WIKIPEDIA
25 years of the free encyclopedia

Information retrieval

Information retrieval (**IR**) in computing and information science is the task of identifying and retrieving information system resources that are relevant to an information need. The information need can be specified in the form of a search query. In the case of document retrieval, queries can be based on full-text or other content-based indexing. Information retrieval is the science[1] of searching for information in a document, searching for documents themselves, and also searching for the metadata that describes data, and for databases of texts, images or sounds. Cross-modal retrieval implies retrieval across modalities.

Automated information retrieval systems are used to reduce what has been called information overload. An IR system is a software system that provides access to books, journals and other documents; it also stores and manages those documents. Web search engines are the most visible IR applications.

Overview

An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval, a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevance.

An object is an entity that is represented by information in a content collection or database. User queries are matched against the database information. However, as opposed to classical SQL queries of a database, in information retrieval the results returned may or may not match the query, so results are typically ranked. This ranking of results is a key difference of information retrieval searching compared to database searching.[2]

Depending on the application the data objects may be, for example, text documents, images,[3] audio,[4] mind maps[5] or videos. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates or metadata.

Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query.[6]

History

The idea of using computers to search for relevant pieces of information was popularized in the article *As We May Think* by Vannevar Bush in 1945.[7] It would appear that Bush was inspired by patents for a 'statistical machine' – filed by Emanuel Goldberg in the 1920s and 1930s – that searched for documents stored on film.[8] The first description of a computer searching

> there is ... a machine called the Univac ... whereby letters and figures are coded as a pattern of magnetic spots on a long steel tape. By this means the text of a document, preceded

for information was described by Holmstrom in 1948,[9] detailing an early mention of the Univac computer. Automated information retrieval systems were introduced in the 1950s: one even featured in the 1957 romantic comedy *Desk Set*. In the 1960s, the first large information retrieval research group was formed by Gerard Salton at Cornell. By the 1970s several different retrieval techniques had been shown to perform well on small text corpora such as the Cranfield collection (several thousand documents).[7] Large-scale retrieval systems, such as the Lockheed Dialog system, came into use early in the 1970s.

> by its subject code symbol, can be recorded ... the machine ... automatically selects and types out those references which have been coded in any desired way at a rate of 120 words a minute
>
> —J. E. Holmstrom, 1948

In 1992, the US Department of Defense along with the National Institute of Standards and Technology (NIST), cosponsored the Text Retrieval Conference (TREC) as part of the TIPSTER text program. The aim of this was to look into the information retrieval community by supplying the infrastructure that was needed for evaluation of text retrieval methodologies on a very large text collection. This catalyzed research on methods that scale to huge corpora. The introduction of web search engines has boosted the need for very large scale retrieval systems even further.

By the late 1990s, the rise of the World Wide Web fundamentally transformed information retrieval. While early search engines such as AltaVista (1995) and Yahoo! (1994) offered keyword-based retrieval, they were limited in scale and ranking refinement. The breakthrough came in 1998 with the founding of Google, which introduced the PageRank algorithm,[10] using the web's hyperlink structure to assess page importance and improve relevance ranking.

During the 2000s, web search systems evolved rapidly with the integration of machine learning techniques. These systems began to incorporate user behavior data (e.g., click-through logs), query reformulation, and content-based signals to improve search accuracy and personalization. In 2009, Microsoft launched Bing, introducing features that would later incorporate semantic web technologies through the development of its Satori knowledge base. Academic analysis[11] have highlighted Bing's semantic capabilities, including structured data use and entity recognition, as part of a broader industry shift toward improving search relevance and understanding user intent through natural language processing.

A major leap occurred in 2018, when Google deployed BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) to better understand the contextual meaning of queries and documents. This marked one of the first times deep neural language models were used at scale in real-world retrieval systems.[12] BERT's bidirectional training enabled a more refined comprehension of word relationships in context, improving the handling of natural language queries. Because of its success, transformer-based models gained traction in academic research and commercial search applications.[13]

Simultaneously, the research community began exploring neural ranking models that outperformed traditional lexical-based methods. Long-standing benchmarks such as the **T**ext **RE**trieval **C**onference (TREC), initiated in 1992, and more recent evaluation frameworks Microsoft MARCO(**MA**chine **R**eading **CO**mprehension) (2019)[14] became central to training and evaluating retrieval systems across multiple tasks and domains. MS MARCO has also been adopted in the TREC Deep Learning Tracks, where it serves as a core dataset for evaluating advances in neural ranking models within a standardized benchmarking environment.[15]

As deep learning became integral to information retrieval systems, researchers began to categorize neural approaches into three broad classes: **sparse**, **dense**, and **hybrid** models. Sparse models, including traditional term-based methods and learned variants like SPLADE, rely on interpretable representations and inverted indexes to enable efficient exact term matching with added semantic signals.[16] Dense models, such as dual-encoder architectures like ColBERT, use continuous vector embeddings to support semantic similarity beyond keyword overlap.[17] Hybrid models aim to combine the advantages of both, balancing the lexical (token) precision of sparse methods with the semantic depth of dense models. This way of categorizing models balances scalability, relevance, and efficiency in retrieval systems.[18]

As IR systems increasingly rely on deep learning, concerns around bias, fairness, and explainability have also come to the picture. Research is now focused not just on relevance and efficiency, but on transparency, accountability, and user trust in retrieval algorithms.

