

An Evolution of Search

by John D. Holt and David J. Miller

Search 2009

The technology of information retrieval systems continues to evolve, and in particular, the technology of search has continued to evolve. A new stage in the evolution of search has arrived with the advent of entity-based searching. This paper provides a brief review of some of the earlier stages of search evolution in the context of the evolutionary pressures of the concurrent improvement of both precision and recall.

Early Boolean Search

Efficient mechanical Boolean search of records is a 19th century invention. The extension of Boolean search from records (structured data) to text documents was initially accomplished by the simple expedient of creating a summary record of the text using a controlled vocabulary and a taxonomy. The records under search could refer to books, papers, notes, works of art or even public records. The subset of records that satisfied the Boolean predicate was selected, and the records that did not satisfy the predicate were left behind. In the case of needle-sort cards in buckets, this is a literal description of the process. The scope of the Boolean predicate was by necessity the record or fields on the record. The searcher could use the AND, OR and AND NOT Boolean operations in the construction of the search predicate.

The resolving power provided by the mechanical search approach was very low. The number of controlled vocabulary terms or taxonomy leaves was limited to the length of the card edges. In addition to searching through cards, print indices could be searched as well. A much larger number of

John D. Holt and David J. Miller are both senior architects in the Lexis Nexis Risk and Information Analytics Group. Holt can be reached by email at John.D.Holt@lexisnexis.com; Miller can be reached at David.J.Miller@lexisnexis.com

controlled vocabulary terms could be used in the construction of an index such as the *Readers Guide to Periodic Literature*, first published in 1901. An industrious researcher could perform the AND, OR and AND NOT operations by keeping lists of the document references and manually performing the operation upon those lists.

Boolean Search and the Digital Computer

The digital computer allowed for a large increase in the number of controlled vocabulary terms by the middle of the 20th century, and the number of terms that could be used expanded from about a hundred to several thousands. The search was still performed upon an extract of the original papers and upon indices. Computer memories were too small to seriously consider large collections of text or structured records.

The primary and secondary memory of the digital computer increased in the last quarter of the 20th century to the point where the complete text of papers, articles or large collections of structured data could be stored in computer systems. Additional operations such as adjacency became required to allow users to specify search predicates consisting of word patterns. These word patterns provided the context necessary to disambiguate the individual words.

A deep understanding of the document collection was required to properly specify the patterns of words. If the patterns were too loose, the number of documents that satisfied the predicate became too large for effective review. The resulting documents could sometimes be sequenced to allow the searcher to stop the review of the answer documents before all of the documents were examined. Unfortunately, most of the time there were no ways to naturally sequence the answer documents to provide an early termination of the review.

The search system stemmed the words to varying degrees to reduce the word variants that a searcher would need to enter. The system generally would allow the word strings to include wild card characters to provide additional flexibility. Some systems introduced word equivalents for common spelling variations. The notion of word equivalents eventually expanded to include alternate names of famous organizations.

The search of structured and semi-structured data collections presented similar problems. To achieve a result set of reasonable size, a highly specific search would need to be formulated. Unfortunately, a highly specific search was unlikely to return all of the records of interest. The negative impact of recording error upon structured and semi-structured data searches is higher than the impact of recording errors in text documents because there is much higher degree of redundancy found in text documents.

Relevance Ranking the Result

Gerald Salton is generally credited with the observation that more frequently occurring terms are less distinguishing than less frequently occurring terms. The inverse term frequency of the search terms can be used as weights to rank the documents. Most text search systems today use some variation of the vector space model where each word in the collection is considered an attribute and therefore a dimension. The search expression is treated as a vector of terms, and each document in the collection is also treated as a vector of terms. The similarity of the search vector to each document vector is calculated using the weights of the matching terms to the score the importance of the match.

The document ranking provided good results with very loose searches. The searcher no longer needed a deep understanding of the word patterns that were likely to be found in the relevant documents. The searcher was still required to know the vocabulary of the subject.

The presence of phrases and wild-carded string expressions complicates the process of ranking, as it is not obvious how to determine the weights efficiently. Index terms assigned to documents can also be used in the search process but can be difficult to use in the relevance ranking calculation.

Statistical indexing can also be performed upon text documents. The

statistical indexing approach uses the number of occurrences of content words and noun-noun phrases to calculate a statistical digest of the document. The search can be restricted to the set of statistical index terms to provide improved precision. The ranking can then be performed using the document terms.

Relevance ranking can be extended to semi-structured and structured information searching. The search again is treated as a vector, and each record in the collection is also a vector. The degree of similarity between the search and each of the records in the collection can be measured, and the most similar records are returned to the researcher. In practice, the usual technique is to calculate the degree of mismatch between the search and each of the records in the collection instead of the similarity.

