

# Why is Finding What You Want So Difficult?

## A Look at Information Retrieval Technology

*Jimmy Lin*

---

The explosive growth of the World Wide Web has given people potential access to more information than they have ever had before. The presence of this vast amount of digital information available online presents a new challenge to the field of information retrieval (IR), which is concerned with developing technologies that enable users to find the information they desire from a vast, often unorganized, collection of data. However, current available systems rarely fulfill users' information access needs. At best, these systems serve as a springboard from which users can conduct manual browsing, and, at worst, they return totally irrelevant results. The full potential of the World Wide Web as a vast repository of knowledge cannot be realized without a satisfactory means of retrieving the desired knowledge.

This paper will first provide a brief overview of current information retrieval technology, the attempt to expose the inadequacy of the current paradigm by presenting some alternative approaches that may prove more fruitful in satisfying users' queries. Also, extending information retrieval, which focuses primarily on text documents, to include the World Wide Web, which is a chaotic mosaic of multimedia elements, will be discussed.

### **Classic Information Retrieval**

The central tenet of information retrieval, which has basically remained unchanged over the last three decades, is colloquially known as the bag-of-words paradigm. This approach assumes that the set of words in a document is representative of its content and meaning, and if a document shares common words with the query, then it is likely to be a relevant answer to that query. This assumption serves as the underlying foundation of the traditional key word search engine, in which a user describes his or her information need by enumerating key words that pertain to it, often modified by Boolean operators (AND, OR, NOT), sometimes with weights (for importance).

The two most popular methods of implementing information retrieval systems are through inverted indices and key word matrices (Figure 1). In an inverted index, each unique word serves as the key to a list of pointers that point back to every occurrence of that word in the data collection.

Inverted indices usually support query via Boolean operators. Such operations can be realized by considering documents containing a particular key word as sets: thus, AND, OR, or NOT translates into the intersection, union, and inverse of sets, respectively. For example, the query “movie AND directors” becomes the intersection of the set of documents containing the word movie and the set of documents containing the word directors.

In a key word matrix, each document is condensed into a high-dimensional vector (typically thousands of dimensions) whose components represent the individual key words weighted according to some algorithm. A simple algorithm is to score each component directly as a function of term frequency (the number of occurrences of a particular word in a particular document). Another popular, but more complex scheme is known as Term Frequency Inverse Document

Frequency (TFIDF), which also takes into account the distribution of a particular word across all documents in addition to term frequency. This basic idea is that if a word appears frequently in a document, it must be an important key word, unless it also appears frequently in other documents.

This approach assigns low importance to common function words that appear in almost all documents, such as and or the, which contribute little to the content of a document. Retrieval using a key word matrix thus becomes an exercise in linear algebra—finding the document vector ‘closest’ to the query vector. A variety of methods have been developed to measure distance in this high dimensional space<sup>1</sup>, but one of the most popular and effective derives from the law of cosines, by measuring the angle subtended by the two vectors.

Breakthroughs in the field of information retrieval over the last three decades have neither addressed nor challenged the underlying bag-of-words assumption. Instead, research has focused on sophisticated weighting functions, often motivated by information theory and techniques that deal with large numbers of high dimensional vectors.<sup>1</sup> For example, principle component analysis (i.e., latent semantic indexing) is a technique for reducing the dimensionality of document vectors by finding first order correlation

among the components.<sup>2, 3, 4</sup> Also, various clustering techniques can decrease the number of documents by grouping similar documents together into a representative vector.<sup>1,5,6</sup>

Current information retrieval systems are inadequate in fulfilling users’ information needs because they operate on a fundamentally invalid assumption. Although a document can sometimes be described by its component key words, it always contains far more information due to the richness and expressive power of natural language. Documents that share common key words with the query may not be good answers to the query, while documents that share no common key words with the query may actually be great answers to the query. Problems with the bag-of-words paradigm can be grouped into four categories, discussed in detail below.

## Morphological variation

The bag-of-words assumption does take into account morphological variations of words. Morphology can be classified into two types: inflectional, which prescribes changes due to syntax without alteration of the part of speech, and derivation, which alters the part of speech, often transforming the word meaning as a result. Inflectional morphology includes singular and plural variations, possessives, comparative and superlative form (good, better, best), and verb conjugation. Derivational morphology includes addition of suffixes (e.g., -ize, -tion). Arguably, a request for a particular key word should extend to (at least a part of) its morphological variations. For example, a query for “wolves” should also return documents with the singular “wolf”; a query for “organize” should also return documents concerning “organization”. The decision to extend information retrieval to morphological variants generally improves recall at the expense of precision.