Applications

Areas where information retrieval techniques are employed include (the entries are in alphabetical order within each category):

General applications

- Digital libraries
- Information filtering
  - Recommender systems
- Media search
  - Blog search
  - Image retrieval
  - 3D retrieval
  - Music retrieval
  - News search
  - Speech retrieval
  - Video retrieval
- Search engines
  - Site search
  - Desktop search
  - Enterprise search
  - Federated search
  - Mobile search
  - Social search
  - Web search

Domain-specific applications

- Expert search finding
- Genomic information retrieval

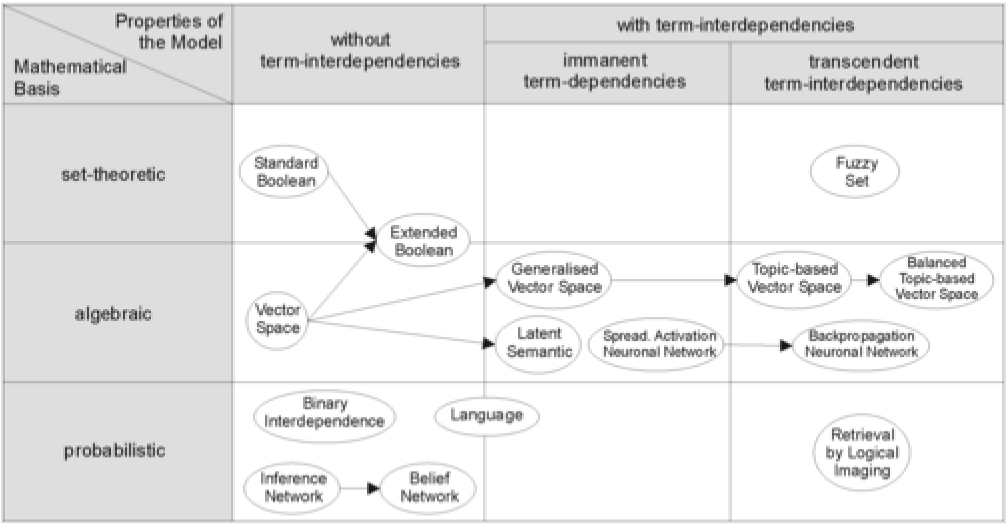
- Geographic information retrieval
- Information retrieval for chemical structures
- Information retrieval in software engineering
- Legal information retrieval
- Vertical search

Other retrieval methods

Methods/Techniques in which information retrieval techniques are employed include:

- Cross-modal retrieval
- Adversarial information retrieval
- Automatic summarization
    - Multi-document summarization
- Compound term processing
- Cross-lingual retrieval
- Document classification
- Spam filtering
- Question answering

Model types



Properties of the Model Mathematical Basis	without term-interdependencies	with term-interdependencies	
		immanent term-dependencies	transcendent term-interdependencies
set-theoretic	Standard Boolean Extended Boolean		Fuzzy Set
algebraic	Vector Space	Generalised Vector Space Latent Semantic Spread. Activation Neuronal Network	Topic-based Vector Space Balanced Topic-based Vector Space Backpropagation Neuronal Network
probabilistic	Binary Interdependence Inference Network Belief Network	Language	Retrieval by Logical Imaging

Categorization of IR-models (translated from German entry, original source Dominik Kuropka (http://www.logos-verlag.de/cgi-bin/engbuchmid?isbn=0514&lng=eng&id=))

In order to effectively retrieve relevant documents by IR strategies, the documents are typically transformed into a suitable representation. Each retrieval strategy incorporates a specific model for its document representation purposes. The picture on the right illustrates the relationship of some common models. In the picture, the models are categorized according to two dimensions: the mathematical basis and the properties of the model.

First dimension: mathematical basis

- *Set-theoretic* models represent documents as sets of words or phrases. Similarities are usually derived from set-theoretic operations on those sets. Common models are:
  - Standard Boolean model
  - Extended Boolean model
  - Fuzzy retrieval
- *Algebraic models* represent documents and queries usually as vectors, matrices, or tuples. The similarity of the query vector and document vector is represented as a scalar value.
  - Vector space model
  - Generalized vector space model
  - (Enhanced) Topic-based Vector Space Model
  - Extended Boolean model
  - Latent semantic indexing a.k.a. latent semantic analysis
- *Probabilistic models* treat the process of document retrieval as a probabilistic inference. Similarities are computed as probabilities that a document is relevant for a given query. Probabilistic theorems like Bayes' theorem are often used in these models.
  - Binary Independence Model
  - Probabilistic relevance model on which is based the okapi (BM25) relevance function
  - Uncertain inference
  - Language models
  - Divergence-from-randomness model
  - Latent Dirichlet allocation
- *Feature-based retrieval models* view documents as vectors of values of *feature functions* (or just *features*) and seek the best way to combine these features into a single relevance score, typically by learning to rank methods. Feature functions are arbitrary functions of document and query, and as such can easily incorporate almost any other retrieval model as just another feature.
- *Data fusion models:* Data fusion in information retrieval combines results from multiple search systems or retrieval models to improve performance. By merging ranked lists, it leverages the strengths of diverse approaches, often enhancing recall and precision. Common methods include score normalization and voting techniques like CombSUM or Borda count. This meta-search strategy is particularly effective when individual systems have complementary coverage or when query difficulty varies, producing a more robust and reliable final ranking sparsity.[19][20]

