What is Information Retrieval

Jacob Murel Ph.D., Senior Technical Content Creator

Meredith Syed, Technical Content, Editorial Lead, IBM

28 August 2024

Information retrieval addresses data retrieval for user queries. It powers search tools such as library catalogs and web search engines.

Information retrieval (IR) is a broad field of computer science and information science. We can generally define it as finding unstructured data within a large collection in order to satisfy a particular information need.[1] An IR system—information retrieval system—provides material in response to a given query. The system searches collections for items relevant to the user's query. It then returns those items to the user, typically in list form sorted per computed relevance.[2]

IR systems and techniques power an array of search tools, such as web search engines and digital library catalogs.

Information retrieval versus data retrieval

Note that many online sources contrast IR systems with data retrieval: IR systems retrieve unstructured information, such as text documents and web pages; data retrieval, by contrast, deals in structured data, as found in relational database management systems. By extension, data retrieval uses a structured query language (SQL) for conducting search queries.

This distinction between IR as unstructured and non-relational versus data retrieval as structured and relational, however, is more equivocal than many online sources suggest. IR systems index, and thereby structure, information. For instance, while it is true that IR traditionally deals with raw text document retrieval, some IR systems use XML to represent and index texts. Research literature often describes XML-based systems as a branch of IR called *structured retrieval* or *semi-structured retrieval*.[3] Additionally, literature has explored the use of relational IR models for decades.[4]

The distinction between IR and data retrieval is thus more ambiguous than traditionally held. Indeed, given that data is, by definition, information, structured data retrieval is perhaps better understood as a type of information retrieval.

Information retrieval versus recommender systems

Note that IR is distinct from recommender systems. Machine learning recommendation techniques—such as collaborative filtering and content-based filtering—can perhaps be understood as a form of information filtering, a sub-task of IR systems. Nevertheless, IR and recommender systems are distinct. IR traditionally requires a user query; recommendation engines typically retrieve objects without a user query.[5]

How information retrieval systems work

Different IR models represent information in different ways. The chosen form of document representation largely determines how the model searches and retrieves information. Nevertheless, indexing, weighting and relevance feedback are three information retrieval techniques common across IR models.

Indexing

Indexing essentially amounts to metadata creation.[6] Many people have encountered an index at the back of a printed book. It is a structured set of words compiled from the given print document that allows readers to readily access passages on given topics. The IR index is similar. An IR index (or *inverted index*) is a data structure sourced from a set of documents intended to improve search query results.[7]

Index construction requires first parsing a document for feature extraction. For instance, say we are creating an IR system for text-based documents. As is common in natural language processing (NLP), we prepare the collection of documents with various preprocessing techniques, such as tokenization and stop word removal. The IR system then represents this processed collection of documents as an organized data structure. One such structure is a dictionary in which each document has an ID pointed to by the words (or index terms) that appear therein.[8] Another potential data structure for a text retrieval system is a vector space model, such as a bag of words.[9] Both these approaches extract words as features, which are then used to retrieve and rank documents in response to user queries.

Weighting

How does a search system rank approximate or exact matches for a given query? Approaches to the ranking and retrieval of information depend on both the type of information retrieval model and form of document representation used in the system. Index terms, however, play a key role in how an IR system ranks documents in response to queries. But not all index terms are equal. IR systems thus utilize different methods to weight index terms per their perceived importance.

IR systems using vector space models, such as bag of words, may use term frequency-inverse document frequency (TF-IDF). TF-IDF is a variation of bag of words that accounts for a word's prevalence throughout each document in the text set. The more documents in which a given word appears, the greater TF-IDF reduces that word's weight. Other approaches include singular value decomposition (SVD) and latent semantic analysis (LSA), the latter which is a common topic modeling approach.[10]

Such weighting approaches affect how IR systems rank documents in response to queries. But different types of IR models use these weights for ranking in different ways.

