Glossary

ASCII: Standard binary codes to represent occidental characters in one byte.

Ad hoc retrieval: standard retrieval task in which the user specifies his information need through a query which initiates a search (executed by the information system) for documents which are likely to be relevant to the user.

All-pairs or spatial-join query: a query that requests all the *pairs* of objects that are within the specified distance from their partner.

Amdahl's law: Using N processors, the maximal speedup S obtainable for a given problem is related to f, the fraction of the problem that must be computed sequentially. The relationship is given by: $S_N \leq \frac{1}{f+(1-f)/N} \leq \frac{1}{f}$.

Belief network: a probabilistic model of document retrieval based on interpreting documents, user queries, and index terms as nodes of a Bayesian network. This model is distinct from the inference network model.

Bit-parallelism: a speed-up technique based on exploiting the fact that the processor performs some operations in parallel over all the bits of the computer word.

Block addressing: a technique used to reduce the size of the lists of occurrences by pointing to text blocks instead of exact positions.

Boolean model: a classic model of document retrieval based on classic set theory.

Browsing: interactive task in which the user is more interested in exploring the document collection than in retrieving documents which satisfy a specific information need.

CACM collection: a reference collection composed of all the 3204 articles published in the Communications of the ACM from 1958 to 1979.

CISI collection: a reference collection composed of 1460 documents selected from a previous collections assembled at ISI.

Clustering: the grouping of documents which satisfy a set of common properties. The aim is to assemble together documents which are related among themselves. Clustering can be used, for instance, to expand a user query with new and related index terms.

Coding: the substitution of text symbols by numeric codes with the aim of encrypting or compressing text.

Collection: a group of items, often documents. In (digital) libraries this designates all the works included, usually selected based on a collection management plan.

Compression of text: the study of techniques for representing text in fewer bytes or bits.

Content-based query: query exploiting data content.

Conversion: changing from one form to another, as in converting from analog to digital (also called "digitization"), or paper to online (as in "retrospective conversion" of a card catalog to an online catalog, or old books to scanned images.

Cystic Fibrosis collection: a reference collection composed of 1239 documents indexed with the term *cystic fibrosis* in the National Library of Medicine's MEDLINE database.

DLI: Digital Libraries Initiative, a program of the US National Science Foundation, for research and development related to digital libraries, which began with $24M of funding split across 6 universities for 1994-98, and which will continue from 1998 onward with roughly double that amount of support.

DTD: Document Type Definition: SGML definition for a markup language.

Data cartridge: data structure and associated methods to represent and query a particular multimedia data type.

Data retrieval: the retrieval of items (tuples, objects, Web pages, documents) whose contents satisfy the conditions specified in a (regular expression like) user query.

Database industry: the organizations, including commercial, government, and not-for-profit sectors, who produce, provide access to, and market databases of bibliographic, reference, and full-text information.

Database producers: the organizations, including commercial, government, and not-for-profit sectors, who create electronic abstracting and indexing tools as well as other reference and full-text databases.

Database vendors: the organizations, primarily in the commercial and not-for-profit sectors, who license databases from their producers and provide search software and a front end for consumer access to the information contained.

Digital library: the combination of a collection of digital objects (repository); descriptions of those objects (metadata); a set of users (patrons or target audience or users); and systems that offer a variety of services such as capture, indexing, cataloging, search, browsing, retrieval, delivery, archiving, and preservation.

Digital object: some string of bits that is viewed as an entity in its own right (e.g., a full-text document) though it may be a part of another digital object (e.g., an image that is part of a book), often with associated ”metadata” and sometimes with terms and conditions (especially on access).

Digital preservation: ensuring that a digital object continues to be accessible and useful over a long period of time, which usually requires both media conversion (copying from one old tape format to a new tape format before the old tapes are no longer readable) and format conversion (changing from some file structure or encoding to a newer one that will continue to be used and understood).

Directory: a usually hierarchical categorization of concepts in a domain of knowledge.

Distributed computing: The application of multiple computers connected by a network to solve a single problem.

Distributed information retrieval: The application of distributed computing techniques to solve information retrieval problems.

Document surrogate: a representation of a document such as the title and a short summary. Surrogates are common to display the answers to a user query.

Document: a unit of retrieval. It might be a paragraph, a section, a chapter, a Web page, an article, or a whole book.

E measure: an information retrieval performance measure, distinct from the harmonic mean, which combines recall and precision.