Second dimension: properties of the model

- *Models without term-interdependencies* treat different terms/words as independent. This fact is usually represented in vector space models by the orthogonality assumption of term vectors or in probabilistic models by an independency assumption for term variables.
- *Models with immanent term interdependencies* allow a representation of interdependencies between terms. However the degree of the interdependency between two terms is defined by the model itself. It is usually directly or indirectly derived (e.g. by dimensional reduction) from the co-occurrence of those terms in the whole set of documents.
- *Models with transcendent term interdependencies* allow a representation of interdependencies between terms, but they do not allege how the interdependency between

two terms is defined. They rely on an external source for the degree of interdependency between two terms. (For example, a human or sophisticated algorithms.)

Third Dimension: representational approach-based classification

In addition to the theoretical distinctions, modern information retrieval models are also categorized on how queries and documents are represented and compared, using a practical classification distinguishing between sparse, dense and hybrid models.[16]

- ***Sparse*** models utilize interpretable, term-based representations and typically rely on inverted index structures. Classical methods such as TF-IDF and BM25 fall under this category, along with more recent learned sparse models that integrate neural architectures while retaining sparsity.[21]
- ***Dense*** models represent queries and documents as continuous vectors using deep learning models, typically transformer-based encoders. These models enable semantic similarity matching beyond exact term overlap and are used in tasks involving semantic search and question answering.[22]
- ***Hybrid*** models aim to combine the strengths of both approaches, integrating lexical (tokens) and semantic signals through score fusion, late interaction, or multi-stage ranking pipelines.[23]

This classification has become increasingly common in both academic and the real world applications and is getting widely adopted and used in evaluation benchmarks for Information Retrieval models.[18][21]

Performance and correctness measures

The evaluation of an information retrieval system' is the process of assessing how well a system meets the information needs of its users. In general, measurement considers a collection of documents to be searched and a search query. Traditional evaluation metrics, designed for Boolean retrieval or top-k retrieval, include precision and recall. All measures assume a ground truth notion of relevance: every document is known to be either relevant or non-relevant to a particular query. In practice, queries may be ill-posed and there may be different shades of relevance.

Libraries for searching and indexing

- Lemur
- Lucene
  - Solr
  - Elasticsearch
- Manatee
- Manticore search
- Sphinx
- Terrier Search Engine
- Xapian

Timeline

- Before the **1900s**

  1801: Joseph Marie Jacquard invents the Jacquard loom, the first machine to use punched cards to control a sequence of operations.
  1880s: Herman Hollerith invents an electro-mechanical data tabulator using punch cards as a machine readable medium.
  1890 Hollerith cards, keypunches and tabulators used to process the 1890 US census data.

- **1920s–1930s**

  Emanuel Goldberg submits patents for his "Statistical Machine", a document search engine that used photoelectric cells and pattern recognition to search the metadata on rolls of microfilmed documents.

- **1940s–1950s**

  late 1940s: The US military confronted problems of indexing and retrieval of wartime scientific research documents captured from Germans.

  > **1945**: Vannevar Bush's *As We May Think* appeared in *Atlantic Monthly*.
  > **1947**: Hans Peter Luhn (research engineer at IBM since 1941) began work on a mechanized punch card-based system for searching chemical compounds.

  1950s: Growing concern in the US for a "science gap" with the USSR motivated, encouraged funding and provided a backdrop for mechanized literature searching systems (Allen Kent *et al.*) and the invention of the citation index by Eugene Garfield.
  1950: The term "information retrieval" was coined by Calvin Mooers.[24]
  1951: Philip Bagley conducted the earliest experiment in computerized document retrieval in a master thesis at MIT.[25]
  1955: Allen Kent joined Case Western Reserve University, and eventually became associate director of the Center for Documentation and Communications Research. That same year, Kent and colleagues published a paper in American Documentation describing the precision and recall measures as well as detailing a proposed "framework" for evaluating an IR system which included statistical sampling methods for determining the number of relevant documents not retrieved.[26]
  1958: International Conference on Scientific Information Washington DC included consideration of IR systems as a solution to problems identified. See: *Proceedings of the International Conference on Scientific Information, 1958* (National Academy of Sciences, Washington, DC, 1959)
  1959: Hans Peter Luhn published "Auto-encoding of documents for information retrieval".