## Lexical relationships

In real-word documents, different words are used to represent the same or similar concepts, and readers are implicitly assumed to have knowledge of common word relationships. A successful information retrieval system must therefore take into account these important relationships, which are ignored under the traditional bag-of-words paradigm:

- Synonyms. A search for movies should also return documents regarding films.
- Antonyms. A search for (NOT big) dogs should also return documents regarding small dogs.
- Hypernyms/Hyponyms (is a kind of). A search



Figure 1

for dogs should return documents about poodles.

- Meronyms/Holonyms (is a part of). A search for gills should return documents regarding animals that use gills.

## Context

Due to the complexity of natural language, the same word often has many different meanings in different contexts. For example, mouse can either refer to a pointing device or a furry rodent, depending on whether it is referred to in conjunction with a computer or a cat. As another example, a bank is either the side of a river or a financial institution. Each distinct use of the same word is referred to as a word sense and can only be distinguished by context. Thus, blindly matching key words in the traditional information retrieval paradigm will lead to poor precision because the approach pays attention to the specific word sense relevant to the user query.

## Syntactic variation

The richness of natural language provides many different ways to convey the same meaning. For example, sentences can be written in either the active or the passive voice:

- Groves of palm and orange trees line the shore.
- The shore is lined with groves of palm and orange trees.

As another example, a fact can be stated in two entirely different ways:

- The population of Taiwan is over 22 million.
- Over 22 million people live in Taiwan.

A successful information retrieval system must handle all these syntactic variations, which are ignored by the bag-of-words paradigm.

## Richness and Expressiveness of Natural Language

In many circumstances, key words are simply not expressive enough to describe the users' information needs. For example, suppose a user wanted to find the answer to the following question, What country in Africa has the largest population? Using a traditional key word search engine, she enters the query "largest AND population AND country AND Africa", in hopes of getting, "Nigeria is the largest country in Africa in terms of population" as an answer.

Instead, she gets the following irrelevant results:

- Of any country in Africa, it has the largest elephant population.
- Its population contains the largest segment

of poor laborers of any country in Africa.

- Half of the city's population gathered to witness the largest funeral ever held in the country, or anywhere in Africa.

This occurs because the answer to her question is stated in a different form:

- Nigeria is the most populous of all African countries.

As another example, suppose a user wished to find out what spiders eat by posing the query "spiders AND eat." He would not only get information about what spiders eat, but what eats spiders, and many other irrelevant results:

- Creepers eat insects and spiders found on and in the crevices of bark.
- Although spiders feed mostly on insects, some spiders capture and eat tadpoles, small frogs, small fish, and mice.
- Enemies of spiders include snakes, frogs, toads, lizards, birds, fish, and other animals that also eat insects.
- Spider crabs are eaten for food, especially in East Asia.

The inability of the bag-of-words paradigm to answer such queries underscores the necessity of understanding the document contents in order to perform successful information retrieval. In most cases, the collection of key words occurring in a document is not expressive enough to convey the meaning of that document.

## Potential Solutions

Given the inadequacy of current information retrieval technology and the invalid assumptions of the bag-of-words paradigm, the following section presents some potential solutions that may better fulfill the user's information access needs.

## Stop words

The most basic technique used in improving the performance of an information retrieval system is to ignore words that do not contribute to the content of a document. In English, these are function words such as and, a, or the.

## Stemming

Stemming recovers the stem, or base form, of a word—this is accomplished by either stripping off suffixes (e.g., cats to cat) or by a static lookup in a word list for irregular forms (e.g., men into man). Stemming operates on the assumption that words with the same stem have similar (or the same) meanings.

The usage of stemming generally leads to more effective information retrieval. Recovering the stem of a word generated through inflection-

## Bibliography

1. Rijsbergen CJ. Information Retrieval; 1979  
{<http://www.dcs.gla.ac.uk/Keith/Preface.html>}
2. Berry MW, Dumais, ST, O'Brien, GW (1995) Using Linear Algebra for Intelligent Information Retrieval. *SIAM Review*. 1995; 37(4) 1995, 573–595.
3. Berry MW, Dumais, ST, Shippy, A. A Case Study of Latent Semantic Indexing; 1995.
4. Deerwester S, Dumais ST, Landauer TK, Furnas GW, Harshman RA. Indexing by Latent Semantic Analysis. *Society for Information Science*; 41(6): 391–407.
5. Cormack, R.M. (1971). A review of classification. *Journal of the Royal Statistical Society, Series A*, 134, 321–353.
6. Dorofeyuk, A.A. (1971). Automatic Classification Algorithms (Review). *Automation and Remote Control*, 32, 1928–1958.
7. WordNet – a Lexical Database for English.