Edit distance: (between two strings) minimum number of insertions, deletions and replacements of characters necessary to make two strings equal.

Efficiency: A measure of parallel algorithm performance given by: $\phi = \frac{S}{N}$, where S is speedup and N is the number of processors.

Entropy: measure of information defined on the statistics on the characters of a text.

Exact match: mechanism by which only the objects satisfying some well specified criteria, against object attributes, are returned to the user as a query answer.

Extended Boolean model: a set theoretic model of document retrieval based on an extension of the classic Boolean model. The idea is to interpret partial matches as Euclidean distances represented in a vectorial space of index terms.

Extended pattern: a general pattern allowing rich expressions such as wild cards, classes of characters, and others.

Faceted query: a query which is divided in topics and/or facets, each of which should be present in documents in the answer.

Feature: information extracted from an object and used during query processing.

Federated search: support for finding items that are scattered among a distributed collection of information sources or services, typically involving sending queries to a number of servers and then merging the results to present in an integrated, consistent, coordinated format.

Filtering: retrieval task in which the information need of the user is relatively static while *new* documents constantly enter the system. Typical examples are news wiring services and electronic mail lists.

Full text: a logical view of the documents in which all the words which compose the text of the document are used as indexing terms.

Fuzzy model: a set theoretic model of document retrieval based on fuzzy theory.

Generalized vector space model: a generalization of the classic vector model based on a less restrictive interpretation of term-to-term independence.

Global analysis: a reference to techniques of identifying document and term relationships through the analysis of all the documents in a collection. Global analysis is used, for instance, to build thesauri.

Granularity: The amount of computation relative to the amount of communication performed by a parallel program.

Guided tour: a sequence of navigational choices (usually in a hypertext) aimed at presenting the nodes in a logical order for some goal.

HTML: Hypertext markup language of the Web, based on SGML.

Harmonic mean: an information retrieval performance measure which combines recall and precision.

Heaps' Law: an empirical rule which describes the vocabulary growth as a function of the text size. It establishes that a text of n words has a vocabulary of size $O(n^\beta)$ for $0 < \beta < 1$.

Huffman coding: an algorithm for coding text in which the most frequent symbols are represented by the shortest codes.

Human-Computer Interfaces: (HCI) the study of interfaces which assist the user with information seeking related tasks such as: query formulation, selection of information sources, understanding of search results, and tracking of the retrieval task.

Hypertext model: a model of information retrieval based on representing document relationships as edges of a generic graph in which the documents are the nodes.

Independence of index terms: see *term-to-term independence.*

Index point: the initial position of a text element which can be searched for, for example a word.

Index term: (or keyword) a pre-selected term which can be used to refer to the content of a document. Usually, index terms are nouns or noun groups. In the Web, however, some search engines use all the words in a document as index terms.

Index: a data structure built on the text to speed up searching.

Inference network: a probabilistic model of document retrieval based on interpreting documents, user queries, and index terms as nodes of a Bayesian network.

Information retrieval: (IR) part of computer science which studies the retrieval of information (not data) from a collection of written documents. The retrieved documents aim at satisfying a *user information need* usually expressed in natural language.

Information retrieval performance: an evaluation of an information system in terms of the quality of the answers it generates with regard to a set of test queries. The quality of the answer set is evaluated by comparing the documents in it with those in a set of documents (provided by specialists) known to be relevant to the test query in focus.

Informative feedback: information to the user about the relationship between the query specification and the documents retrieved.

Interoperability: the working together of a number of computer systems, typically for a common purpose, such as when a number of digital libraries "support federated searching", often enabled by standards and agreed-upon conventions including data formats and protocols.

Inverted file: a text index composed of a vocabulary and a list of occurrences.

KWIC: KeyWords In Context, a technique that displays the occurrences of query terms within the context of the documents retrieved.

Keyword: see Index term.

Kohonen's feature map: a bi-dimensional map whose regions represents the main themes in a document or in a collection.

Latent semantic indexing: an algebraic model of document retrieval based on a singular value decomposition of the vectorial space of index terms.

Levenshtein distance: see Edit distance.

Lexicographical order: order in which the words are listed in a dictionary or telephone guide.

Local analysis: a reference to techniques of identifying document and term relationships through the analysis of the documents retrieved by a given user query.

Local context analysis: a technique of query expansion which combines local and global analysis.

Logical view of documents: the representation of documents and Web pages adopted by the system. The most common form is to represent the text of the document by a set of indexing terms or keywords.