- **1960s**:

  early 1960s: Gerard Salton began work on IR at Harvard, later moved to Cornell.
  1960: Melvin Earl Maron and John Lary Kuhns[27] published "On relevance, probabilistic indexing, and information retrieval" in the Journal of the ACM 7(3):216–244, July 1960.
  1962:
  * Cyril W. Cleverdon published early findings of the Cranfield studies, developing a model for IR system evaluation. See: Cyril W. Cleverdon, "Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems". Cranfield Collection of Aeronautics, Cranfield, England, 1962.
  * Kent published *Information Analysis and Retrieval*.
  1963:
  * Weinberg report "Science, Government and Information" gave a full articulation of the idea of a "crisis of scientific information". The report was named after Dr. Alvin Weinberg.

* Joseph Becker and Robert M. Hayes published text on information retrieval. Becker, Joseph; Hayes, Robert Mayo. *Information storage and retrieval: tools, elements, theories*. New York, Wiley (1963).
1964:
* Karen Spärck Jones finished her thesis at Cambridge, *Synonymy and Semantic Classification*, and continued work on computational linguistics as it applies to IR.
* The National Bureau of Standards sponsored a symposium titled "Statistical Association Methods for Mechanized Documentation". Several highly significant papers, including G. Salton's first published reference (we believe) to the SMART system.
mid-1960s:

* National Library of Medicine (NLM) developed MEDLARS Medical Literature Analysis and Retrieval System, the first major machine-readable database and batch-retrieval system.
* Project Intrex at MIT.
1965: J. C. R. Licklider published *Libraries of the Future*.
1966: Don Swanson was involved in studies at University of Chicago on Requirements for Future Catalogs.

late 1960s: F. Wilfrid Lancaster completed evaluation studies of the MEDLARS system and published the first edition of his text on information retrieval.

1968:

* Gerard Salton published *Automatic Information Organization and Retrieval*.
* John W. Sammon, Jr.'s RADC Tech report "Some Mathematics of Information Storage and Retrieval..." outlined the vector model.

1969: Sammon's "A nonlinear mapping for data structure analysis (http://studentnet.cs.manchester.ac.uk/pgt/COMP61021/reference/Sammon.pdf) Archived (https://web.archive.org/web/20170808172524/http://studentnet.cs.manchester.ac.uk/pgt/COMP61021/reference/Sammon.pdf) 2017-08-08 at the Wayback Machine" (IEEE Transactions on Computers) was the first proposal for visualization interface to an IR system.

- **1970s**

early 1970s:

* First online systems—NLM's AIM-TWX, MEDLINE; Lockheed's Dialog; SDC's ORBIT.
* Theodor Nelson promoting concept of hypertext, published *Computer Lib/Dream Machines*.

1971: Nicholas Jardine and Cornelis J. van Rijsbergen published "The use of hierarchic clustering in information retrieval", which articulated the "cluster hypothesis".[28]
1975: Three highly influential publications by Salton fully articulated his vector processing framework and term discrimination model:

* *A Theory of Indexing* (Society for Industrial and Applied Mathematics)
* *A Theory of Term Importance in Automatic Text Analysis* (JASIS v. 26)
* *A Vector Space Model for Automatic Indexing* (CACM 18:11)

1978: The First ACM SIGIR conference.
1979: C. J. van Rijsbergen published *Information Retrieval* (Butterworths). Heavy emphasis on probabilistic models.
1979: Tamas Doszkocs implemented the CITE natural language user interface for MEDLINE at the National Library of Medicine. The CITE system supported free form query input, ranked output and relevance feedback.[29]

- **1980s**

1980: First international ACM SIGIR conference, joint with British Computer Society IR group in Cambridge.
1982: Nicholas J. Belkin, Robert N. Oddy, and Helen M. Brooks proposed the ASK (Anomalous State of Knowledge) viewpoint for information retrieval. This was an

important concept, though their automated analysis tool proved ultimately disappointing.

1983: Salton (and Michael J. McGill) published *Introduction to Modern Information Retrieval* (McGraw-Hill), with heavy emphasis on vector models.

1985: David Blair and Bill Maron publish: *An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System*.[30]

1986: Donald A.B. Lindberg M.D., NLM Director, implemented a direct health professional interface to MEDLINE and other MEDLARS databases. Grateful Med, a pun on the Grateful Dead, was adapted from Microsearch, an ELHILL user interface that assembled query language prior to connecting to the NLM mainframe. After retrieving the query results, Grateful Med disconnected from the mainframe to keep search costs low.[31]

mid-1980s: Efforts to develop end-user versions of commercial IR systems.

> **1985–1993**: Key papers on and experimental systems for visualization interfaces.
>
> Work by Donald B. Crouch, Robert R. Korfhage, Matthew Chalmers, Anselm Spoerri and others.

1989: First World Wide Web proposals by Tim Berners-Lee at CERN.