Structured information has far fewer attributes than text documents where each unique word can be considered a different attribute. Since each attribute is treated as a dimension, the structured information record has a much lower dimensionality than a typical text document. This lower dimensionality of the structured information record enables the use of more sophisticated measures for similarity. In the case of the document, a term was either present or it was not present. For structured information, each attribute can be evaluated for the degree of similarity between the attribute values. For example, two strings can be compared to each other using an edit distance measure to determine the degree of similarity.

Smarter Data

The problem of vocabulary mismatch can be ameliorated by adding indexing terms. Indexing terms also remove some of the need for the searcher to anticipate the word patterns. Indexing alone falls short of achieving complete disambiguation.

Additional terms can be added to the search index of the document collection to supplement an original sequence of words. The sequence “Falkland Islands” can be supplemented with “Malvinas” as an alternate term. This is a simple example of a data enhancement. In the instant example, a search looking for “Malvinas” in the same paragraph as “Exocet” would find a story about the Falklands war and the use of an Exocet missile to sink the HMS Sheffield.

Semantic indexing and entity recognition are tools that can be used to make the documents easier to find. The user now only needs to specify the concepts or entities that must be discussed for the document to be returned. Unfortunately, semi-structured data such as public record data does not lend itself to semantic indexing techniques.

Entity recognition is a difficult problem. A natural language parsing process can be used to say that a particular string references an entity or concept, but cannot say anything about the entity referenced. An ontology is required to resolve the reference. An ontology for concepts, historical events, historically important people and organizations and places can be constructed by human efforts from historical accounts, biographies, gazetteers and maps. But what of an ontology of ordinary people and organizations or even of emerging concepts?

Discovery of Entities in Data

Text documents, semi-structured data (for example, bibliography entries) and structured data can all be processed to discover entities. Entities can be individual people, transient organizations (for example, the co-authors of a paper), organizations, places or events. The document text or record refers to an entity, but is not itself the entity.

Some entities can have concrete identities. Place names can be tied to particular geographic areas. A geographic area can be definitively identified with coordinates and boundary information. Time entities, like 7 December 1941, can be definitively identified. Some organizations can be definitively identified, though in practice fewer than one would think.

In general, very few non-geographic and non-chronological entity references can be definitively identified as a reference to a particular entity. What can be done is to group or associate the references into internally consistent sets of references. A particular set of references then describes an entity, and an entity is described by a set of references. The preferred error is to have more than one set of references describe a particular entity. In the domain of document collections, the process of the creation of sets of references to the entities is a form of clustering. In the domain of records management, the process of the creation of sets of references to the entities is known as record linkage.

A reference to an entity has attributes. A set of references that describe an entity will in practice have some subset of attributes in common for a subset of the references. It is quite possible and desirable that there not be any attributes that are in common for all of the references. There need only be enough commonality among the disparate references to link all of the references together.

For example, consider the case where there are three sources of information (source A, source B and source C), and each source has a variety of fields. There are three different pairings of the sources, A-B, A-C and B-C. As long as any two pairings have distinguishing fields in common the records can be linked. Each record is a reference to an entity.

Once the references (either records or documents or passages of documents) have been gathered into sets, the fact of set membership can be used for searching and reporting. The use of the set for reporting is simply the aggregation of the references that make up the set into a virtual document or record. There are two methods of leveraging the set of references to improve the search process for finding the documents or records of interest. Given a found record of interest, the information extracted from the other members of the set of references can be used to form a search that is both more precise and more inclusive. Alternatively, instead of using the individual records or documents as the object of the search, the entity is directly used as the object of the search by using the complete virtual document or record.

Example

Consider a collection of bibliographies and the abstracts and the keywords for each of the referenced papers. We can discover the ad hoc or transient organizations of co-authors and enhance the collection by adding entity information indicating the group of co-authors responsible for a particular paper. Group membership can be used to disambiguate the names of the individual authors. The journals and conference proceedings referenced can be categorized by clustering the keyword lists from the papers published in the journal or conference proceeding.

In some cases, there will be additional information available, such as

e-mail addresses of one or more of the authors, the academic affiliations of the authors and a credit to the grant and sometimes the grant identification.

Crawling the websites of academic institutions gathering lists of then-current faculty can augment the bibliography collection. Most faculty will maintain a list of their publications as well, and these lists will provide additional statistical information. Note that without human intervention it is not possible to assign a semantic meaning to a list of bibliographic references on a faculty member's web page. It may be a list of publications authored by the faculty member or it may be a suggested reading list.

The group entities are discovered using statistical record linking or clustering. The individual entities can be disambiguated using group membership as well as other attributes. The statistical linking or clustering will use the list of authors, the journal or conference proceeding, the cluster identity of the journal or conference proceeding, the publication date, idiomatic expressions found in the abstract and any other attributes that are available on a sufficient number of the records in the collection.