MARC: a standardized record format used by libraries and bibliographic utilities for sharing and storing cataloguing information about bibliographic materials.

MIMD: A parallel computer architecture consisting of multiple instruction streams and multiple data streams.

MISD: A parallel computer architecture consisting of multiple instruction streams and a single data stream.

Meta search: A search technique common on the World Wide Web where a single point of entry is provided to multiple heterogeneous back-end search engines. A meta search system sends a user's query to the back-end search engines, combines the results, and returns a single, unified hit-list to the user.

Metadata: Attributes of data or a document, usually descriptive as author or content, often broken up into categories or facets, typically maintained in a catalog, sometimes recorded according to a scheme like the Dublin Core or MARC.

Model for IR: a set of premisses and an algorithm for ranking documents with regard to a user query. More formally, a IR model is a quadruple [$\mathbf{D}$, $\mathbf{Q}$, $\mathcal{F}$, $R(q_i, d_j)$] where $\mathbf{D}$ is a set of logical views of documents, $\mathbf{Q}$ is a set of user queries, $\mathcal{F}$ is a framework for modeling documents and queries, and $R(q_i, d_j)$ is a ranking function which associates a numeric ranking to the query q_i and the document d_j.

Modeling: the part of IR which studies the algorithms (or *models*) used for ranking documents according to a system assigned likelihood of relevance to a given user query.

Multidimensional Scaling: (MDS) a method to map objects into points, trying to preserve distances.

Multimedia Database Management System: system able to represent, manage, and store multimedia data.

Multimedia Information Retrieval System: Information Retrieval system handling multimedia data.

Multimedia data: data combining several different media, such as text, images, sound and video.

Multitasking: The simultaneous execution of multiple, independent tasks. On a single processor machine the tasks share the processor and their execution is interleaved sequentially. On a machine with multiple processors the tasks may execute concurrently.

***n*-gram:** any substring of length n.

Nearest-neighbor query: a query that requests the spatial object closest to the specified object.

Neural networks: an algebraic model of document retrieval based on representing query, index terms, and documents as a neural network.

Non-overlapping lists model: a model of retrieving structured documents through indexing structures implemented as non-overlapping lists.

OPAC: (Online Public Access Catalogue) library management system software which provides user access to information contained in a library collection.

Object-relational database technology: technology that extends the relational model with the main features of the object-oriented one.

Occurrence list: a data structure which assigns to each text word the list of its positions in the text.

Online Public Access Catalogue: see OPAC.

Panning: (see also zooming) action of a movie camera that scans sideways across a scene. This effect can be simulated in a screen window even in the absence of the camera.

Parallel computing: The simultaneous application of multiple processors to solve a single problem. Each processor works on a different part of the problem.

Parallel information retrieval: The application of parallel computing to solve information retrieval problems.

Pattern: set of syntactic features that describe the text segments to be matched, ranging from simple words to regular expressions.

Precision: an information retrieval performance measure that quantifies the fraction of retrieved documents which are known to be relevant.

Probabilistic model: a classic model of document retrieval based on a probabilistic interpretation of document relevance (to a given user query).

Proximal nodes model: a model of retrieving structured documents through a hierarchical indexing structure.

Query expansion: a process of adding new terms to a given user query in an attempt to provide better contextualization (and hopefully retrieve documents which are more useful to the user).

Query preview: a low-cost, rapid-turnaround visualization of what the results of many variations on a query might be.

Query: the expression of the user information need in the input language provided by the information system. The most common type of input language allows simply the specification of keywords and of a few Boolean connectives.

Range query: A query that requests all the spatial objects that intersect the given range.

Recall: an information retrieval performance measure that quantifies the fraction of known relevant documents which were effectively retrieved.

Reference collection: a collection of documents used for testing information retrieval models and algorithms. A reference collection usually includes a set of documents, a set of test queries, and a set of documents known to be relevant to each query.

Regular expression: general pattern that allows to express alternative strings, repetitions and concatenations of substrings.

Repository: a physical or digital place where objects are stored for a period of time, from which individual objects can be obtained if they are requested and their terms and conditions are satisfied.

Retrieval task: the task executed by the information system in response to a user request. It is basically of two types: *ad hoc* and *filtering*.

Retrieval unit: the type of objects returned by an information retrieval system as the response to a query, e.g. documents, files, Web pages, etc.

SGML: Standard markup metalanguage. HTML is a markup language based in SGML.

