INFORMATION RETRIEVAL

Lecture Notes

Massimo Melucci

University of Padova

2025

Contents

1

Introduction

> I have only made this letter longer because I have not had the time to make it shorter.
>
> Pascal (1657)

1.1 Memex, Search Engines, and Generative Machines

Information Retrieval (IR) means the set of models, methods and computer systems for the representation and retrieval of all and only the information relevant to any information need of any user in any context. IR has a long history when compared to other computer science disciplines. Indeed, the first hints of the organization, representation and retrieval of information date back to the forties of the last century when Bush wrote that

> Science may implement the ways in which man produces, stores, and consults the record of the race. It might be striking to outline the instrumentalities of the future more spectacularly, rather than to stick closely to methods and elements now known and undergoing rapid development, as has been done here.

Surprisingly, this statement was made in from 1945 by Bush who was scientific adviser to the presidency of the United States of America during and after World War II. He also wrote:

> Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin

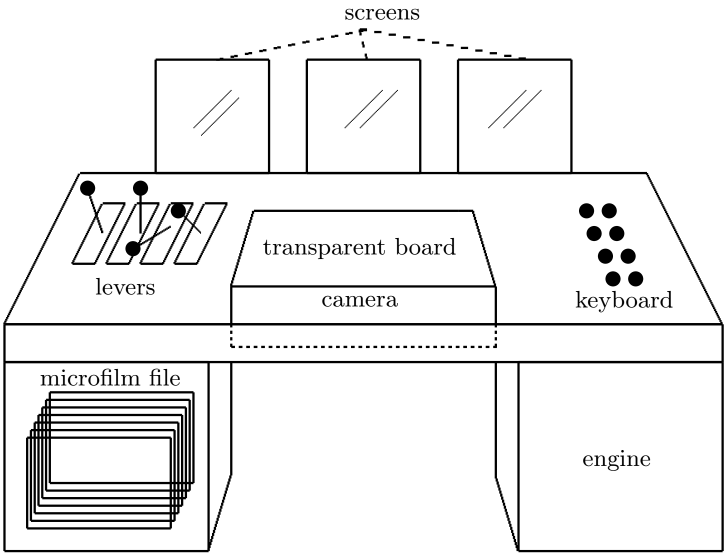

Fig. 1.1. – The memex imagined by Bush (1945)

> one at random, “ memex” will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.

memex

The hardware components of the memex depicted in fig. 1.1 bear little resemblance to those of a modern information processing system, although they appear to be arranged on a desktop and have many of the functionalities of commonly used technological devices. The aspect that makes Bush’s contribution still relevant is not so much the visionary one as the concreteness that transpired from the list of technologies already available at that time to create the memex. However, it took about twenty years before computer systems could be seen with some of the features of the memex – one of these systems was the hypertextual one.

After Memex and Bush, there has been a long series of real inventions that are characterizing this century and are the result of decades of study and research. Figure 1.2 summarizes some of the innovations since the forties; The figure emphasizes methodological innovations of a software nature while further references are given in the bibliography suggested at the end of this chapter.

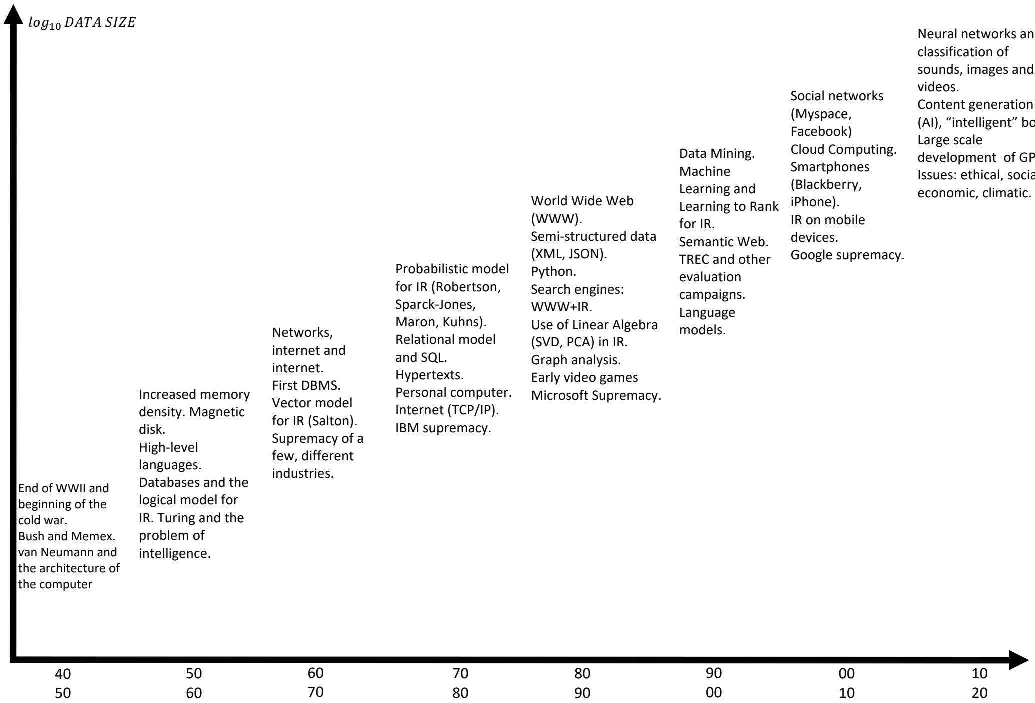

Fig. 1.2. – A short history of the technology behind IR

Because of the significance of the World Wide Web (web) in IR and the technologies exploiting search functions, it is important to start from hypertext. A hypertext is a network of nodes containing data connected by links of different types and functions. Hypertext has marked a significant stage of the evolution of information technologies from the memex to date, since it has been a tool capable of emulating the mechanism of the mental association. The first studies on hypertext were completed in the Sixties when the mouse and the earliest terminals equipped with a screen appeared. It was Nelson (1987) to coin the term "hypertext" and to create Xanadu which was a hypertext system for literary texts. In the same period numerous industrial products appeared and the sectors of electronic publishing, digital learning and cooperative work were developed. Before Nelson, Englebart (1986) created a data management system, called NLS, which was a hypertext *ante litteram* for software documentation and design having many of the characteristics of the current electronic books and magazines.

hypertext

oN Line System (NLS)

disorientation
cognitive overload
informational myopia

Despite its great potential, hypertext is not in itself an efficient and effective tool for finding information due to three main problems: first, the disorientation of those who visit the hypertext, second, the cognitive overload to which the person in question is subject for the management of the different paths of access to the information contained in the nodes and, finally, the informational myopia, i.e. the tendency to look at the information contained in the nodes without understanding the reasons for the links between the nodes placed by the author. Disorientation and cognitive overload are the reasons that lead the person who visits the hypertext to take a long time to orient himself among the network of connections, making it so inefficient. Information myopia is the reason why hypertext is ineffective: visitors do not understand or agree with the semantics given to links by the author and use these links ineffectively and inconsistently.

Berners-Lee has been a physicist at the European Centre for Nuclear Research in Geneva (CERN). He published the introductory article to web in 1989 and presented to his superiors a project to replace file system-based computer systems for the exchange of documentation within CERN with a hypertext, thus following Engelbart's NLS aim. The distribution of the hypertext on a computer network was the innovation brought about with the web and not so much the hypertext network itself. Indeed, in his report of the 1989, the CERN physicist described the communication protocol between servers and clients, called HyperText Transfer Protocol (HTTP), and a language for organizing and presenting documentation, called HyperText Markup Language (HTML).

