Information retrieval

From Wikipedia, the free encyclopedia

Information retrieval (**IR**) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing.

Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications.

Contents

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 relevancy.

An object is an entity that is represented by information in a database. User queries are matched against the database information. Depending on the application the data objects may be, for example, text documents, images,[1] audio,[2] mind maps[3] 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.[4]
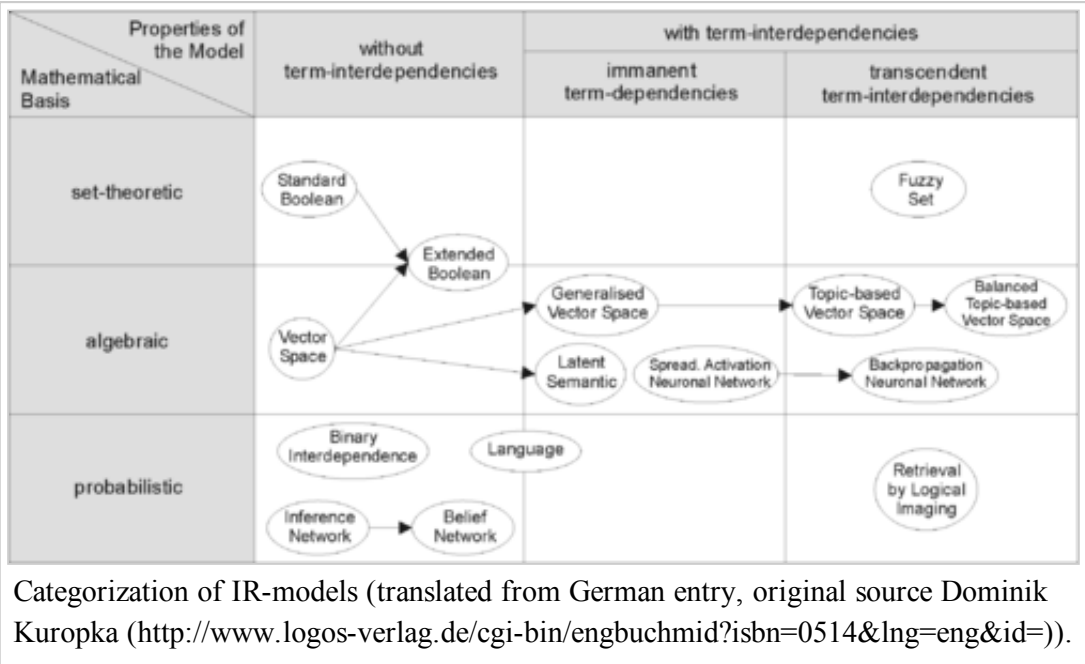
History

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.[5] The first automated information retrieval systems were introduced in the 1950s and 1960s. By 1970 several different techniques had been shown to perform well on small text corpora such as the Cranfield collection (several thousand documents).[5] Large-scale retrieval systems, such as the Lockheed Dialog system, came into use early in the 1970s.

> " But do you know that, although I have kept the diary [on a phonograph] for months past, it never once struck me how I was going to find any particular part of it in case I wanted to look it up? "
>
> —Dr Seward, Bram Stoker's *Dracula*, 1897

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.

Model types

For effectively retrieving 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

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=)).

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 the 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 a yet another feature.

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 an external source for the degree of interdependency between two terms. (For example a human or sophisticated algorithms.)

Performance and correctness measures

Many different measures for evaluating the performance of information retrieval systems have been proposed. The measures require a collection of documents and a query. All common measures described here assume a ground truth notion of relevancy: 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 relevancy.

Precision

Precision is the fraction of the documents retrieved that are relevant to the user's information need.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

In binary classification, precision is analogous to positive predictive value. Precision takes all retrieved documents into account. It can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called *precision at n* or *P@n*.

Note that the meaning and usage of "precision" in the field of Information Retrieval differs from the definition of accuracy and precision within other branches of science and statistics.

Recall

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

In binary classification, recall is often called sensitivity. So it can be looked at as *the probability that a relevant document is retrieved by the query*.

It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by computing the precision.

Fall-out

The proportion of non-relevant documents that are retrieved, out of all non-relevant documents available:

$$\text{fall-out} = \frac{|\{\text{non-relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{non-relevant documents}\}|}$$

In binary classification, fall-out is closely related to specificity and is equal to $(1 - \text{specificity})$. It can be looked at as *the probability that a non-relevant document is retrieved by the query*.

It is trivial to achieve fall-out of 0% by returning zero documents in response to any query.

F-measure

The weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score is:

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}.$$

This is also known as the F_1 measure, because recall and precision are evenly weighted.