# Precision and Recall—Metrics for Evaluating Information Retrieval System

The effectiveness of an information retrieval system is usually defined in terms of two metrics: precision and recall. Precision is the fraction of documents in a query result that are relevant. Recall is the fraction of all relevant documents that are retrieved in a particular query. These two metrics are theoretically orthogonal—to achieve perfect recall (and very low precision) a system merely has to return all documents in the data collection for any query, and to achieve perfect precision (with no recall) a system merely has to return zero documents for any query. However, in most current-generation information retrieval systems there is an inverse relationship between the two measures. Having tighter search parameters will most likely filter out irrelevant results, thus raising precision, but often has the unfortunate side effect of rejecting relevant results, lowering recall. Implementing looser search parameters will produce the reverse tradeoff of sacrificing precision for recall.

al morphology greatly raises recall at little or no cost to precision. Recovering the stem of a word generated through derivational morphology may raise recall at the expense of recall. Another consideration in stemming is the degree to which it is conducted. For example, should organizational be stemmed to organization or organize? Once again, this choice represents a precision-recall tradeoff. Exhaustive stemming increases recall at the cost of precision.

## Query expansion and word sense disambiguation

A technique to increase the effectiveness of information retrieval systems is query expansion, in which related words (e.g., synonyms, hypernyms) of a key word are added to the query. For example, if the user searches with “movie,” “film” is automatically added.

Query expansion operates on the premise that a database of word relationships exists. A popular tool for this purpose is WordNet,<sup>7</sup> a freely available lexical database providing information regarding the basic word relationships discussed earlier. However, since WordNet aims to be a general-purpose database, it has far too many relationships and word associations for most purposes—leading to low recall. Often, more specific lexical databases must be handcrafted, which is an ex-tremely time-consuming task.

Effective information re-trieval and query expansion can only be achieved by contextual word sense, or determining which sense of a par-

ticular word is being used both in the document and in the query. In response to queries, precision can be improved by determining which word sense the user is requesting. Furthermore, knowledge of the relevant word sense serves as an effective control mechanism for query expansion. For example, determining whether a particular use of bank refers to a river bank or financial institution will either lead to query expansion with “vault,” “money,” “teller,” etc., or “water,” “current,” “river,” etc. Carefully controlled query expansion of the correct word sense can greatly improve the recall of an information retrieval system. However, blindly performing query expansion on every word sense is very dangerous, and may have devastating effects on recall.<sup>8, 9, 10, 11</sup>

Demystifying word sense is usually accomplished by creating co-occurrence profiles for each individual sense. For example, when cat occurs in the vicinity of mouse, the latter word probably refers to a furry rodent, but when computer occurs in the vicinity of mouse, the latter word most likely refers to a pointing device. These co-occurrence profiles are a method of representing the context of a word, and can be learned semiautomatically through statistical and machine learning techniques.

## Adding natural language

A major fault with the bag-of-words paradigm is that the approach ignores the richness and complexity of natural language. Information retrieval systems cannot achieve high precision and recall without (at least some fundamental) understanding of the documents. Correspondingly, natural language is the best query method for information access. Under such a scheme, information seekers would formulate their queries in a natural language, such as English, and receive the exact answer to their questions without the need for any additional browsing. Compared to current search interfaces that depend on complex Boolean operators, natural language is far more intuitive and easy to use. As we have demonstrated above, there is no way to express the query “What country in Africa has the largest population” using Boolean operators under a key word-based formulation. Thus, a successful information retrieval system must integrate natural language technology to understand the documents and query.

Developing practical natural language-understanding technology and suitable representational structures are the two major bottlenecks in developing the next generation of information retrieval systems. Practical is the operative word in melding these two key

8. Gonzalo J, Verdejo F, Chugur I, Ciagarran. Indexing with WordNet synsets can improve text retrieval. Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems; 1998.

9. Richardson R, Smeaton A F. Using WordNet in a knowledge-based approach to information retrieval. Proceedings of the 17th BCS/IRSG Colloquium on Information Retrieval; 1995.