In 1994 the technologies of data management systems that had already been available since the Sixties for the large database management systems were the answer to the request for greater efficiency and effectiveness coming from the users of the web. That's when search engines appeared. By leveraging the technologies of data management systems, search engines gave the possibility of querying a computer system capable of returning web pages in response to users' queries in a short time. The features of the search engines appeared in the Nineties are still available, although with some differences, for example, the advanced search mode is now less visible than before, since it is used by a marginal minority of users. Moreover, current engines integrate different functionalities such as e-mail or online socializing, thus becoming an integral part of plenty of information technology applications.

Large Language Model (LLM)

Towards the end of the 2010s, the notion of LLM was coined and large language models could be developed. A LLM is an information structure by which content, especially textual content can be generated through a series of word predictions to be concatenated to the words already generated. The large language models have changed the way content is produced and used; instead of having human authors writing the documents, there is a machine that autonomously produces text, images, sounds and videos; instead of receiving ordering lists of references to found documents, such as links to web pages, the user receives a document generated at the time of the request and therefore not already written by human authors and indexed by IR system.

1.2 Information Retrieval

Information is what modifies knowledge, which is the coordinated set of facts identified, chosen and acquired by a person in the course of his or her life in relation to one or more topics when solving problems or carrying out tasks.

task
problem

A distinction is being made between "task" and "problem" because the former is a part of the work that is assigned to others or that someone sets out to do, while the latter is a situation, chance, fact that presents difficulties to be faced and solved. Therefore, the task has an executive meaning, while the problem has a design meaning and can be articulated in a series of tasks once the problem has been articulated as sub-problems.

data

Information takes on a concrete form through data, i.e. symbols that, once processed, form the knowledge of those who learn the information contained in them and interpret it. Examples of tasks are:

- the search for the best information resources related to a topic (resource finding), *resource finding*
- the identification of a web page of which you know exists, but you do not remember the Uniform Resource Locator (URL) (homepage finding), *homepage finding*
- the answer to questions related to the "who-how-when" of an event (question answering), *question answering*
- the search for information made with the contribution of members of one's social groups (social search), *social search*
- the identification of experts in a given field (expert search) *expert search*
- the search for industrial patents (patent search), *patent search*

chemical search

genomic retrieval

- the retrieval of scientific information (chemical search, genomic retrieval),
- the search for multimedia information (images, videos, music, written text and spoken text).

information need

Anomalous State of Knowledge (ASK)

An information need is the set of circumstances in which a person has a problem to solve or a task to perform and requires important, useful, or necessary information to solve the problem or perform the task. An information need can be seen as a state of knowledge deficient called in the literature ASK, i.e. a state in which the user does not have the knowledge to operate and make decisions related to the activities he has to conduct.

is the property that makes information important, useful or necessary to satisfy a person's need for information. Information needs and relevance are essential for the person who only can judge whether information is relevant to his or her information needs. Relevance cannot automatically be determined by a machine and with reasonable computational costs because relevance depends on context, i.e., what is relevant to one person, in one place, in one period, for one task or one problem, may no longer be relevant to another person, in another place or time, for another task or problem.

context

An example of apparently simple, yet difficult to automatically manage context dependence, occurs when the person scrolls through a list of bibliographic references containing possibly relevant information, judges one reference relevant and, then, judges the next one to be irrelevant because the previous one has in the meantime already satisfied his or her information need. Instead, the person her/himself could have considered the subsequent reference useful if the two documents had been assessed in reverse order.

pertinence

aboutness

Note that relevance is different from pertinence or aboutness. The latter two refer to the relationship of something to something else whereas relevance refers to the relationship between information and information need. Thus, information may be pertinent, but not very thorough for a person's purposes. On the other hand, there may be little pertinent, yet very relevant information, such as a book about Molecular Biology which contains information very relevant to Microelectronics for some researchers, although it might superficially be considered as little pertinent because of its different subjects.

Information Retrieval (IR)

IR identifies all the activities used to find information relevant to an information need of a certain user. Relevant information is almost always found together with irrelevant information – the hardness of IR lies precisely in having to find all and only relevant information and , at the same time in getting rid of all and only irrelevant information. "Information Retrieval" and the acronym thereof are worldwide adopted. The corresponding expressions in languages other than English are scarcely used, even in the domestic scientific literature, although they would be preferable from a purely linguistic point of view.

document

In IR, "document" is used to refer to a persistent and uniquely identifiable container of data; Examples of documents are books, book chapters, scientific articles, newspaper or magazine articles, videos, speeches and images. The use of "document" has now become consolidated even for those objects that in general have little to do with the notion of document commonly understood. If, according to common sense, a book or a newspaper article can be understood as documents, because they have the role of "documenting", in IR, even a fragment of an image, an audio sequence or a web page containing advertisements, if they are equipped with an identifier, they are considered "documents" because the information represented by those data contributes to the formation of a person's knowledge.

identifier

The identifier of a document allows access to the document independently of its content, particularly when there are identical copies of the same document produced at different times that need to be treated as distinct documents.

collection

A collection of documents is the set of documents to be represented, described, stored and managed automatically for IR purposes. The collection of documents is a homogeneous set in terms of content and collection criteria of the documents that are part of it when designed for a specific domain such as Medicine or Mathematics. Therefore, the criteria by which the documents are part of it are known and predefined; indeed, an element is part of a whole when the criteria of belonging established a priori by the designer of the collection are met.

size of the collection

A document collection may be of arbitrary size, but it is destined to grow. The size of the collection depends on three elements: the size of the individual documents, the number of documents that are present in the collection at any given time, and the speed at which the collection is fed and grows over time.

As said, a computer system is not able to understand a person's information need as long as the latter remains closed in his mind and not completely and precisely represented. A representation of an information *query*

need is anyway necessary yet not sufficient to the system to find relevant information. The typical form of representation of an information need is called query. One speaks of natural language queries or, in the specific case of written expressions, of free-text queries when a user expresses the information need with the language used to express himself in his or her own language.[1].

media

Documents and queries can contain data from one or more media, such as:

- text, still or moving images, sound, music, spoken speech;
- environmental parameters (e.g. noise and temperature);
- software applications (e.g. apps for mobile devices).

Text is certainly the most used and “simple” *medium.*[2] However, it presents several pitfalls, the most important of which are synonymy and polysemy.

synonymy

Synonymy indicates the condition of substitutability of one linguistic element with another in the given context, without resulting in an alteration of meaning; for example, “edge” can be substituted by “border” or “slope” depending on context.

polysemy

Polysemy indicates coexistence of different meanings in a word; for example “bank” can indicate the land alongside or sloping down to a river or lake or a financial establishment that uses money deposited by customers for investment, pays it out when required, makes loans at interest, and exchanges currency.Oxford Dictionary

language

In case of written text or spoken speech, more than one language can be used to write documents or queries. If different languages are used together, the language used for a query may be different from the language used in a document. Moreover, a single document or a specific query can be drafted in different languages even in the same text as in the following cases:

- a query can be expressed in Italian, as ”reperimento dell’informazione”, or in English, as ” Information Retrieval”;
- A sentence, such as ”information retrieval system”, contains words from different languages;

[1] Note that, in the field of relational databases, “query” is used to indicate a sequence of statements written in Structured Query Language (SQL).