The general formula for non-negative real β is:

$$F_\beta = \frac{(1 + \beta^2) \cdot (\text{precision} \cdot \text{recall})}{(\beta^2 \cdot \text{precision} + \text{recall})}.$$

Two other commonly used F measures are the F_2 measure, which weights recall twice as much as precision, and the $F_{0.5}$ measure, which weights precision twice as much as recall.

The F-measure was derived by van Rijsbergen (1979) so that F_β "measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision". It is based on van Rijsbergen's effectiveness measure $E = 1 - \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}}$. Their relationship is $F_\beta = 1 - E$ where

$$\alpha = \frac{1}{1 + \beta^2}.$$

Average precision

Precision and recall are single-value metrics based on the whole list of documents returned by the system. For systems that return a ranked sequence of documents, it is desirable to also consider the order in which the returned documents are presented. By computing a precision and recall at every position in the ranked sequence of documents, one can plot a precision-recall curve, plotting precision $p(r)$ as a function of recall r. Average precision computes the average value of $p(r)$ over the interval from $r = 0$ to $r = 1$:[6]

$$\text{AveP} = \int_0^1 p(r)dr$$

That is the area under the precision-recall curve. This integral is in practice replaced with a finite sum over every position in the ranked sequence of documents:

$$\text{AveP} = \sum_{k=1}^{n} P(k)\Delta r(k)$$

where k is the rank in the sequence of retrieved documents, n is the number of retrieved documents, $P(k)$ is the precision at cut-off k in the list, and $\Delta r(k)$ is the change in recall from items $k-1$ to k.[6]

This finite sum is equivalent to:

$$\text{AveP} = \frac{\sum_{k=1}^{n}(P(k) \times \text{rel}(k))}{\text{number of relevant documents}}$$

where $\text{rel}(k)$ is an indicator function equaling 1 if the item at rank k is a relevant document, zero otherwise.[7] Note that the average is over all relevant documents and the relevant documents not retrieved get a precision score of zero.

Some authors choose to interpolate the $p(r)$ function to reduce the impact of "wiggles" in the curve.[8][9] For example, the PASCAL Visual Object Classes challenge (a benchmark for computer vision object detection) computes average precision by averaging the precision over a set of evenly spaced recall levels {0, 0.1, 0.2, ... 1.0}:[8][9]

$$\text{AveP} = \frac{1}{11} \sum_{r \in \{0, 0.1, \ldots, 1.0\}} p_{\text{interp}}(r)$$

where $p_{\text{interp}}(r)$ is an interpolated precision that takes the maximum precision over all recalls greater than r:

$$p_{\text{interp}}(r) = \max_{\tilde{r}:\tilde{r} \geq r} p(\tilde{r}).$$

An alternative is to derive an analytical $p(r)$ function by assuming a particular parametric distribution for the underlying decision values. For example, a *binormal precision-recall curve* can be obtained by assuming decision values in both classes to follow a Gaussian distribution.[10]

R-Precision

Precision at **R**-th position in the ranking of results for a query that has **R** relevant documents. This measure is highly correlated to Average Precision. Also, Precision is equal to Recall at the **R**-th position.

Mean average precision

Mean average precision for a set of queries is the mean of the average precision scores for each query.

$$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AveP(q)}}{Q}$$

where Q is the number of queries.

Discounted cumulative gain

DCG uses a graded relevance scale of documents from the result set to evaluate the usefulness, or gain, of a document based on its position in the result list. The premise of DCG is that highly relevant documents appearing lower in a search result list should be penalized as the graded relevance value is reduced logarithmically proportional to the position of the result.

The DCG accumulated at a particular rank position p is defined as:

$$\text{DCG}_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}.$$

Since result set may vary in size among different queries or systems, to compare performances the normalised version of DCG uses an ideal DCG. To this end, it sorts documents of a result list by relevance, producing an ideal DCG at position p ($IDCG_p$), which normalizes the score:

$$\text{nDCG}_p = \frac{DCG_p}{IDCGp}.$$

The nDCG values for all queries can be averaged to obtain a measure of the average performance of a ranking algorithm. Note that in a perfect ranking algorithm, the DCG_p will be the same as the $IDCG_p$ producing an nDCG of 1.0. All nDCG calculations are then relative values on the interval 0.0 to 1.0 and so are cross-query comparable.

Other Measures

- Mean reciprocal rank
- Spearman's rank correlation coefficient

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 citation indexing (Eugene Garfield).

1950: The term "information retrieval" appears to have been coined by Calvin Mooers.[11]

1951: Philip Bagley conducted the earliest experiment in computerized document retrieval in a master thesis at MIT.[12]

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.

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[13] 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 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" (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."[14]

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.

- **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

  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*[15] with emphasis on visualization and multi-reference point systems.

  late 1990s: Web search engines implementation of many features formerly found only in experimental IR systems. Search engines become the most common and maybe best instantiation of IR models.

