



Informational Retrieval Glossary

cataphora

Boolean

Boolean search is a common technique which allows terms to be logically connected using AND, OR and NOT to refine a query. For example: "(mercury OR saturn) AND NOT automobile" would return all documents containing either the word mercury or the word saturn and not containing the word automobile.

Bayesian Classification

In general, bayesian classification is based on the statistical probability of a class and the features associated with that class. This type of classification utilizes a training set composed of classes that have correctly assigned features. Once the probabilities of the training set features and classes have been stored, new data is compared against the training set. During this comparison of the "learned" classification of the training set with the new data, the new data's features are calculated and the new data are assigned classes whose probability of matching the training set's classes and features is highest.

Categorization

In general, categorization is the grouping of objects, people or ideas on the basis of some kind of "similarity". As applied to electronic discovery, it describes the grouping of documents according to some desired criteria. Categorization may be by topic or by legal criteria. Categorizing documents about specific products, or documents that relate to sales in a given country are examples of categorizing by topic. Legal criteria might include categorizing responsive, non-responsive, and privileged documents.

Clustering

Clustering is the grouping of information by some category or statistical similarity. This is done through various grammatical, semantic, and even punctuation algorithms designed by combination to detect topics rather than just individual keywords. Statistical clustering can be done by counting words and their frequency, then grouping those documents with similar statistics together in a cluster. When files are determined to be about the same or similar topics, they are clustered together, and usually displayed in some kind of graphical relationship that facilitates reviewing similar documents together.

Concept

Concept search attempts to find documents that address some concept that a user is interested in. To do so, it goes beyond *keyword* search for documents that contain a specified word or phrase, and tries to find other documents that address the underlying concept. For example, a concept search for "fiber" might return documents that refer to the concept of fiber using alternative terms such as cloth, material, cotton etc.

Keyword

Keyword search looks for documents that contain a specific word or phrase. Keyword searches may be further refined by using *Boolean* operators.

Latent Semantic Indexing

Latent Semantic Indexing has involves extracting multiple concepts from the data collections through a statistical semantic analysis of each file. The theory is that unstructured files comprise *latent* concepts that are not readily recognized and remain hidden until a more precise lexicon is developed out of the whole collection. These concepts then form a dictionary (lexicon) for the collection that can be weighted for both frequency of occurrence and relevance. At that point each file in the collection is compared to the concepts list, and it is assigned a *fingerprint* (or value) that uniquely defines the file according to those criteria. Searches can then be conducted by requesting files that are statistically similar, i.e. that have similar fingerprints, under the presumption they will be not just similar but conceptually related as well.

Linguistic Techniques

Search or categorization techniques that are based on analysis of language features of documents, in contrast with *statistical* techniques. *Ontologies* are an example of a linguistic technique.

Neural Network

A neural network is a computer program whose operation is loosely inspired by the way a human or animal brain works (though the neural network is much, much simpler). A neural network can be "trained" by giving it sample inputs and the correct outputs



Informational Retrieval Glossary

cataphora

associated with these. The network can analyze the difference between the answers it is generating and the "correct" answers. It can then automatically adjust its internal workings, until its answers on the training set adequately match the given outputs. The idea is that you can now feed it new inputs (the answers to which are unknown) and it should now be able to provide the correct outputs for these. For purposes of electronic discovery, the inputs might be information about documents and the outputs a *categorization* of those documents.

Ontology

An ontology is an arrangement of words, phrases and search terms under a *concept*. Here is a simple example:

AIRCRAFT CONCEPT

- Boeing
- 747
- Cessna
- Glider

By reading documents and establishing whether they contain any of the four terms listed under the AIRCRAFT CONCEPT, we can determine whether any of the documents discuss the concept of aircraft. This process can be automated, so that a computer does the work. If the computer finds a document that contains one or more of the four terms, it concludes that the document is (at least partially) about aircraft. The document might also discuss other concepts, but a reference to the concept of aircraft is clearly present in the document.

Relevant

"Having some reasonable connection with, and in regard to evidence in trial, having some value or tendency to prove a matter of fact significant to the case." Finding relevant documents more effectively than our competitors is a key value that we provide.

Related

Documents that may be related to each other in a number of ways, such as addressing the same issue, or being created by a significant *actor* at a crucial time. Most electronic discovery software can find only documents which are similar to each other, or which belong to the same *email thread*. Cataphora *Discussions* are the first technology to pull together related documents, not just similar ones.

Statistical

Categorization may be based on statistical analysis of the similarity of documents. For this, a document is mathematically represented by a set of features such as the occurrence of words, or their proximity to other words in the documents. Different weights (levels of importance) may be assigned to the various features. Documents are then deemed to be similar (and therefore belong to the same category), based on the degree to which their features resemble each other.

Taxonomy

Taxonomy is the practice and science of classification. Taxonomies, which are composed of *taxonomic units* known as *taxa* (singular *taxon*), are frequently hierarchical in structure, commonly displaying parent-child relationships.

Vector Space Modeling

Vector Space Modeling (VSM) is a concept that first came into favor in the early 1970s and it has provided some additional guidance in automated document review even to this day. It is based on building vectors that describe the relationships between each search query and each file in the collection. Each vector, by its magnitude and direction then maps to other files that are closest to it in relation to the same *feature* as emphasized by the search query. Each file thus becomes a compilation of *features* that place it in a multi-dimensional construct. That construct can be realized in a graphical display depicting all the relationships as vector lines between and among separate files.