Using Entities to Improve Search Results

A search finds the set of records that match (or best match, if a relevance-ranked search) a predicate. The notion of expanding or augmenting a search has been shown to be an effective method of improving recall. The typical method of expansion is to use a statistical thesaurus. The query terms are used to look up thesaurus entries. The associated terms are pooled and a subset of terms is selected based upon criteria such as the degree of commonality.

A collection of entity information can be used as a statistical thesaurus for the purpose of query expansion. The expansion of the query with information from the entity may not be quite as simple or as straightforward as traditional query expansion. The number of terms in the traditional expanded search will vary with the number of matching thesaurus entries. Unfortunately, the number of expansion terms available from an entity-based thesaurus will vary by the number of entities related to the initial set of search terms. A simple expansion will only be practical when the number of initial entities involved is reasonably small, such as a few hundred entities.

Another form of query expansion is a "More Documents like this Document" search. The researcher runs an initial search that is fairly broad, browses the list of answer documents and finds a suitable seed document. The researcher then requests that the search system use the descriptive terms and phrases of the seed document as a second search to create a highly relevant document. A similar process can be used with entities. It is important to recall that our entities are really just references and there may be multiple sets of these references to each actual entity in our collection.

The ability to specify more complex search predicates is another means of improving the search results. Consider the search where you might want to find authors who had published papers on the same disparate set of topics. The search system could simply use the least frequently occurring topic in the set to get the initial list of author entities and use the list of author entities to retrieve the corresponding publication records and then filter the list of publication records by the remaining topics.

The expressive range of this approach for efficient search and retrieval is limited to the information found in the individual records in our collection and upon search criteria that produce intermediate result sets of manageable size. Information that is derived from some aggregate operations upon sets of records will require another approach. Search criteria that create intermediate result sets that are too large will also require a different approach.

Using Entities as the Objects of the Search

As was noted above, we really do not know the entities. What we do know are internally consistent sets of records that reference an entity or entities. The sets are constructed to prefer an error where more than one set refers to a particular entity and to avoid an error where a set refers to more than one particular entity. It is convenient to adopt the fiction that these sets of references are the entities.

A document can be constructed for each of the entities in the form of a report. An author group entity report can contain the reports of any individual author entities that can be unambiguously linked to the author group. The report can show the group core members, those authors that have participated at a very high rate, peripheral members that have participated in a plurality

of the publications and ad hoc members. The report need never actually exist, but can be virtual.

The virtual document is more than just the union of the individual records that comprise the virtual document. A variety of aggregations can also be included in the virtual document. In our present example, aggregations such as the average number of publications per year or the average length of time between publications can be included in the virtual document.

The creation of a set of complex derived attributes, such as aggregates, during the creation of the database supports efficient search and retrieval. If the data collection is small enough, these operations can be performed during the search process. However, for any reasonably sized collection there will be far too many intermediate results.

The search expression is matched against the virtual document or record. The virtual record has a significant number of attributes. It can be both convenient and effective to leverage the availability of these attributes by using a forms-based search interface. The system presents the researcher with a form. The entry boxes on the form solicit the attribute values from the researcher. The researcher now only needs to know the attribute values of interest. The system can match these attribute values against the attribute values for the entities and deliver the closest matching entities. It is important to note that it is not necessary for a single record to exist that contains all or even a plurality of the attribute values supplied.

Semi-structured information and text documents contain large numbers of text strings as attribute values. In the instant case, the abstract is an example of an attribute with a large number of text strings. The search form mechanism discussed above is inadequate when the attribute has a large number of text strings.

It can be helpful to support extensions to the search form interface to

solicit the specification that the strings entered on the form should be from different or the same component records. The system can support a search specification of more complex relationships between text strings because the system assembled the virtual document.

Conclusion

Entity search is another step in the evolution of information retrieval systems. Entity search builds upon Boolean and relevance ranking techniques. Entity search provides improvements in both precision and recall over traditional Boolean and relevance ranked search techniques.

Boolean search techniques require the researcher to be knowledgeable of the words and expressions used in the document or record collection. Precise results can be obtained, but at the cost of a significant drop in recall. Recall can be achieved, but only at a significant drop in precision.

Relevance ranking via statistical techniques can be used to improve apparent precision in some cases. However, the statistical techniques do not apply well to searching structured and semi-structured data with attribute values.

The linking or clustering of the documents or records into sets of references that describe an entity can be used for much more than just reporting on an entity. The information from the set can be used in some cases to improve recall by broadening the search. Alternatively, and more powerfully, the entity can become the object of the search.

A search expression that specifies a set of attribute values can be used when the entity is the object of the search. Both precision and recall are improved. Precision is improved because the entities returned are all consistent with the attribute values supplied in the search. Recall is improved because the combination of entity values specified in the search expression need not appear in any particular underlying reference document or record. ■