[2] Note that *media* is a Latin word and the plural forrmog *medium.* Although the plural form should be distinguished from the singular form, “in the sense ‘television, radio, and the press collectively’, it behaves as a collective noun (like staff or clergy, for example), which means that it is now acceptable in standard English for it to take either a singular or a plural verb.” Oxford Dictionary

Language	2023	2025
English	55.5%	49.3%
Spanish	5.0%	6.0%
Russian	4.9%	3.9%
German	4.3%	5.6%
French	4.4%	4.4%
Japanese	3.7%	5.1%
Portuguese	2.4%	3.8%
Turkish	2.3%	1.8%
Italian	1.9%	2.7%
Persian	1.8%	1.2%
Other	13.8%	16.2%

(a) Distribution of languages used in the visible web pages

User Language	2020
English	25.9%
Chinese	19.4%
Spanish	7.9%
Japanese	2.6%
Portuguese	3.7%
German	2.0%
Arabic	5.2%
French	3.3%
Russian	2.5%
Other	23.1%

(b) Distribution of languages used in queries sent to search engines

Fig. 1.3. – Distribution of languages

- A word like ”file” can exist in the vocabulary of a language with a certain meaning and, at the same time, in that of another language, but with a different meaning.

multilinguism

Multilinguism in IR is the presence of multiple languages within a collection, document or query. Therefore, an IR system may have to deal with a multilingual document or a multilingual collection of documents which may or may not be multilingual. English and the various national variations thereof are the one most present in the web and, in general, in the cataloguing databases, but it is not the only one and will probably be less and less majority. In this regard, fig. 1.3 shows the distribution of the languages of the web.[3]

Cross-Language IR (CLIR)

Multilinguism is connected to CLIR which is the retrieval of documents in one or more languages to answer queries in different languages. CLIR requires the translation of a text from one language to another by using dictionaries, for example. refined, do not eliminate the influence of the ambiguity of the language on the result of the translation; Although machine translation has been the subject of research for decades, even at an industrial level, the accuracy ánd appreciable are in some cases. However, in IR, accuracy ís only necessary for translating individual words, since translating grammatical structures does not affect the overall effectiveness of IR.

[3] `https://en.wikipedia.org/wiki/Languages_used_on_the_Internet`, February 22, 2025

Geographic Location	$P(C)$
Africa	0.037
Antarctica	0.003
Asia	0.171
Europe	0.367
Latin America and the Caribbean (LAC)	0.053
Northern America (NA)	0.324
Oceania	0.045

Fig. 1.1. – Distribution of documents across the main geographical locations including those in an unknown area. LAC stands for Latin America and the Caribbean and NA stands for Northern America.

geograophical location

The phenomenon of multilinguisim is naturally connected to that of geograophical location. Although the web pages are mainly written in English and a few other languages, the organizations of the authors of the pages may located in different geographical areas. Consider the document collection built from Wikipedia in order to allow the experiments carried out at the Fair Ranking track of the 2021 Text REtrieval Conference (TREC)'s edition. The documents have been referred to one or more geographical locations inmplemented as subcontinents as listed in the first column of Table 1.1 where the second column reports the weight of geographical locations.[4] The weight of a geographical location was calculated by summing the scores of the retrieved documents assigne to the geographical locations whereas the scores are the sum of the weights of the terms describing the topics of the experimental collection. The reported data clearly show the unbalance among the subcontinents and the association between weights and economic development levels to a degree that the more developed the subcontinent the larger the *a priori* probability that a document of a certain subcontinent can be retrieved.

Information Retrieval System (IRS)
user

A IRS is a computer system or part of a computer system designed and built to automatically perform IR tasks requested by a user. A user is thus a person who uses a IR system. An IR system carries out its tasks for collections of documents of any size, ensuring the description of the information content of the documents and the rapid retrieval of those that represent information relevant to the information needs expressed through end-user queries. The functional architecture of a IR system, shown in fig. 1.4, and the imaginative one in fig. 1.5, include several parts, both those relating to the collection and representation of the information content of documents and queries, and those relating to interaction with the end user.

[4] Geographical location was unavailable for about 40% documents.

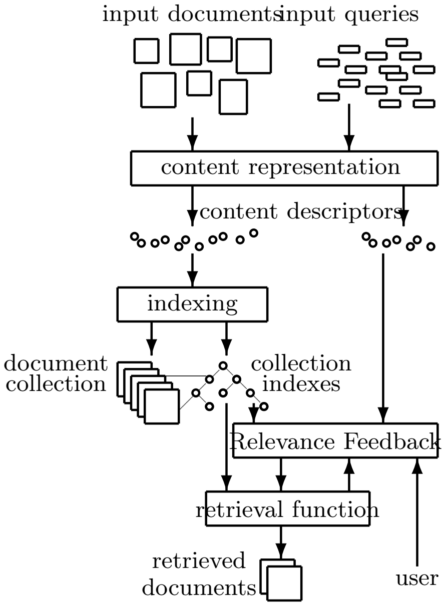

Fig. 1.4. – Functional architecture of an Information Retrieval System

Information needs can also be represented starting from user behavior: for example, a sequence of visited web pages can represent the user's interest in the topics they deal with, and can give useful information on the user's information needs who has visited them. As user behaviour may provide information about information needss, queries are not necessarily textual when the information found is contained in non-textual documents such as images, videos or sound, i.e. the need for information could be expressed only through images, videos or sounds, or through user behavior such as eye eye tracking, mouse tracking or (CTD).

user behavior

eye tracking

mouse tracking

Click-Through Data (CTD)

A descriptor is a piece of data that expresses the salient aspects of the information content of a document, a query or in general of another piece of data. A descriptor of a text is called keyword or index term (in English, *keyword* or *index term*): the first meaning refers to single words with a key role in the description of the content; On the other hand, "index term" refers to groups of two or more key words or phrases linked by grammatical structures, such as "information retrieval", "information retrieval system" and "retrieval of information". If the document and query are of a different kind, such as images or pieces of music, the descriptors can be significant fragments of an image or

descriptor

keyword

index term

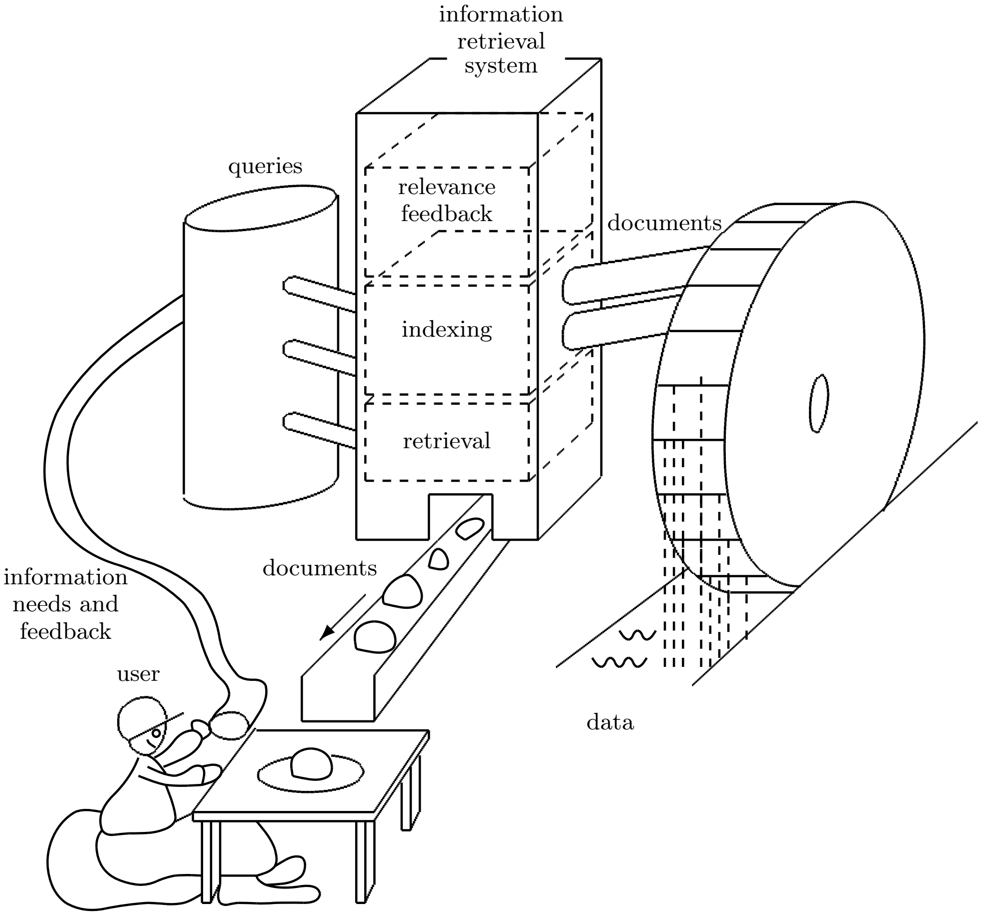

Fig. 1.5. – Information Retrieval System

the audio spectrum of a piece of music. If a document contains different media, the descriptors will be of different types.