SIMD: A parallel computer architecture consisting of a single instruction stream and multiple data streams.

SISD: A traditional sequential computer architecture consisting of a single instruction stream and a single data stream.

SMP: (Symmetric Multiprocessor) A common MIMD computer architecture where all of the parallel processors have equal access to the same system resources, and any processor can execute any task. The processors may operate independently, or they may cooperate to execute a parallel task concurrently.

SQL3: the upcoming SQL standard for relational databases.

STARTS (Stanford Proposal for Internet Meta-Searching): A protocol for distributed, heterogeneous search developed at Stanford University in cooperation with a consortium of search product and search service vendors.

Scather/Gather: a browsing strategy which clusters the local documents in the answer set dynamically into topically-coherent groups and presents the user with descriptions of such groups.

Search history: a mechanism for tracking the history of a user session or of a collection of user sessions. The search history should show what the available choices are at any given point, what moves have been made in the past, short term tactics, and annotations on choices made along the way.

Search starting point: “where” or “how” an information seeking task is initiated. Search interfaces should provide the users with good ways to get started.

Semi-static text: a text collection which does not change too frequently.

Semi-structured data: data whose structure may not match, or only partially match, the structure prescribed by the data schema.

Sequential or on-line text searching: the problem of finding a pattern in a text without using any precomputed information on the text.

Signature file: a text index based on storing a signature for each text block to be able to filter out some blocks quickly.

Source selection: The process of selecting one or more document collections from a set of document collections where the selected collections are most likely to contain documents relevant to the query. Commonly used in distributed IR systems.

Spatial Access Methods: file structures for fast storage and retrieval of spatial objects like points, lines, polygons.

Speedup: A measure of performance for parallel algorithms defined in Section **??**.

Statistical text compression: a reference to techniques and methods of text compression based on probability estimates for the rates of occurrence of each symbol.

Stemming: a technique for reducing words to their grammatical roots.

Stopwords: words which occur frequently in the text of a document. Examples of stopwords are articles, prepositions, and conjunctions.

Suffix tree and suffix array: text indices based on a lexicographical arrangement of all the text suffixes.

Syntax tree: structural interpretation of a query, where the nodes are the operators and the subtrees are the operands.

Tag: a string which is used to mark the beginning or ending of structural elements in the text.

Term-to-term independence: a fundamental assumption underlying the classic vector model which states that the vectors used for representing index terms form an orthonormal basis. Such assumption is usually interpreted to mean pairwise orthogonality (but not in the generalized vector space model).

Text structure: information present in a text apart from its content, which relates its different portions in a semantically meaningful way.

Thesaurus: a data structure composed of (1) a pre-compiled list of important words in a given domain of knowledge and (2) for each word in this list, a list of related (synonym) words.

TREC collection: a reference collection which contains over a million documents and which has been used extensively in the TREC conferences. The TREC collection has been organized by NIST and is becoming a standard for comparing IR models and algorithms.

User information need: a natural language declaration of the informational need of a user. For instance, *find documents which discuss the political implications of the Monica Lewinsky scandal in the results of the 1998 elections for the U.S. Congress.*

User relevance feedback: an interactive process of obtaining information from the user about the relevance and the non-relevance of retrieved documents.

Vector model: a classic model of document retrieval based on representing documents and queries as vectors of index terms. The model adopts as foundation the notion of term-to-term independence.

Vocabulary: the set of all the different words in a text.

WAIS: (Wide Area Information Service): suite of protocols designed to publish information and to allow querying databases in Internet.

Working memory load: information about the decisions and choices made during an interactive user session. A system with a well designed interface should keep this information around to facilitate the task of the user (for instance, allowing the user to easily resume temporarily abandoned strategies).

XML: A subset of SGML, defined for the Web. In XML is easier to define new markup languages.

Z39.50: a national standard that has become an international standard. It is a protocol for client-server communication with retrieval systems, allowing a single client to interact with one or more systems, or retrieval systems (or gateways) to communicate with other retrieval systems. It supports connection-oriented sessions, having a system explain itself, submission of queries, obtaining information, the results lists, and retrieved documents.

Zipf's Law: an empirical rule that describes the frequency of the text words. It states that the i-th most frequent word appears as many times as the most frequent one divided by i^θ, for some $\theta \geq 1$.

Zooming: (see also panning) action of a movie camera that can either move in for a closeup or back away to get a wider view. This effect can be simulated in a screen window even in the absence of the camera.