- **1990s**

1992: First TREC conference.

1997: Publication of Korfhage's *Information Storage and Retrieval*[32] with emphasis on visualization and multi-reference point systems.

1998: Google is founded by Larry Page and Sergey Brin. It introduces the PageRank algorithm, which evaluates the importance of web pages based on hyperlink structure.[10]

1999: Publication of Ricardo Baeza-Yates and Berthier Ribeiro-Neto's *Modern Information Retrieval* by Addison Wesley, the first book that attempts to cover all IR.

- **2000s**

2001: Wikipedia launches as a free, collaborative online encyclopedia. It quickly becomes a major resource for information retrieval, particularly for natural language processing and semantic search benchmarks.[33]

2009: Microsoft launches Bing, introducing features such as related searches, semantic suggestions, and later incorporating deep learning techniques into its ranking algorithms.[11]

- **2010s**

2013: Google's Hummingbird algorithm goes live, marking a shift from keyword matching toward understanding query intent and semantic context in search queries.[34]

2018: Google AI researchers release BERT (Bidirectional Encoder Representations from Transformers), enabling deep bidirectional understanding of language and improving document ranking and query understanding in IR.[12]

2019: Microsoft introduces MS MARCO (Microsoft **MA**chine **R**eading **CO**mprehension), a large-scale dataset designed for training and evaluating machine reading and passage ranking models.[14]

- **2020s**

2020: The **ColBERT** (Contextualized Late Interaction over BERT) model, designed for efficient passage retrieval using contextualized embeddings, was introduced at SIGIR 2020.[35][17]

2021: SPLADE is introduced at SIGIR 2021. It's a sparse neural retrieval model that balances lexical and semantic features using masked language modeling and sparsity regularization.[36]

2022: The **BEIR** benchmark is released to evaluate zero-shot IR across 18 datasets covering diverse tasks. It standardizes comparisons between dense, sparse, and hybrid IR models.[21]

Major conferences

- SIGIR: Special Interest Group on Information Retrieval
- ECIR: European Conference on Information Retrieval
- CIKM: Conference on Information and Knowledge Management
- WWW: International World Wide Web Conference

Awards in the field

- Tony Kent Strix award
- Gerard Salton Award
- Karen Spärck Jones Award

See also

- Adversarial information retrieval – Information retrieval strategies in datasets
- Computer memory – Component that stores information
- Controlled vocabulary – Method of organizing knowledge
- Cross-language information retrieval
- Data mining – Process of extracting and discovering patterns in large data sets
- Data retrieval – Way to obtain data from a database
- Human–computer information retrieval (HCIR)
- Information extraction – Machine reading of unstructured documents
- Information seeking – Type of activity in information science
  - Information seeking § Compared to information retrieval
  - Collaborative information seeking
  - Social information seeking
- Information Retrieval Facility – Organization in Vienna, Austria 2006–2012
- Knowledge visualization – Set of techniques for creating images, diagrams, or animations to communicate a message
- Multimedia information retrieval
- Personal information management – Tools and systems for managing one's own data
- Pearl growing – Type of search strategy
- Query understanding – Search engine processing step
- Relevance (information retrieval) – Measure of a document's applicability to a given subject or search query
- Relevance feedback – Data used in information retrieval and recommendation systems
- Retrievability – Property of being able to access something
- Rocchio classification – Classification model in machine learning
- Search engine indexing – Method for data management
- Special Interest Group on Information Retrieval – Subgroup of the Association for Computing Machinery
- Subject indexing – Classifying a document by index terms

- Temporal information retrieval – Area of research related to information retrieval centered on timeliness
- tf–idf – Estimate of the importance of a word in a document
- XML retrieval – Content-based retrieval of XML documents
- Web mining – Process of extracting and discovering patterns in large data sets