Relevance feedback

How might a system improve its search results? That is, how might a system fine-tune a user's search and increase the number of returned relevant documents?

Relevance feedback is a common information retrieval technique for improving search results. Relevance feedback essentially gathers information about user response to an initial set of query results. The system then reweights item relevance in light of the user's responses. It then returns a new set of search results incorporating the initial query and the user's feedback to that initial set of query results.

Relevance feedback typically involves user's providing explicit responses on the relevance of retrieved documents. Implicit feedback is a variation that deduces item relevance by observing user behavior—or example, which website links a user clicks in a search results page. Pseudo-relevance feedback assumes that an initial query's first *n* retrieved documents are relevant. It then gathers additional features common throughout those documents to modify the query further.[11]

Types of information retrieval techniques

There are numerous types of information retrieval models. To provide anything in the way of an exhaustive summary requires a much larger discussion. Nevertheless, IR textbooks and encyclopedic overviews often overlap in mentioning three general IR methodologies: Boolean, algebraic and probabilistic.

Boolean model

Boolean models are perhaps the most straightforward, even simplistic, IR models. They use a dictionary structure of index terms as described previously. The model then ranks documents according to the presence of words from a user's query in the retrieved documents. For instance, if a user gives the query, "jazz AND dancing," the Boolean model retrieves only those documents that contain the words *jazz* and *dancing* in combination. Boolean models thereby account only for the presence or absence of words in a document; partial matches do not exist in Boolean retrieval systems. Text preprocessing techniques such as stemming and lemmatization can solve this issue of morphological variants—such as documents that contain *dance*, *dances*, or *dancer*, rather than only the user's query *dancing*.

https://www.ibm.com/think/topics/information-retrieval

As mentioned, Boolean models only consider the presence and absence of words. This binary decision criterion lacks a grading scale to determine which documents are most pertinent to a user's query. One potential solution is to grade documents based on the frequency of user query terms therein. In other words, the more a document mentions *jazz* and *dancing*, the more pertinent the model considers it for the user's query. Increased term frequency does not necessarily indicate greater relevance however. Despite this potential drawback, Boolean models have been used in many IR systems given their ease of implementation.[12]

Algebraic model

Boolean document retrieval inhibits any form of partial matching. Algebraic and probabilistic models address this matter by assigning non-binary weights to index terms.

One representative algebraic model is the vector space model. In this approach, the IR system represents documents and queries as vectors in a multi-dimensional vector space. In this space, index terms will likely be features of the vector space, and queries and documents are plotted across this space according to the presence and frequency with which they contain index terms. The IR system computes similarity between a search query and documents according to their proximity in vector space.

There are a number of metrics for determining proximity in a vector space model, such as Jaccard and dot product. Perhaps one of the most common, however, is cosine similarity, represented by the formula:

$$\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Here, x and y signify two vectors in the vector space. The cosine similarity score can be any value between -1 and 1. The higher the cosine score, the more alike two items are considered.

The IR vector space model returns documents in order according to their measured degree of similarity. In this way, algebraic IR systems, such as the vector space model, allow for partial matching, potentially providing a more precise or nuanced form of information retrieval.[13]

https://www.ibm.com/think/topics/information-retrieval

Probabilistic model

Probabilistic models also allow for partial matching between user queries and documents. Probabilistic models function on that assumption that a given query has an ideal set of retrieved information system resources. This ideal set is, admittedly, unknown. But index term semantics can characterize the properties of this set.

Like algebraic models, probabilistic models use index term presence and frequency to determine similarity between queries and documents. But probabilistic models differ in that they consider additional factors. For example, they may account for index term co-frequency—how often index terms co-occur in a document—in relation to the document's full-text length, or how often a single index term occurs over all of the query terms in a given query. These are only some potential factors considered—a more detailed discussion requires more thorough understanding of probability theory.