10. Voorhees E M. Query expansion using lexical-semantic relations. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 61-69. London: Springer-Verlag; 1994.

11. Voorhees E M. Using Wordnet for Text Retrieval. Wordnet, an Electronic Lexical Database. Cambridge, Massachusetts: MIT Press; 1998.

technologies. Current-generation natural language systems have yet to reach sufficient maturity to parse real-word documents, and such a system is not likely to be available any time in the near future. Thus, even state-of-the-art systems can only produce a partial understanding of documents, at best. Furthermore, storing and retrieving the richness of natural language relations in a suitable representation becomes a serious computational problem, because the description of a document may potentially be larger than the document itself.

START is a natural language question-answering system developed at the MIT Artificial Intelligence Laboratory that has been answering real-word English questions on the World Wide Web since 1993. The system attempts to serve as a common natural language interface to knowledge available on the World Wide Web. START attempts to overcome the limitations of current technology with natural language annotations, which are collections of natural language sentences and phrases, simple enough for the computer to analyze, that describe the contents of various information segments.<sup>12, 13, 14</sup> These segments are not limited to text and can include images and other multimedia content. START analyzes and matches natural English queries with these annotations, which contain pointers back to the information segments they summarize. If an annotation matches a query, then the original information segment is returned to the user as the answer. With this mechanism, natural language information retrieval can be accomplished without understanding the entire data document set. However, no method exists to automatically create annotations for documents that contain arbitrarily complex text and multimedia content. Presently, these annotations must be generated manually, after a human reads over the target text. Although generalization mechanisms allow application of a single annotation over an entire class of knowledge through parameterization, each data source must be integrated manually, which is still relatively time consuming.

What is in store for the future? Information access can only be successful on a large scale (e.g., the World Wide Web) if it is accomplished automatically, without human assistance. Thus, there is a growing trend towards reducing or totally eliminating human involvement. Although the START system performs marvelously as a natural language information access interface, its domain is relatively limited, and extending its knowledge base is a tedious task. Current research in the field has focused on linguistically

motivated information retrieval,<sup>15, 16</sup> which attempts to perform partial parses of the documents and then store those parses in a simplified form suitable for retrieval. Experiments utilizing this approach have yielded limited success, as this is one of the frontiers of information retrieval. However, the integration of natural language and information retrieval is well motivated, and remains one of the most interesting and promising areas of research today.

## Conclusions

The ability to perform natural language information retrieval on the World Wide Web is not a trivial application of natural language information retrieval technology. The Web—a chaotic mosaic of not only text fragments, images, and other multimedia segments—requires the integration of other technologies such as image recognition, pattern matching, signal processing, and voice recognition in order to handle the multimedia contents. Furthermore, an implicit assumption made in the above discussion is the well-edited structure and grammatical correctness of text documents, which is certainly not true of many Web documents that contain short, choppy, and poorly written segments.

A complete solution to the general information access problem will require a multidisciplinary approach and simultaneous breakthroughs in the field of natural language processing and other fields. Applying this technology to the World Wide Web will require the integration of even more technologies, but the dream is tantalizing: the ability to intuitively access any piece of knowledge available in the continually growing global online community will revolutionize the way humans view knowledge itself. It is the actual realization of the cliché, all of the world's knowledge at your fingertips.

12. Katz B. Using English for Indexing and Retrieving. *Artificial Intelligence at MIT: Expanding Frontiers*. In P. H. Winston and S. A. Shellard, editors. Cambridge, MA: MIT Press; 1990.
13. Katz B. From Sentence Processing to Information Access on the World Wide Web. *AAAI Spring Symposium on Natural Language Processing for the World Wide Web*; 1997.
14. Katz B. Annotating the World Wide Web Using Natural Language. *Proceedings of the 5th RIAO Conference of Computer Assisted Information Searching on the Internet*; 1997.
15. Arampatzis A, van der Weide T P, Koster CHA, van Bommel P. An Evaluation of Linguistically Motivated Indexing Schemes. *Proceedings of the BCS-IRSG 2000 Colloquium on IR Research*; 2000.

## Recommended Readings

16. Arampatzis A, van der Weide TP, Koster CHA, van Bommel P. Linguistically Motivated Information Retrieval. *Encyclopedia of Library and Information Science*, New York, NY: Marcel Dekker, Inc.; 2000.