References

1. Luk, R. W. P. (2022). "Why is information retrieval a scientific discipline?". *Foundations of Science*. **27** (2): 427–453. doi:10.1007/s10699-020-09685-x (https://doi.org/10.1007%2Fs10699-020-09685-x). hdl:10397/94873 (https://hdl.handle.net/10397%2F94873). S2CID 220506422 (https://api.semanticscholar.org/CorpusID:220506422).
2. Jansen, B.J.; Rieh, S. (2010). "The Seventeen Theoretical Constructs of Information Searching and Information Retrieval" (http://www.bernardjjansen.com/uploads/2/4/1/8/24188166/jansen_theoretical_constructs.pdf) (PDF). *Journal of the American Society for Information Sciences and Technology*. **61** (8): 1517–34. doi:10.1002/asi.21358 (https://doi.org/10.1002%2Fasi.21358).
3. Goodrum, Abby A. (2000). "Image Information Retrieval: An Overview of Current Research" (https://inform.nu/Articles/Vol3/v3n2p63-66.pdf) (PDF). *Informing Science*. **3** (2): 063–066. doi:10.28945/578 (https://doi.org/10.28945%2F578).
4. Foote, Jonathan (1999). "An overview of audio information retrieval". *Multimedia Systems*. **7**: 2–10. CiteSeerX 10.1.1.39.6339 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.39.6339). doi:10.1007/s005300050106 (https://doi.org/10.1007%2Fs005300050106). S2CID 2000641 (https://api.semanticscholar.org/CorpusID:2000641).
5. Beel, Jöran; Gipp, Bela; Stiller, Jan-Olaf (2009). *Information Retrieval On Mind Maps — What Could It Be Good For?* (https://fossies.org/linux/docear/docear_plugin_core/resources/demo/docear_example_pdfs/Information%20Retrieval%20on%20Mind%20Maps%20--%20What%20could%20it%20be%20good%20for.pdf) (PDF). Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom'09). IEEE. doi:10.4108/ICST.COLLABORATECOM2009.8298 (https://doi.org/10.4108%2FICST.COLLABORATECOM2009.8298). ISBN 978-963-9799-76-9.
6. Frakes, William B.; Baeza-Yates, Ricardo (1992). *Information Retrieval Data Structures & Algorithms* (https://web.archive.org/web/20130928060217/http://www.scribd.com/doc/13742235/Information-Retrieval-Data-Structures-Algorithms-William-B-Frakes). Prentice-Hall, Inc. ISBN 978-0-13-463837-9. Archived from the original (https://www.scribd.com/doc/13742235/Information-Retrieval-Data-Structures-Algorithms-William-B-Frakes) on 2013-09-28.
7. Singhal, Amit (2001). "Modern Information Retrieval: A Brief Overview" (http://singhal.info/ieee2001.pdf) (PDF). *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*. **24** (4): 35–43.
8. Mark Sanderson & W. Bruce Croft (2012). "The History of Information Retrieval Research" (https://doi.org/10.1109%2Fjproc.2012.2189916). *Proceedings of the IEEE*. **100**: 1444–51. doi:10.1109/jproc.2012.2189916 (https://doi.org/10.1109%2Fjproc.2012.2189916).
9. JE Holmstrom (1948). " 'Section III. Opening Plenary Session" (https://books.google.com/books?id=M34lAAAAMAAJ&q=univac). *The Royal Society Scientific Information Conference, 21 June-2 July 1948: Report and Papers Submitted*: 85.
10. Brin, Sergey; Page, Lawrence (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine" (http://ilpubs.stanford.edu:8090/361/1/1998-8.pdf) (PDF). *Computer Networks and ISDN Systems*. **30** (1–7): 107–117. doi:10.1016/S0169-7552(98)00110-X (https://doi.org/10.1016%2FS0169-7552%2898%2900110-X).

11. Uyar, Ahmet; Aliyu, Farouk Musa (2015-01-01). "Evaluating search features of Google Knowledge Graph and Bing Satori: Entity types, list searches and query interfaces" (https://www.emerald.com/insight/content/doi/10.1108/oir-10-2014-0257/full/html). *Online Information Review*. **39** (2): 197–213. doi:10.1108/OIR-10-2014-0257 (https://doi.org/10.1108%2FOIR-10-2014-0257). ISSN 1468-4527 (https://search.worldcat.org/issn/1468-4527).
12. Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". arXiv:1810.04805 (https://arxiv.org/abs/1810.04805) [cs.CL (https://arxiv.org/archive/cs.CL)].
13. Gardazi, Nadia Mushtaq; Daud, Ali; Malik, Muhammad Kamran; Bukhari, Amal; Alsahfi, Tariq; Alshemaimri, Bader (2025-03-15). "BERT applications in natural language processing: a review" (https://doi.org/10.1007%2Fs10462-025-11162-5). *Artificial Intelligence Review*. **58** (6): 166. doi:10.1007/s10462-025-11162-5 (https://doi.org/10.1007%2Fs10462-025-11162-5). ISSN 1573-7462 (https://search.worldcat.org/issn/1573-7462).
14. Bajaj, Payal; Campos, Daniel; Craswell, Nick; Deng, Li; Gao, Jianfeng; Liu, Xiaodong; Majumder, Rangan; McNamara, Andrew; Mitra, Bhaskar; Nguyen, Tri; Rosenberg, Mir; Song, Xia; Stoica, Alina; Tiwary, Saurabh; Wang, Tong (2016). "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset". arXiv:1611.09268 (https://arxiv.org/abs/1611.09268) [cs.CL (https://arxiv.org/archive/cs.CL)].
15. Craswell, Nick; Mitra, Bhaskar; Yilmaz, Emine; Rahmani, Hossein A.; Campos, Daniel; Lin, Jimmy; Voorhees, Ellen M.; Soboroff, Ian (February 2024). "Overview of the TREC 2023 Deep Learning Track" (https://www.microsoft.com/en-us/research/publication/overview-of-the-trec-2023-deep-learning-track/). *Text REtrieval Conference (TREC)* – via Microsoft.
16. Kim, Dohyun; Zhao, Lina; Chung, Eric; Park, Eun-Jae (2021). "Pressure-robust staggered DG methods for the Navier-Stokes equations on general meshes". arXiv:2107.09226 (https://arxiv.org/abs/2107.09226) [math.NA (https://arxiv.org/archive/math.NA)].
17. Khattab, Omar; Zaharia, Matei (2020-07-25). "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT" (https://dl.acm.org/doi/10.1145/3397271.3401075). *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '20. New York, NY, USA: Association for Computing Machinery. pp. 39–48. doi:10.1145/3397271.3401075 (https://doi.org/10.1145%2F3397271.3401075). ISBN 978-1-4503-8016-4.
18. Lin, Jimmy; Nogueira, Rodrigo; Yates, Andrew (2020). "Pretrained Transformers for Text Ranking: BERT and Beyond". arXiv:2010.06467 (https://arxiv.org/abs/2010.06467) [cs.IR (https://arxiv.org/archive/cs.IR)].
19. Shaw, Joseph A; Fox, Edward A. (1994) Combination of Multiple Searches. TREC 1994: 105-108
20. Wu, Shengli (2012). *Data Fusion in Information Retrieval*. Springer. pp. 1–212. ISBN 978-3-642-28865-4.
21. Thakur, Nandan; Reimers, Nils; Rücklé, Andreas; Srivastava, Abhishek; Gurevych, Iryna (2021). "BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models". arXiv:2104.08663 (https://arxiv.org/abs/2104.08663) [cs.IR (https://arxiv.org/archive/cs.IR)].
22. Lau, Jey Han; Armendariz, Carlos; Lappin, Shalom; Purver, Matthew; Shu, Chang (2020). Johnson, Mark; Roark, Brian; Nenkova, Ani (eds.). "How Furiously Can Colorless Green Ideas Sleep? Sentence Acceptability in Context" (https://aclanthology.org/2020.tacl-1.20/). *Transactions of the Association for Computational Linguistics*. **8**: 296–310. doi:10.1162/tacl_a_00315 (https://doi.org/10.1162%2Ftacl_a_00315).
23. Arabzadeh, Negar; Yan, Xinyi; Clarke, Charles L. A. (2021). "Predicting Efficiency/Effectiveness Trade-offs for Dense vs. Sparse Retrieval Strategy Selection". arXiv:2109.10739 (https://arxiv.org/abs/2109.10739) [cs.IR (https://arxiv.org/archive/cs.IR)].

