

本資料由成功大學資訊工程系蔣榮先教授整理提供

Boolean query

A query that is a Boolean combination of terms. Some examples are INFORMATION AND RETRIEVAL, VISION OR SIGHT, and CLINTON AND (NOT GORE).

Classification

The process of deciding the appropriate category for a given document. Examples are deciding what newsgroup an article belongs in, what folder an email message should be directed to, or what is the general topic of an essay.

Cluster

A grouping of representations of similar documents. In a vector space model, one can perform retrieval by comparing a query vector with the centroids of clusters. One can continue search in those clusters that are in this way most promising.

Collaborative Filtering

The process of filtering documents by determining what documents other users with similar interests and/or needs found relevant. Also called "social filtering".

Collection

A group of documents that a user wishes to get information from. See also test collection.

Collection Fusion

The problem of combining the search results from multiple collections. This could be tricky since some measures such as IDF will differ across collections, and, if one retrieves a fixed number of documents, it is unclear how many to take from each collection.

Content-Based Filtering

The process of filtering by extracting features from the text of documents to determine the documents' relevance. Also called "cognitive filtering".

Cosine Similarity

See similarity.

Document

A piece of information the user may want to retrieve. This could be a text file, a WWW page, a newsgroup posting, a picture, or a sentence from a book.

Indexing

The process of converting a collection into a form suitable for easy search and retrieval.

Information Extraction

A related area that attempts to identify semantic structure and other specific types of information from unrestricted text.

Information Filtering

Given a large amount of data, return the data that the user wants to see. This is the standard problem in IR.

Information Need

What the user really wants to know. A query is an approximation to the information need.

Information Retrieval

The study of systems for indexing, searching, and recalling data, particularly text or other unstructured forms.

Inverse Document Frequency

Abbreviated as IDF, this is a measure of how often a particular term appears across all of the documents in a collection. It is usually defined as $\log(\text{collection size/number of documents containing the term})$. So common words will have low IDF and words unique to a document will have high IDF. This is typically used for weighting the parameters of a model.

Inverted File

A representation for a collection that is essentially an index. For each word or term that appears in the collection, an inverted file lists each document where it appears. This representation is especially useful for performing Boolean queries.

Precision

A standard measure of IR performance, precision is defined as the number of relevant documents retrieved divided by the total number of documents retrieved. For example, suppose there are 80 documents relevant to widgets in the collection. System X returns 60 documents, 40 of which are about widgets. Then X's precision is $40/60 = 67\%$. In an ideal world, precision is 100%. Since this is easy to achieve (by returning just one document), a system attempts to maximize both precision and recall simultaneously.

Precoordination of terms

The process of using compound terms to describe a document. For example, this page may be indexed under the term "information retrieval glossary".

Postcoordination of terms

The process of using single terms to describe a document which are then combined (or coordinated) based on a given query. For example, this page may be indexed under the words INFORMATION, RETRIEVAL, and GLOSSARY. We'd then have to combine these terms based on a query like "INFORMATION and RETRIEVAL".

Probabilistic Model

Any model that considers the probability that a term or concept appears in a document, or that a document satisfies the information need. A Bayesian inference net is a good framework for this style of model. The INQUERY system is the most successful example.

Query

A string of words that characterizes the information that the user seeks. Note that this does not have to be an English language question.

Query Expansion

Any process which builds a new query from an old one. It could be created by adding terms from other documents, as in relevance feedback, or by adding synonyms of terms in the query (as found in a thesaurus).

Question Answering

The problem of finding the exact answer to a user's natural language question in a large collection.

Recall

A standard measure of IR performance, recall is defined as the number of relevant documents retrieved divided by the total number of relevant documents in the collection. For example, suppose there are 80 documents relevant to widgets in the collection. System X returns 60 documents, 40 of which are about widgets. Then X's recall is $40/80 = 50\%$. In an ideal world, recall is 100%. However, since this is trivial to achieve (by retrieving all of the documents), a system attempts to maximize both recall and precision simultaneously.

Relevance

An abstract measure of how well a document satisfies the user's information need. Ideally, your system should retrieve all of the relevant documents for you. Unfortunately, this is a subjective notion and difficult to quantify.

