. 2nd ed. San Francisco: Morgan Kaufmann; 1999.
14. Zipf G. Human behavior and the principle of least effort. Reading: Addison-Wesley; 1949.

Information Retrieval Models

Giambattista Amati
Fondazione Ugo Bordoni, Rome, Italy

Synonyms

Ad hoc retrieval models; Document term weighting; Term-document matching function

Definition

A model of information retrieval (IR) selects and ranks the relevant documents with respect to a user's query. The texts of the documents and the queries are represented in the same way, so that document selection and ranking can be formalized by a matching function that returns a retrieval status value (RSV) for each document in the collection. Most of the IR systems represent document contents by a set of descriptors, called terms, belonging to a vocabulary $\mathbf{V}$.

An IR model defines the query-document matching function according to four main approaches:

- The estimation of the probability of user's relevance *rel* for each document $\mathbf{d}$ and query $\mathbf{q}$ with respect to a set $R\mathbf{q}$ of training documents

$$\text{Prob}\left(rel|\mathbf{d},\mathbf{q},R_{\mathbf{q}}\right)$$

- The computation of a similarity function between queries and documents in a vector space

$$SIM\left(\mathbf{d},\mathbf{q}\right)$$

- The estimation of the probability of generating the document $\mathbf{d}$ from a query $\mathbf{q}$

$$\mathbf{p}\left(\mathbf{d}|\,\mathbf{q}\right)$$

- The amount of information carried by the query terms in the document, that is, the number of bits that are necessary to code the number X_i of occurrences of the query terms $\mathbf{t}_i \in \mathbf{q}$ in the document:

$$-\log_2 \text{Prob}\left(\mathbf{d}|X_1,\ldots,X_{\mathbf{q}}\right)$$

Historical Background

The history of information retrieval has evolved in parallel to that of the development of IR models. In general, an IR system is mainly identified with its retrieval function that is employed to rank documents, because it is the retrieval *effectiveness* that matters in IR systems. In the early days, automatic indexing was the main focus of the IR research with the aim to help the manual classification performed by librarians. Early works of IR concerned the construction of effective methods for keywords selection to represent succinctly the documents, and the matching between documents and queries was more simply performed by a Boolean search of the query terms in the index of such surrogates of the documents. A full exploitation of more complex models, such as the probabilistic model, was only achieved in the 1990s with the birth of the BM25 ranking formula [1]. Before the appearance of the BM25, most of the theoretical investigation was devoted to the vector space model, the first parameter-free model for ranking documents [2]. It was only in the late 1990s that other new models appeared on the scene, such as the language models [3] and the information theoretic models that include the Divergence From Randomness models [4] and models based on survival functions [5]. The Boolean model remains attractive because it requires less computational cost both in terms of size of the indices and response time, and a few models tried to extend the Boolean model with fuzzy or logical operators, yet the potentialities of such algebraic and logical models are not fully exploited.

Foundations

IR models can be classified into four types: probabilistic models, algebraic and logical mod-