24. Mooers, Calvin N. (1951). *The Theory of Digital Handling of Non-numerical Information and its Implications to Machine Economics* (https://babel.hathitrust.org/cgi/pt?id=mdp.39015034570591;view=1up;seq=3). *Zator Technical Bulletin* (Technical report). 48., cited in Fairthorne, R. A. (1958). "Automatic Retrieval of Recorded Information" (https://doi.org/10.1093%2Fcomjnl%2F1.1.36). *The Computer Journal*. **1** (1): 37. doi:10.1093/comjnl/1.1.36 (https://doi.org/10.1093%2Fcomjnl%2F1.1.36).
25. Doyle, Lauren; Becker, Joseph (1975). *Information Retrieval and Processing*. Melville. ISBN 978-0-471-22151-7.
26. Perry, James W.; Kent, Allen; Berry, Madeline M. (1955). "Machine literature searching X. Machine language; factors underlying its design and development". *American Documentation*. **6** (4): 242–254. doi:10.1002/asi.5090060411 (https://doi.org/10.1002%2Fasi.5090060411).
27. Maron, Melvin E. (2008). "An Historical Note on the Origins of Probabilistic Indexing" (http://yunus.hacettepe.edu.tr/~tonta/courses/spring2008/bby703/maron-on-probabilistic%20indexing-2008.pdf) (PDF). *Information Processing and Management*. **44** (2): 971–2. doi:10.1016/j.ipm.2007.02.012 (https://doi.org/10.1016%2Fj.ipm.2007.02.012).
28. N. Jardine, C.J. van Rijsbergen (December 1971). "The use of hierarchic clustering in information retrieval". *Information Storage and Retrieval*. **7** (5): 217–240. doi:10.1016/0020-0271(71)90051-9 (https://doi.org/10.1016%2F0020-0271%2871%2990051-9).
29. Doszkocs, T.E.; Rapp, B.A. (1979). "Searching MEDLINE in English: a Prototype User Interface with Natural Language Query, Ranked Output, and relevance feedback" (https://www.osti.gov/biblio/5047496). *Information choices and policies : 42nd annual meeting, Minneapolis, Minnesota, October 14-18, 1979*. American Society for Information Science. Vol. 16. Knowledge Industry Publications. pp. 131–9. OCLC 271407392 (https://search.worldcat.org/oclc/271407392). OSTI 5047496 (https://www.osti.gov/biblio/5047496).
30. Blair, D.C.; Maron, M.E. (1985). "An evaluation of retrieval effectiveness for a full-text document-retrieval system". *Communications of the ACM*. **28** (3): 289–299. doi:10.1145/3166.3197 (https://doi.org/10.1145%2F3166.3197).
31. Dorsch JL, Faughnan JG, Humphreys BL (June 2022). "Grateful Med: Direct access to MEDLINE for health professionals with personal computers" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9196098). *Inf Serv Use*. **42** (2): 151–160. doi:10.3233/ISU-220147 (https://doi.org/10.3233%2FISU-220147). PMC 9196098 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9196098). PMID 35720429 (https://pubmed.ncbi.nlm.nih.gov/35720429).
32. Korfhage, Robert R. (1997). *Information Storage and Retrieval* (https://archive.org/details/informationstora00korf/page/368). Wiley. pp. 368 pp (https://archive.org/details/informationstora00korf/page/368). ISBN 978-0-471-14338-3.
33. "History of Wikipedia" (https://en.wikipedia.org/wiki/History_of_Wikipedia), *Wikipedia*, 2025-02-21, retrieved 2025-04-09
34. Sullivan, Danny (2013-09-26). "FAQ: All About The New Google "Hummingbird" Algorithm" (https://searchengineland.com/google-hummingbird-172816). *Search Engine Land*. Retrieved 2025-04-09.
35. Khattab, Omar; Zaharia, Matei (2020). "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT". arXiv:2004.12832 (https://arxiv.org/abs/2004.12832) [cs.IR (https://arxiv.org/archive/cs.IR)].