media

indexing

Indexing consists of describing the information content of the documents in the collection. By indexing, an IR system assigns descriptors to each query or document. The result is a set of reference tables, called “indexes” which are sorted according to appropriate criteria, recorded in the memory of a computer system and searched to find the identifiers of the documents in the collection.

posting

A posting is the assignment of a descriptor to a document. The notion of posting is similar yet not identical to that of label or tag commonly used to describe the information content of documents usually available

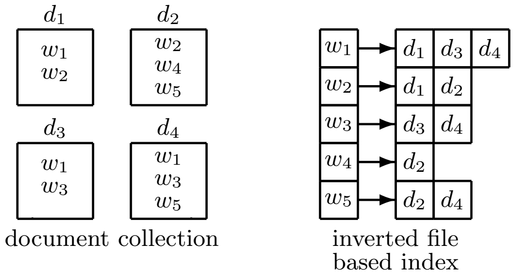

Fig. 1.6. – Document collection and document index

through the web. A posting differs from a tag in the presence of various statistical data that measure the contribution of the descriptor to the document.

posting list

A posting list is an information structure stored in the memory of the IR system, which collects postings related to a descriptor. For example, a posting list for a text document collection is the list of identifiers of documents in which a certain keyword is present.

index

dictionary

An index resulting from indexing a document collection is an information structure that collects the posting lists for a given set of descriptors. An index stores a dictionary, i.e. a set of descriptors of the documents of the indexed collection which may be utilized by users to search for information in the collection. Each descriptor of the dictionary of an index is linked to a posting list.

transposed file

inverted file

The organization of the data of an index is based on a transposed file or inverted file as shown in fig. 1.6. There is an index in correspondence with each part of the document or with each indexing algorithm; for example, if a document may be structured into “author”, “title” and “body” there may be an index for the authors, one for the words in the title and another for the words in the body. If an algorithm that extracts the linguistic root of each word were used, one index would be created in which the descriptors are the words and another index would also created in which the descriptors are the roots. No index is generated for user’s queries except if the queries are stored for future use such as CTD analysis.

Relevance Feedback (RF)

An IR system can implement RF by using the input which is explicitly or implicitly provided by the user about his or her information needs or about the relevance of some documents chosen from those found by the system as response to a query. By means of RF, the system allows the user to refine his query, manually entering descriptors that s/he thinks could prove to be effective in finding relevant documents, and eliminating those that are not considered effective. In case of explicit RF, the user indicates some relevant documents among those returned by the system in response to a query; then, the system extracts significant descriptors from the relevant documents and inserts them into an automatically constructed query in order to finally produce a new list of documents.

1.3 Search Engines

search engine

A search engine for the web (or, for short, "search engine" as is known to a wide audience of users) is a special case of IR system. Like the latter, a search engine knows its collection of documents at all times, since the methods of automatic collection of web pages are designed on the basis of predefined criteria. The fact that the pages collected by the search engine are distributed in the web is little significant if not from the point of view how the collection should be feeded, because the collection of pages managed centrally and separately from the web does not store the original and up-to-date document content. Bottom line: a search engine only stores one or more collection indexes fueled by search agents which harvest content from the web by means of Internet – the contet is kept at the server where it was harvested from by a search agent.

From an application point of view, search engines are not the only special case of IR system. Some applications are little known to the general public which might be unaware of them. However, those applications are no less important than the web search engines such as the IR system used in the medical field. In other cases, those applications are used so frequently that they are considered as an ordinary technology that no longer needs to be noticed such as the IR systems dedicated to search:

- electronic mailboxes,
- file systems,
- web sites,
- local networks (e.g., enterprise search),
- library catalogues,
- digital museums,
- electronic archives.

enterprise search

Search engines exhibit some differences them from an IR system generally understood. Documents are web pages, written in HTML, often

HyperText Markup Language (HTML)

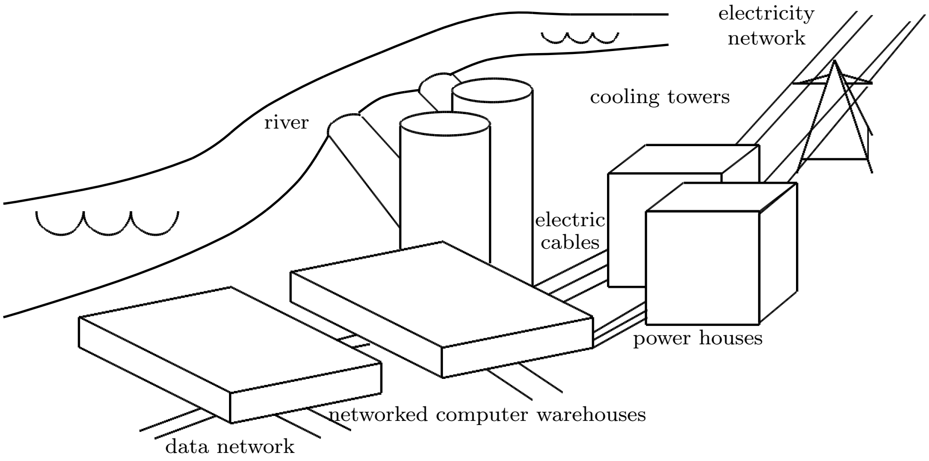

Fig. 1.7. – Data center

automatically generated, and not necessarily products of an editorial process such as books or newspaper articles. Moreover, search engines store only the URL and not the entire content of the pages which remain stored and managed by the server in which they were found. Consequently, a page can be changed over time or even deleted without the search engine or the user can be aware of.