Note that not all probabilistic models consider the same factors when computing document similarity, or probability. For instance, the binary independence model (BIM), the first probabilistic IR model, does not consider term frequency. A model incorporating the topic modeling technique latent Dirichlet allocation (LDA), however, will account for term co-frequency.[14]

Recent research

Bias. Web search engines are perhaps one of the most well-known IR use cases. The text summarization tool PageRank is used to retrieve and rank webpages (HTML documents). Research well establishes the unfortunate reality that search algorithms perpetuate a host of biases, such as racial and gender-based.[15] In response, published experiments explore a host of methods for reducing social bias in IR systems, such as negative sampling[16] and bias-aware algorithms that incorporate penalties for biased results.[17] Mitigating bias is a paramount area for research to develop ethical praxis around IR and even artificial intelligence.

Footnotes

1 Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze, *An Introduction to Information Retrieval*, Cambridge University Press, 2009.

2 Qiaozhu Mei and Dragomir Radev, "Information Retrieval," *The Oxford Handbook of Computational Linguistics*, 2nd edition, Oxford University Press, 2016.

3 Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze, *An Introduction to Information Retrieval*, Cambridge University Press, 2009. Mounia Lalmas and Ricardo Baeza-Yates, "Structured Document Retrieval," *Encyclopedia of Database Systems*, Springer, 2018.

https://www.ibm.com/think/topics/information-retrieval

[4] Robert Crawford, “The relational model in information retrieval,” *Journal of the American Society for Information Science*, Vol. 32, No. 1, 1981, pp. 51-64.

[5] Alejandro Bellogín and Alan Said, “Information Retrieval and Recommender Systems,” *Data Science in Practice*, Springer, 2018.

[6] Jeffrey Pomerantz, *Metadata*, MIT Press, 2015.

[7] Steven Beitzel, Eric Jensen, and Ophir Frieder, “Index Creation and File Structures,” *Encyclopedia of Database Systems*, Springer, 2018.

[8] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze, *An Introduction to Information Retrieval*, Cambridge University Press, 2009.

[9] Qiaozhu Mei and Dragomir Radev, “Information Retrieval,” *The Oxford Handbook of Computational Linguistics*, 2nd edition, Oxford University Press, 2016.

[10] Qiaozhu Mei and Dragomir Radev, “Information Retrieval,” *The Oxford Handbook of Computational Linguistics*, 2nd edition, Oxford University Press, 2016. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, ACM Press, 1999.

[11] Qiaozhu Mei and Dragomir Radev, “Information Retrieval,” *The Oxford Handbook of Computational Linguistics*, 2nd edition, Oxford University Press, 2016. Stefan Büttcher, Charles Clarke, and Gordon Cormack, *Information Retrieval: Implementing and Evaluating Search Engines,* MIT Press, 2016.

[12] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, ACM Press, 1999. Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze, *An Introduction to Information Retrieval*, Cambridge University Press, 2009.

[13] Qiaozhu Mei and Dragomir Radev, “Information Retrieval,” *The Oxford Handbook of Computational Linguistics*, 2nd edition, Oxford University Press, 2016. Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze, *An Introduction to Information Retrieval*, Cambridge University Press, 2009.

[14] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, ACM Press, 1999. Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze, *An Introduction to Information Retrieval*, Cambridge University Press, 2009.

[15] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, NYU Press, 2018.

[16] Amin Bigdeli et al., “A Light-Weight Strategy for Restraining Gender Biases in Neural Rankers,” *Proceedings of the 44th European Conference on Advances in Information Retrieval*, 2022, pp. 47-55.

[17] Dhanasekar Sundararaman and Vivek Subramanian, "Debiasing Gender Bias in Information Retrieval Models," 2022, https://arxiv.org/abs/2208.01755. Shirin Seyed

https://www.ibm.com/think/topics/information-retrieval

Salehi et al., “Bias-aware Fair Neural Ranking for Addressing Stereotypical gender Biases,” Microsoft Research, 2022.