Awards in the field

- Tony Kent Strix award
- Gerard Salton Award

See also

- Adversarial information retrieval
- Collaborative information seeking
- Controlled vocabulary
- Cross-language information retrieval
- Data mining
- European Summer School in Information Retrieval
- Human–computer information retrieval
- Information extraction
- Information Retrieval Facility
- Knowledge visualization
- Multimedia Information Retrieval
- List of information retrieval libraries
- Personal information management
- Relevance (Information Retrieval)
- Relevance feedback
- Rocchio Classification
- Search index
- Social information seeking
- Special Interest Group on Information Retrieval
- Structured Search
- Subject indexing
- Temporal information retrieval
- Tf-idf
- XML-Retrieval
- Key-objects

References

1. Goodrum, Abby A. (2000). "Image Information Retrieval: An Overview of Current Research". *Informing Science* **3** (2).
2. Foote, Jonathan (1999). "An overview of audio information retrieval". *Multimedia Systems* (Springer).
3. Beel, Jöran; Gipp, Bela; Stiller, Jan-Olaf (2009). "Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom'09)". Washington, DC: IEEE. |chapter= ignored (help)
4. Frakes, William B. (1992). *Information Retrieval Data Structures & Algorithms* (http://www.scribd.com/doc/13742235/Information-Retrieval-Data-Structures-Algorithms-William-B-Frakes). Prentice-Hall, Inc. ISBN 0-13-463837-9.
5. Singhal, Amit (2001). "Modern Information Retrieval: A Brief Overview" (http://singhal.info/ieee2001.pdf). *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* **24** (4): 35–43.
6. Zhu, Mu (2004). Missing or empty |title= (help); |chapter= ignored (help)
7. Turpin, Andrew; Scholer, Falk (2006). "User performance versus precision measures for simple search tasks". *Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Seattle, WA, August 06–11, 2006)* (New York, NY: ACM): 11–18. doi:10.1145/1148170.1148176 (https://dx.doi.org/10.1145%2F1148170.1148176). ISBN 1-59593-369-7.
8. Everingham, Mark; Van Gool, Luc; Williams, Christopher K. I.; Winn, John; Zisserman, Andrew (June 2010). "The PASCAL Visual Object Classes (VOC) Challenge"

(http://pascallin.ecs.soton.ac.uk/challenges/VOC/pubs/everingham10.pdf). *International Journal of Computer Vision* (Springer) **88** (2): 303–338. doi:10.1007/s11263-009-0275-4 (https://dx.doi.org/10.1007%2Fs11263-009-0275-4). Retrieved 2011-08-29.

9. Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich (2008). *Introduction to Information Retrieval* (http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html). Cambridge University Press.

10. K.H. Brodersen, C.S. Ong, K.E. Stephan, J.M. Buhmann (2010). The binormal assumption on precision-recall curves (http://icpr2010.org/pdfs/icpr2010_ThBCT8.28.pdf). *Proceedings of the 20th International Conference on Pattern Recognition*, 4263-4266.

11. Mooers, Calvin N.; *Theory Digital Handling Non-numerical Information* (Zator Technical Bulletin No. 48) 5, cited in "information, n.". OED Online. December 2011. Oxford University Press.

12. Doyle, Lauren; Becker, Joseph (1975). *Information Retrieval and Processing*. Melville. pp. 410 pp. ISBN 0-471-22151-1.

13. 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). *Information Processing and Management* **44** (2): 971–972. doi:10.1016/j.ipm.2007.02.012 (https://dx.doi.org/10.1016%2Fj.ipm.2007.02.012).

14. 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://dx.doi.org/10.1016%2F0020-0271%2871%2990051-9).

15. Korfhage, Robert R. (1997). *Information Storage and Retrieval* (http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471143383,descCd-authorInfo.html). Wiley. pp. 368 pp. ISBN 978-0-471-14338-3.

External links

- ACM SIGIR: Information Retrieval Special Interest Group (http://www.acm.org/sigir/)
- BCS IRSG: British Computer Society - Information Retrieval Specialist Group (http://irsg.bcs.org/)
- Text Retrieval Conference (TREC) (http://trec.nist.gov)
- Forum for Information Retrieval Evaluation (FIRE) (http://www.isical.ac.in/~fire)
- Information Retrieval (http://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/)
- Information Retrieval Facility (http://ir-facility.org/)
- Information Retrieval @ DUTH (http://www.nonrelevant.net)
- Introduction to Information Retrieval (online book) by Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Cambridge University Press. 2008. (http://nlp.stanford.edu/IR-book/)

Wikiquote has quotations related to: *Information retrieval*

Retrieved from "http://en.wikipedia.org/w/index.php?title=Information_retrieval&oldid=648491872"

Categories: Information retrieval │ Natural language processing

---

- This page was last modified on 23 February 2015, at 16:06.