Although search engines do not normally store the contents of documents, they do keep the latest versions of indexed pages in a cache memory, which are only one version of those stored in servers and are not necessarily the most recent versions. *cache*

Compared to traditional IR systems, search engines face greater problems of scalability. The mass of information to be indexed and retrieved and that of users to be reached with IR services are of a size that determines high computational costs. The computing infrastructures of the search engine companies require dozens of computing sites and several warehouses for each site, thus raising problems related to energy consumption, cooling and preservation of the natural environment in which the sites are installed: the complexes dedicated to a search engine is called data center (fig. 1.7). *scalability* *data center*

Contrary to what one might think, search engines index and allow users to access only to a small part of the entire web which is depicted by the smallest cube in fig. 1.8. The limited number of accessible pages is partly desired and partly suffered by the companies that manage search

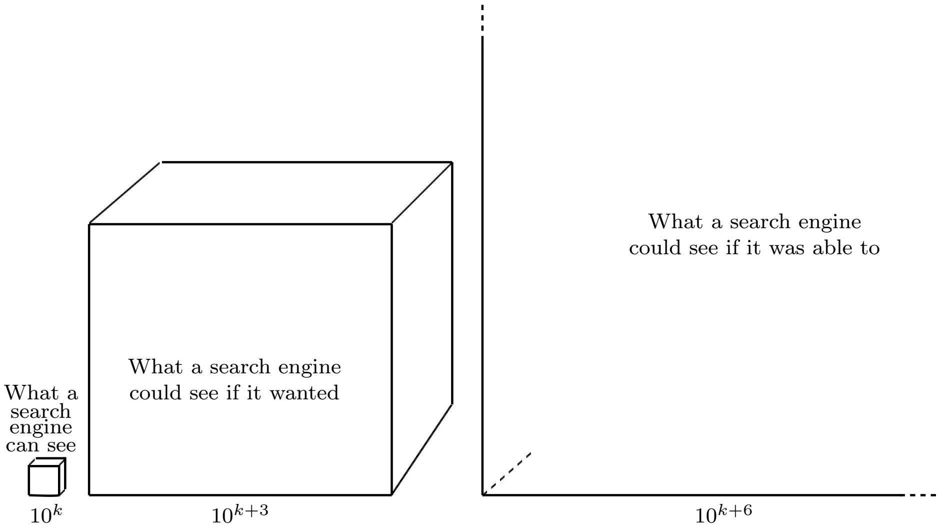
What a search engine can see
10^k
What a search engine could see if it wanted
10^{k+3}
What a search engine could see if it was able to
10^{k+6}

Fig. 1.8. – Accessible, visible and invisible web

engines. A search engine cannot index all the pages of the web possible because of its own technological limits, organizational choices made by the search engine companies, or regulatory constraints imposed by the country that hosts the data centers. As a consequence, many of the pages written in languages other than English or containing unwelcome material are simply ignored.

visible web

The mass of pages that an engine might yet does not index is called visible web which is depicted by the second largest cube of fig. 1.8. The size of the visible web is far greater than the size of the indexed web. Nevertheless, the visible web represents a minority of the whole web

deep web
invisible web

The largest mass of data that users use is called deep web or invisible web which is depicted by the largest cube in Fig. 1.8. The deep web is made up of dynamic pages generated only upon demand such as the pages with the lists of hotels in a city that can be booked on a certain date. The deep web pages are dynamic because the data of interest, e.g. hotel names and addresses, customer reviews are stored in databases and extracted from them to compile the pages. The term "deep web" derives from the idea of depth of the abysses in fig. 1.9.

index size

The index size has been an essential feature of the effectiveness of a search engine since the Nineties. The greater the number of indexed pages, the greater the probability that the pages store relevant information. That is why the number of indexed pages was what distinguished one

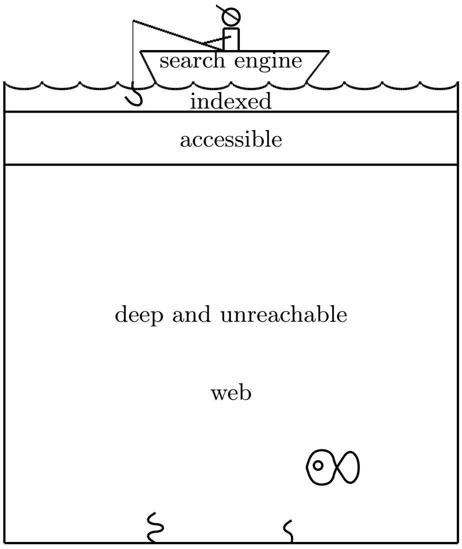

Fig. 1.9. – Deep Web

search engine from others, especially during the early years of the advent of web dearch technologies. Large index size led users to prefer those engines that claimed to be able to index a greater number of pages than competing engines. Aware of the attention that users paid to the number of indexed pages, the search engine companies were used to report the number of indexed pages in the home pages.

The aim of reporting the index size was to attract a large number of users, the latter being an aspect that became crucial over time with the development of digital advertising and electronic commerce on the web, since the major search engines are by now harvesting the largest share of the advertising market. For these reasons, it is not surprising to observe that the few search engines that survived the competition were the ones able to index the largest number of pages.

digital advertising
electronic commerce

1.4 Advancements and Prospects

Since the search engines for the web have appeared, there have been several advances in IR have happened and some prospects for the future have been opened. In this section, we highlight the advances which have in retrospect proved decisive while others could become so.

The development of smartphones since the beginning of this century

smartphones

has certainly represented an important turning point in the IR sector as well. The reasons are many, here they can be summarized like this. On the one hand, the user of a search engine has come into possession of a device with which to access databases without necessarily having to be at a workstation equipped witha Personal Computer (PC) nor having to ask for help from an intermediary. On the other hand, the administrators of the search engine were able to stay connected with the user almost continuously, recording his behavior and then using it for the purpose of modifying queries and offering better results and, especially, advertisements.

social networks

The advent of social networks has allowed users to connect with each other directly as well as in the indirect way offered by the web. The direct connection between users of a social network has implied the autonomous and immediate production of multimedia and multilingual content both in the form of documents and comments and reactions to documents. The IR functions embedded by a social network have gradually taken into account the connections and the nature of the content in order to represent relevance in a different way from what happens with a traditional IR service; for example, the availability of connections between users and the content thereof has allowed the search engine companies to implement innovative ranking algorithms that can take user popularity measures into account.

recommender systems

Recommender systems play a similar role to the role played by IR systems, but with some important differences. The first difference between a recommender system and an IR system is the notion of relevance which stems from the difference between the user need of the latter and the user need of the former. A user of a recommendater system has a problem that generates an economic need *sensu lato*, whether it is satisfied by goods such as electronic items or it is satisfied by services such as entertainment. The other important difference between recommender systems and IR systems is the availability and exploitation of users' reviews on goods and services that allows the so-called collaborative filtering, i.e. the selection of goods and services on the basis of "involuntary" and indirect inter-user collaboration based on the reviews – a user might decide or not decide to buy depending on the other users' reviews.

Large Language Models (LLM)

relevance

The development of LLM changes the way users use the services of IR since the judgment of the relevance of the information represented by a document generated in this way is based only on it and no longer on the comparison with other documents. In addition, the traditional way in which IR systems are evaluated is inadequate in presence of content that is automatically generated only at the time of the request. Indeed, no predefined experimental collection against which to compare different systems can be designed and implemented on the basis of the currently available scientific criteria and knowledge. As a matter of fact, the current system evaluation methodology assumes that relevance assessments can be assigned to documents which have already been authored and made available to the user and that those documents cannot be changed without changing the relevance assessments. If such a change happened the comparison between systems would be infeasible.

In IR, much importance is given to the problems and needs of the users; In fact, systems are designed and built in such a way as to maximize efficiency and in particular effectiveness, i.e. the degree to which a system always and only finds relevant information. The development of web brought out the authoritativeness of documents as an aspect of relevance little considered in previous decades in which collections of documents were built accurately and exhaustively downstream of a document selection process conducted manually by the personnel of the organizations in which the system was operational.

authoritativeness

In the 2020s, in search of methods of IR in which authoritativeness was taken into account, it emerged the need to also take into account the fairness with which the authors of the documents were made visible through the lists of documents retrieved by an IR system. The problem was that an IR system and in particular search engines tend to give greater visibility to authors whose organizations belong to certain social and economic groups. The most visible group are often referred to economically developed countries or to certain genres and languages, thus feeding a vicious circle by which a dominant culture makes itself even more dominant.

1.5 Short Bibliography

Origins and Foundations.