Relevance Feedback

A process of refining the results of a retrieval using a given query. The user indicates which documents from those returned are most relevant to his query. The system typically tries to find terms common to that subset, and adds them to the old query. It then returns more documents using the revised query. This can be repeated as often as desired. Also called "find similar documents" or "query by example".

Robot

See spider.

Routing

Similar to information filtering, the problem of retrieving wanted data from a continuous stream of incoming information (i.e. long-term filtering).

SIGIR

The ACM's special interest group on Information Retrieval. They publish SIGIR Forum and have an annual conference. For more information, check their home page.

Signature File

A representation of a collection where documents are hashed to a bit string. This is essentially a compression technique to permit faster searching.

Similarity

The measure of how alike two documents are, or how alike a document and a query are. In a vector space model, this is usually interpreted as how close their corresponding vector representations are to each other. A popular method is to compute the cosine of the angle between the vectors.

Spider

Also called a robot, a program that scans the web looking for URLs. It is started at a particular web page, and then access all the links from it. In this manner, it traverses the graph formed by the WWW. It can record information about those servers for the creation of an index or search facility. Most search engines are created using spiders. The problem with them is, if not written properly, they can make a large number of hits on a server in a short space of time, causing the system's performance to decay. For more information, look here.

Stemming

The process of removing prefixes and suffixes from words in a document or query in the formation of terms in the system's internal model. This is done to group words that have the same conceptual meaning, such as WALK, WALKED, WALKER, and WALKING. Hence the user doesn't have to be so specific in a query. The Porter stemmer is a well-known algorithm for this task. You can download some source code for this algorithm here. (Unfortunately, I don't remember where I downloaded it from originally.) Be careful: stemming the word PORTER in PORTER STEMMER to PORT would allow hits with documents about boats or wine.

Stopword

A word such as a preposition or article that has little semantic content. It also refers to words that have a high frequency across a collection. Since stopwords appear in many documents, and are thus not helpful for retrieval, these terms are usually removed

from the internal model of a document or query.

Some systems have a predetermined list of stopwords. However, stopwords could depend on context. The word COMPUTER would probably be a stopword in a collection of computer science journal articles, but not in a collection of articles from Consumer Reports.

Term

A single word or concept that occurs in a model for a document or query. It can also refer to words in the original text.

Term Frequency

Abbreviated as TF, the number of times a particular term occurs in a given document or query. This count is used in weighting the parameters of a model.

Test collection

A collection specifically created for evaluating experimental IR systems. It usually comes with a set of queries, and a labelling (decided by human experts) that decides which documents are relevant to each query. TIPSTER is one of the most prevalent test collections currently. Another useful collection for classification is the Reuters text categorization test collection. Here there are no queries, but the documents are news articles labelled with a variety of topic designations.

TIPSTER

An ongoing project where various groups and institutions have pooled their resources to solve problems in routing and information extraction. The framework is such that each team can work on a different piece and simply "plug" their application into the general architecture. The project also has a large test collection available.

TREC

Text REtrieval Conference. This group gives IR researchers a common test collection and a common evaluation system. Hence, systems can be compared and contrasted on the same data. You can visit the conference's home page for information about the conference and on-line versions of the proceedings.

Vector Space Model

A representation of documents and queries where they are converted into vectors. The features of these vectors are usually words in the document or query, after stemming and removing stopwords. The vectors are weighted to give emphasis to terms that exemplify meaning, and are useful in retrieval. In a retrieval, the query vector is compared to each document vector. Those that are closest to the query are considered to be similar, and are returned. SMART is the most famous example of a system that uses a vector space model.

Weighting

Usually referring to terms, the process of giving emphasis to the parameters for more important terms. In a vector space model, this is applied to the features of each vector. A popular weighting scheme is TF*IDF. Other possible schemes are Boolean (1 if the term appears, 0 if not), or by term frequency alone. In a vector model, the weights are sometimes normalized to sum to 1, or by dividing by the square root of the sum of their squares.

References

Faloutsos and Oard, "A Survey of Information Retrieval and Filtering Methods"

Salton and McGill, Introduction to Modern Information Retrieval, McGraw-Hill, 1983