36. Jones, Rosie; Zamani, Hamed; Schedl, Markus; Chen, Ching-Wei; Reddy, Sravana; Clifton, Ann; Karlgren, Jussi; Hashemi, Helia; Pappu, Aasish; Nazari, Zahra; Yang, Longqi; Semerci, Oguz; Bouchard, Hugues; Carterette, Ben (2021-07-11). "Current Challenges and Future Directions in Podcast Information Access" (https://dl.acm.org/doi/10.1145/3404835.3462805). *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '21. New York, NY, USA: Association for Computing Machinery. pp. 1554–65. arXiv:2106.09227 (https://arxiv.org/abs/2106.09227). doi:10.1145/3404835.3462805 (https://doi.org/10.1145%2F3404835.3462805). ISBN 978-1-4503-8037-9.

Further reading

- Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier (2011). *Modern Information Retrieval: The Concepts and Technology behind Search* (https://users.dcc.uchile.cl/~rbaeza/mir2ed/contents.php.html) (2nd ed.). Addison-Wesley. ISBN 978-0-321-41691-9.
- Büttcher, Stefan; Clarke, Charles L.A.; Cormack, Gordon V. (2010). *Information Retrieval: Implementing and Evaluating Search Engines* (https://books.google.com/books?id=epD6AQAAQBAJ&pg=PR7). MIT Press. ISBN 978-0-262-02651-2. OCLC 473652398 (https://search.worldcat.org/oclc/473652398).
- "Information Retrieval System" (https://web.archive.org/web/20200511161049/http://www.lisbdnet.com/information-retrieval-syste). *Library & Information Science Network*. 24 April 2015. Archived from the original (http://www.lisbdnet.com/information-retrieval-syste/) on 11 May 2020. Retrieved 3 May 2020.
- Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich (2008). *Introduction to Information Retrieval* (https://nlp.stanford.edu/IR-book/). Cambridge University Press. ISBN 978-0-521-86571-5.
- ShinJoung, Yeo (2023). *Behind the Search Box: Google and the Global Internet Industry*. University of Illinois Press. ISBN 978-0-252-05417-4. JSTOR jj.4116455 (https://www.jstor.org/stable/jj.4116455). OCLC 1371410330 (https://search.worldcat.org/oclc/1371410330).

External links

- ACM SIGIR: Information Retrieval Special Interest Group (https://sigir.org/)
- BCS IRSG: British Computer Society – Information Retrieval Specialist Group (https://www.bcs.org/membership-and-registrations/member-communities/information-retrieval-specialist-group)
- Text Retrieval Conference (TREC) (https://trec.nist.gov/)
- Forum for Information Retrieval Evaluation (FIRE) (http://www.isical.ac.in/~fire)
- Information Retrieval (https://www.dcs.gla.ac.uk/Keith/Preface.html) (online book) by C. J. van Rijsbergen
- Information Retrieval Wiki (http://ir.dcs.gla.ac.uk/wiki/) Archived (https://web.archive.org/web/20151124065507/http://ir.dcs.gla.ac.uk/wiki) 2015-11-24 at the Wayback Machine
- Information Retrieval Facility (http://ir-facility.org/) Archived (https://web.archive.org/web/20080522151226/http://www.ir-facility.org/) 2008-05-22 at the Wayback Machine
- TREC report on information retrieval evaluation techniques (https://trec.nist.gov/pubs/trec15/appendices/CE.MEASURES06.pdf)
- How eBay measures search relevance (https://innovation.ebayinc.com/tech/engineering/measuring-search-relevance/)

- Information retrieval performance evaluation tool @ Athena Research Centre (http://retrieval.ceti.gr)