The work mentioned at the benginning of this chapter is Bush (1945)'s whereas the document that describes the first mouse is Engelbart (1967)'s. Mooers (1950) coined "Information Retrieval". Isaacson (2014) provides a history and a biography of significant scientists and contributors to

Computer Science and Engineering. Maron and Kuhns (1960) wrote one of the first articles on the then futuristic IR system based on the theory of probability. Salton (1968) brought out the first book on IR in which the algorithms and data structures of modern search engines are described. Belkin (1980) introduced ASK for the first time. The articles collected in the volume edited by Sparck Jones and Willett (1997) are also suggested for an overview of IR until the Nineties when the search engines were invented. A review of the beginnings and the development of IR is Harman (2019)'s.

Methods and Models.

The systems based on statistical methods were consolidated in the Seventies. Sparck Jones, Jardine and van Rijsbergen (1971) published the first overall results on the statistical methods of automatic classification of documents and words in the same volume. Salton and Buckley (1988) reported on an exhaustive experimentation about statistical term weighting and provided a "blueprint" as regards what weighting schemes in which context. Peters and Braschler (2001) and Nie (2010) provided the basic indications on the development problems of methods and systems CLIR. Rocchio (1971) reported on the first RF techniques whereas Harman (1992) described further experiments. An attempt to revisit the models of IR within a single methodological framework for the enthusiasts of Physics and Mathematics has been made by Van Rijsbergen (2004).

Representation and Indexing.

In the 1950s the works of Luhn (1958, 1960) on the automatic generation of summaries and analytical indices of texts were published. A few years later, Cleverdon and Mills (1963) reported on the first experiments of automatic indexing of text from which the development of modern search engines started. In the same year, Salton (1963) wrote an article in which statistical methods are described to measure semantic relationships between words, such as synonymy and polysemy, as well as to build networks of terms and documents. In the late Sixties, Lovins (1968) described the first algorithm to extract the linguistic root of words and in particular to detect the singular form of English words. The article by Hearst (1992) gives an overview of the methods of solving the problems posed by synonymy and polysemy.

Evaluation and Measurement.

Cleverdon et al. (1966) introduced the experimentation methodology based on test collections. Sparck Jones and van Rijsbergen (1976) published the technical report dedicated to the standard method of evaluation of an IR systems still used in the research and industry of IR. Salton (1971) published the results of the tests on the first experimental system based on the vector spaces that outperformed the systems based on Boolean logic and that was eventually incorporated by the early search engines. Buckley and Voorhees (2000), Sanderson and Zobel (2005) and Zobel (1998) discussed evaluation. A survey of a well-known evaluation measure, i.e. the F-measure was written by Christen et al. (2023).

Textbooks.

Kernighan (2017)'s book is a useful introduction to Computer Science for beginners. For a historical perspective of IR, the reader is suggested to refer to the texts written by Salton (1968); Salton and McGill (1983); Salton (1989) and to that of Van Rijsbergen (1979). The books of Frakes and Baeza-Yates (1992) and Manning et al. (2008) are useful to those interested in algorithmic aspects and data structures. The book of Manning and Schütze (1999) is useful for the linguistic aspects of the text, while that of Belew (2000) in some aspects relating to the user and the interaction with the system. Finally, that of Croft et al. (2009) presents the fundamental aspects of IR and some of the specific ones of the search engines. See the texts of Blair (1990), Ingwersen (1992) and Ingwersen and Järvelin (2005) for aspects relating to the processes of representation of the information, the role of the user and the problems relating to the relevance and context. The book of Baeza-Yates and Ribeiro-Neto (2010), in which there are also chapters written by other researchers, is a complete compendium of the discipline. The most recent, multiple authors, research-oriented textbook was written by Alonso and Baeza-Yates (2024).

Multimedia, Multilingualism, web and Other Developments.

The works of Berners-Lee (1989); Berners-Lee and Caillau (1989); Berners-Lee et al. (1992); Berners-Lee et al. (1994); Berners-Lee et al. (2001), Chakrabarti et al. (1999) and Broder (2002) are important to keep track of the advent and developments of the web. Cutting et al. (1992)

and Jain et al. (1999) are two useful reference points for the techniques of clustering to organize and display large masses of documents to users. The works by Manjunath and Ma (1996), Jeon et al. (2003), Flickner et al. (1995) and Faloutsos and Lin (1995) are relevant to the theme of finding and indexing images. Dumais et al. (2003), Joachims (2002) and Silverstein et al. (1999) have written in relation to the exploitation of CTD to represent information needs. The articles of Marchionini and Shneiderman (1988) and Hearst (1997); Hearst and Plaunt (1993) are useful for hypertext access to documents and those of Agosti et al. (1991, 1996) and Salton et al. (1993) for use and automatic generation of the hyperteys in IR. Clarke et al. (2008) wrote about the techniques adopted by the search engines to condense the greatest number of types of documents relevant in the top ten results. As for the LLMs, the reader is invited to read the works of Bengio et al. (2003); Goodfellow et al. (2016) and those of Mikolov et al. (2013c,a,b). The technical report edited by Culpepper et al. (2018) describes the developments of IR. The book of Jurafsky and Martin (2024) is an effective compendium on this topic. One of the first articles on the presence of the bias in the web is by Baeza-Yates (2018) whereas a review was written by Ekstrand et al. (2022) and another was by Zehlike et al. (2022a,b).

1.6 Questions

1.1 Suppose you have the task of writing the final report of a course of study on the subject entitled ”I'm looking for search engines designed to find apps” and write a text of about 500 words describing an information need that arose during the course of the task. Note that the response must contain references to engines that are specifically designed to find applications themselves. The required text must describe the subject matter and the need for information and give the criteria on the basis of which to distinguish a relevant document from a non-relevant one.

1.2 The following text describes the information needs of the question 1.1.

> A large number of users nowadays want or need apps to meet professional, study or entertainment needs. As regards this topic, there are some issues, methods and systems for searching apps (also known as applications for {lap,net,note}books, tablets or pads). Many developers implement apps for different needs. At a rough estimate, about 300,000 apps are currently available. The app market has been increasing at a significant rate since the advent of smartphone and tablets. An app significantly differs from a

> document because the former provides functions whereas the latter provides information. Thus, standard search technologies can hardly be used to implement an app search. A relevant document addresses issues, methods and systems when searching apps; it may refer to standard search technologies; it may address only one of the issues, methods and systems; it may be a survey or a technical paper. Documents that advertise products or companies are not relevant; popular press (e.g. news or RSSs) is not relevant.

Identify a title, a brief description and the criteria on the basis of which the information represented by the data of a document is relevant to the information need.

1.3 The following text reports the title, the brief description and the criteria (narrative) according to which the information represented by the data of a document is relevant to an information need, starting from the text of the question 1.2.

> title: App Search.
>
> description: Issues, methods and systems for searching apps (also known as applications for {lap,net,note}books, tablets or pads).
>
> narrative: A large number of users nowadays want or need apps to meet professional, study or entertainment needs. As regards this topic, there are some issues, methods and systems for searching apps (also known as applications for {lap,net,note}books, tablets or pads). Many developers implement apps for different needs. At a rough estimate, about 300,000 apps are currently available. The app market has been increasing at a significant rate since the advent of smartphones and tablets. An app significantly differs from a document because the former provides functions whereas the latter provides information. Thus, standard search technologies can hardly be used to implement an app search. A relevant document addresses issues, methods and systems when searching for apps; it may refer to standard search technologies; it may address only one of the issues, methods and systems; it may be a survey or a technical paper. Documents that advertise products or companies are not relevant; popular press (e.g. news or RSSs) is not relevant.

Write 5-10 queries in natural language, without the logical operators AND, OR and NOT, that a user could use to find all and only the relevant documents contained in a fixed collection.

relevance

1.4 Please indicate five documents among those listed in the bibliography of this text that have relevance to the topic described in the question 1.2 based on the title; note that a judgment of relevance with respect to the information requirement is not required.

1.5 Consider the following documents relevant to the topic described in the question 1.2:

1. "Exploiting enriched contextual information for mobile app classification" by Zhu et al. (2012);
2. "Climbing the app wall: enabling mobile app discovery through context-aware recommendations" by Karatzoglou et al. (2012);
3. "In search of the ideal app server" by Haber (1995);
4. "Reflections: In search of the killer app" by Pemberton (2001);
5. "Sharing mobile services: beyond the App store model" by Ahmet and Holmquist (2010).

For each of them, indicate a set of keywords based on the title enclosed in quotation marks.

1.6 Define a POSTING table with document, word, and frequency attributes, and add a tuple for each document and each word of the question 1.5. Also express in SQL the queries for the following information needs:

- "I am looking for information about mobile contextual search";
- "I would like documents relevant to mobile apps yet not relevant to context".

What are the advantages and disadvantages of information retrieval when the data is organized in the indicated way?

1.7 Evaluate the computational cost of the approach to the realization of an IR system based on the table and the queries of the question 1.6 considering, for example, the space occupied, the complexity of the indexes and the queries.

1.8 What are the effects on an index of an increase in the number of documents in a collection and of an increase in the size of the individual documents in a collection?

1.9 Several search engines for the web have a function called "more like this" that suggests other relevant documents. Describe the functions made available by some freely chosen search engines and evaluate their effectiveness based on the number of relevant documents retrieved (for this purpose use the queries of the question 1.6). Also define a measure of precision, keeping in mind that accessing web pages similar to those retrieved interrupts the linearity of the list of results returned in response to a query and induces a non-linear list of accesses.

1.10 Measuring the size of the visible web is one of the daily problems of companies that provide IR services for the web. Given that the enumeration of all pages is not possible, describe a technique for estimating the size of the visible web.

1.11 Formulate two queries in natural language in each of which there is at least one polyseme and two others without words in common, but which express the same information need.

1.12 Compare the IR performed with a search engine with that performed with `grep`.

1.13 Reflect on the problems posed by the diversity of languages and media. What are the possible solutions for the former and the potential offered by the resolution of the latter.

1.14 There are those who believe that the search engines currently available through the web are more than sufficient to satisfy the major information needs of most users. Others, on the contrary, believe that there is a significant proportion of users for whom there are many information needs that cannot be satisfied by the search engines currently available. Write a short report describing the two points of view (it is suggested to work in a group, to form a team for each point of view, to collect the opinions of each team, to change teams, to collect other opinions and, finally, to write the report).

1.15 Choose an application that automatically generates text in response to requests such as the one in the question 1.1. Compare the results provided by the chosen application with those provided by a traditional search engine. Underline the differences and similarities.

2

Representation and Indexing

3

Retrieval and Ranking

4

Measurement and Evaluation

5

Query Expansion and Relevance Feedback

6

Computational Methods

7

Retrieval Models

8

Clustering and Link Analysis

9

Learning and Ranking

10

Language Models

References

Maristella Agosti, Roberto Colotti, and Girolamo Gradenigo. A two-level hypertext retrieval model for legal data. In *SIGIR '91: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 316–325, New York, NY, 1991. ACM. ISBN 0-89791-448-1. doi: http://doi.acm.org/10.1145/122860.122892.

Maristella Agosti, Fabio Crestani, and Massimo Melucci. Design and implementation of a tool for the automatic construction of hypertexts for Information Retrieval. *Information Processing and Management*, 32(4):459–476, July 1996.

Zeynep Ahmet and Lars-Erik Holmquist. Sharing mobile services: beyond the app store model. In *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*, MobileHCI '10, pages 383–384, New York, NY, 2010. ACM. ISBN 978-1-60558-835-3. doi: 10.1145/1851600.1851676. URL `http://doi.acm.org/10.1145/1851600.1851676`.

Omar Alonso and Ricardo Baeza-Yates, editors. *Information Retrieval: Advanced Topics and Techniques*, volume 60. Association for Computing Machinery, New York, NY, USA, 1 edition, 2024. ISBN 9798400710506.

R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison Wesley, New York, II edition, 2010.

Ricardo Baeza-Yates. Bias on the web. *Communication of the ACM*, 61 (6):54–61, 2018.

R.K. Belew. *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW.* Cambridge University Press, December 2000.

N.J. Belkin. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5:133–143, 1980.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

T. Berners-Lee. Information management: A proposal. `http://www.w3.org/History/1989/proposal.html`, 1989.

T. Berners-Lee and R. Caillau. WorldWideWeb: Proposal for a hypertext project. `http://www.w3.org/Proposal.html`, 1989.

T. Berners-Lee, R. Cailliau, A. Luotonen, H.F. Nielsen, and A. Secret. The world-wide web. *Communications of the ACM*, 37(8):76–82, 1994.

T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, pages 29–37, May 2001.

T.J. Berners-Lee, Caillau R., Groff J.F., and es B. Pollermann. *World Wide Web: the Information Universe*, volume 2, pages 52–58. Meckler Publishing, Westport, CT, 1992.

D.C. Blair. *Language and representation in information retrieval.* Elsevier, 1990.

Andrei Broder. A taxonomy of Web search. *SIGIR Forum*, 36:3–10, September 2002. ISSN 0163-5840. doi: http://doi.acm.org/10.1145/792550.792552. URL `http://doi.acm.org/10.1145/792550.792552`.

Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of SIGIR*, pages 33–40, 2000.

V. Bush. As we may think. *Atlantic Monthly*, 176(1):101–108, 1945.

S. Chakrabarti, B.E. Dom, S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the web's link structure. *IEEE Computer*, 32(8):60–67, August 1999.

Peter Christen, David J. Hand, and Nishadi Kirielle. A review of the f-measure: Its history, properties, criticism, and alternatives. *ACM Comput. Surv.*, 56(3), October 2023. ISSN 0360-0300. doi: 10.1145/3606367. URL `https://doi.org/10.1145/3606367`.

Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of SIGIR*, pages 659–666, 2008.

C. Cleverdon, J. Mills, and M. Keen. *ASLIB Cranfield Research Project: factors determining the performance of indexing systems.* ASLIB, 1966.

C.W. Cleverdon and J. Mills. The testing of index language devices. *ASLIB Proceedings*, 15(4):106–130, 1963.

W.B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Addison Wesley, 2009. `http://ciir.cs.umass.edu/downloads/SEIRiP.pdf`.

J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker. Report from the third strategic workshop on information retrieval in lorne (swirl 2018). *SIGIR Forum*, 2018.

D.R. Cutting, D.R. Karger, J.O. Pedersen, and J.W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of SIGIR*, pages 318–329, 1992.

Susan Dumais, Edward Cutrell, JJ Cadiz, Gavin Jancke, Raman Sarin, and Daniel C. Robbins. Stuff I've seen: a system for personal information retrieval and re-use. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '03, pages 72–79, New York, NY, 2003. ACM. ISBN 1-58113-646-3. doi: http://doi.acm.org/10.1145/860435.860451. URL `http://doi.acm.org/10.1145/860435.860451`.

Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. Fairness in information access systems. *Foundations and Trends in Information Retrieval*, 16(1–2):1–177, 2022.

Douglas Engelbart. X-Y position indicator for a display system. Patent 3,541,541, Stanford Research Institute, Menlo Park, CA, June, 21 1967.

Doug Englebart. The augmented knowledge workshop. In *Proceedings of the ACM Conference on The history of personal workstations*, HPW '86, pages 73–83, New York, NY, 1986. ACM. ISBN 0-89791-176-8. doi: 10.1145/12178.12184. URL `http://doi.acm.org/10.1145/12178.12184`.

Christos Faloutsos and King-Ip Lin. Fastmap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. *SIGMOD Rec.*, 24(2):163–174, May 1995. ISSN 0163-5808. doi: 10.1145/568271.223812. URL `http://doi.acm.org/10.1145/568271.223812`.

Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele, and Peter Yanker. Query by image and video content: The QBIC system. *Computer*, 28(9):23–

32, September 1995. ISSN 0018-9162. doi: 10.1109/2.410146. URL `http://dx.doi.org/10.1109/2.410146`.

W.B. Frakes and R. Baeza-Yates, editors. *Information Retrieval: Data Structures and Algorithms.* Prentice Hall, Englewood Cliffs, NJ, 1992.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* MIT Press, 2016. http://www.deeplearningbook.org.

Lynn Haber. In search of the ideal app server. *Datamation*, 41(3): 63–64, February 1995. ISSN 0011-6963. URL `http://dl.acm.org/citation.cfm?id=209401.209411`.

D. Harman. Inverted files. In W.B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: data structures and algorithms.*, chapter 3. Prentice Hall, Englewood Cliffs, NJ, 1992.

Donna Harman. Information retrieval: The early years. *Foundations and Trends in Information Retrieval*, 13(5):425–577, 2019.

M. Hearst. Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

M.A. Hearst and C. Plaunt. Subtopic structuring for full-length document access. In R. Korfhage, E. Rasmussen, and P. Willett, editors, *Proceedings of SIGIR*, pages 59–68, 1993.

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, 1992. Association for Computational Linguistics. doi: 10.3115/992133.992154. URL `http://dx.doi.org/10.3115/992133.992154`.

P. Ingwersen. *Information Retrieval Interaction.* Taylor Graham Publishing, London, UK, UK, 1992.

P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context.* Springer, 2005.

Walter Isaacson. *The Innovators.* Simon & Schuster, 2014.

A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):265–323, September 1999.

N. Jardine and C.J. van Rijsbergen. The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval*, 7:217–240, 1971.

J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 119–126, New York, NY, 2003. ACM. ISBN 1-58113-646-3. doi: 10.1145/860435.860459. URL `http://doi.acm.org/10.1145/860435.860459`.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of SIGKDD*, pages 133–142, 2002.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing.* Prentice Hall, 2024. doi: \url{https://web.stanford.edu/\~jurafsky/slp3/}.

Alexandros Karatzoglou, Linas Baltrunas, Karen Church, and Matthias Böhmer. Climbing the app wall: enabling mobile app discovery through context-aware recommendations. In *Proceedings of CIKM*, CIKM '12, pages 2527–2530, New York, NY, 2012. ACM. ISBN 978-1-4503-1156-4. doi: 10.1145/2396761.2398683. URL `http://doi.acm.org/10.1145/2396761.2398683`.

Brian W. Kernighan. *Understanding the Digital World: What You Need to Know about Computers, the Internet, Privacy, and Security.* Princeton University Press, 2017.

J. Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31, 1968.

H.P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, April 1958.

H.P. Luhn. Keyword-in-context index for technical literature. *American Documentation*, 11(4):288–294, 1960.

B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18 (8):837–842, August 1996. ISSN 0162-8828. doi: 10.1109/34.531803. URL `http://dx.doi.org/10.1109/34.531803`.

C. Manning, P. Raghavan, and H. Schütze. *An Introduction to Information Retrieval.* Cambridge University Press, 2008.

C.D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing.* The MIT Press, 1999.

G. Marchionini and B. Shneiderman. Finding facts vsa. browsing knowledge in hypertext systems. *IEEE Computer*, 21(1):70–80, 1988.

M.E. Maron and J.L. Kuhns. On relevance, probabilistic indexing and retrieval. *Journal of the ACM*, 7:216–244, 1960.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of ICLR Workshop*, 2013a.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey A. Dean. Computing numeric representations of words in a high-dimensional space. `https://patents.google.com/patent/US9037464B1/en`, 2013b.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, 2013c.

Calvin Mooers. Coding, information retrieval, and the rapid selector. *Journal of Documentation*, 1(4):225–229, 1950.

T.H. Nelson. Literary machines: The report on, and of, project XANADU concerning word processing, electronic publishing, hypertext, thinkertoys, tomorrow's intellectual revolution, and certain other topics including knowledge, education and freedom, 1987.

Jian-Yun Nie. *Cross-Language Information Retrieval.* Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2010.

Oxford Dictionary. *The New Oxford American Dictionary*, 2005.

Blaise Pascal. Le provinciali, 1657.

Steven Pemberton. Reflections: In search of the killer app. *Interactions*, 8(4):64, July 2001. ISSN 1072-5520. doi: 10.1145/379537.379564. URL `http://doi.acm.org/10.1145/379537.379564`.

C. Peters and M. Braschler. Cross-Language System Evaluation: the CLEF Campaigns. *Journal of the American Society for Information Science and Technology*, 52(12):1067–1072, 2001.

J.J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, chapter 14, pages 313–323. Prentice-Hall, Englewood Cliffs, NJ, 1971.

G. Salton. Associative document retrieval techniques using bibliographic information. *Journal of the ACM*, 10:440–457, 1963.

G. Salton. *Automatic Information Organization and Retrieval.* Mc Graw Hill, New York, NY, 1968.

G. Salton. *The SMART Retrieval System. Experiments in Automatic Document Processing.* Prentice-Hall, NJ, 1971.

G. Salton. *Automatic Text Processing.* Addison-Wesley, 1989.

G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York, NY, 1983.

G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of SIGIR*, pages 49–58, 1993.

Mark Sanderson and Justin Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of SIGIR*, pages 162–169, 2005.

Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, September 1999. ISSN 0163-5840. doi: 10.1145/331403.331405. URL `http://doi.acm.org/10.1145/331403.331405`.

K. Sparck Jones. *Automatic Keyword Classification*. Butterworths, 1971.

K. Sparck Jones and C.J. van Rijsbergen. Information retrieval test collections. *Journal of Documentation*, 32(1):59–75, March 1976.

K. Sparck Jones and P. Willett. *Readings in Information Retrieval*. Morgan Kaufmann, San Francisco, CA, 1997.

Cornelis Joost Van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.

Cornelis Joost Van Rijsbergen. *The Geometry of Information Retrieval*. Cambridge University Press, UK, 2004.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking, part i: Score-based ranking. *ACM Comput. Surv.*, 55(6), December 2022a. ISSN 0360-0300. doi: 10.1145/3533379. URL `https://doi.org/10.1145/3533379`.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking, part ii: Learning-to-rank and recommender systems. *ACM Comput. Surv.*, 55(6), December 2022b. ISSN 0360-0300. doi: 10.1145/3533380. URL `https://doi.org/10.1145/3533380`.

Hengshu Zhu, Huanhuan Cao, Enhong Chen, Hui Xiong, and Jilei Tian. Exploiting enriched contextual information for mobile app classification. In *Proceedings of CIKM*, CIKM '12, pages 1617–1621, New York, NY, 2012. ACM. ISBN 978-1-4503-1156-4. doi: 10.1145/2396761.2398484. URL `http://doi.acm.org/10.1145/2396761.2398484`.

Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of SIGIR*, pages 307–314, 1998.

Index