The Weaknesses of Full-Text Searching

by Jeffrey Beall

This paper provides a theoretical critique of the deficiencies of full-text searching in academic library databases. Because full-text searching relies on matching words in a search query with words in online resources, it is an inefficient method of finding information in a database. This matching fails to retrieve synonyms, and it also retrieves unwanted homonyms. Numerous other problems also make full-text searching an ineffective information retrieval tool. Academic libraries purchase and subscribe to numerous proprietary databases, many of which rely on full-text searching for access and discovery. An understanding of the weaknesses of full-text searching is needed to evaluate the search and discovery capabilities of academic library databases.

Jeffrey Beall is a Metadata Librarian/Assistant Professor, Auraria Library, University of Colorado Denver, 1100 Lawrence Street, Denver, CO 80204, USA <jeffrey.beall@ucdenver.edu>.

Introduction

Definition of Full-Text Searching

Full-text searching is the type of search a computer performs when it matches terms in a search query with terms in individual documents in a database and ranks the results algorithmically. This type of searching is ubiquitous on the Internet and includes the type of natural language search we typically find in commercial search engines, Web site search boxes, and in many proprietary databases. The term full-text searching has several synonyms and variations, including keyword searching, algorithmic searching, stochastic searching, and probabilistic searching.

Metadata-Enabled Searching

There is one other main type of online searching. This is metadata-enabled searching, which is also called deterministic searching. In this type of search, searchers pre-select and search individual facets of an information resource, such as author, title, and subject. In this type of search, the system matches terms in the search with terms in structured metadata and generates results, often a browse display sorted alphanumerically. Author, title, and subject searches in online library catalogs are examples of this type of search.

Importance

Understanding the weaknesses of full-text searching is important for academic libraries for several reasons. First, academic libraries purchase or subscribe to numerous proprietary databases, including many full-text databases. When they decide whether to pay for a particular database, libraries need to evaluate the search engine or system that accompanies the database. When these databases provide only full-text searching and not metadata-enabled searching, resource discovery within the resource may be difficult, putting libraries in the position of paying for content that is hard to find. Library-created databases, such as institutional repositories, are another area where an understanding of the weaknesses of full-text searching is needed. Providing only full-text access to a library's digital objects may not provide resource discovery of sufficient quality for the collection's users. Academic libraries need to evaluate these collections and the available search engines and systems and select the best one for their particular databases. Finally, much current debate centers on the need for online library catalogs versus the ability to access academic library materials through a commercial search engine. A thorough knowledge of the weaknesses of full-text searching adds to the debate and helps academic librarians in the evaluation, recommendation and design of library database search engines.

Objective

The purpose of this article is to list and describe the chief weaknesses of full-text searching. We limit the scope of this article to true full-text searching that automatically matches words entered in the search box with words in resources a database contains to generate results. This study does not include in its analysis new, semantic search engines such as Hakia, which stores metadata for each Web page indexed and uses that metadata, along with word matching, to generate search results. Indeed, many popular search engines do incorporate metadata into their searches. For example, the Google advanced search allows for limiting search results to a specific language. This search limit is generated by language metadata that the search engine assigns to each Web page it indexes (the accuracy of this automatically-generated language metadata may not always be high).

Still, the great majority of the searches performed on the Internet are of the type this paper seeks to study: full-text searching that matches words in a search box with words in online documents or online text. This study is not a comparison of full-text searching and metadata-enabled searching. Both of these two types of searching have their various strengths and weaknesses. This article seeks chiefly to describe the weaknesses of full-text searching.

This paper is a theoretical critique of full-text searching and focuses on the type of searching done in academic libraries. It describes and categorizes the ways in which full-text searching can fail, failures that most searchers have likely encountered themselves. While outside the scope of this paper, quantitative research that measures the extent of these problems would be valuable and would further inform the debate.

Previous Studies

Most information retrieval and information discovery has transitioned from searching dominated by metadata-enabled searching (academic library card catalogs) to the present full-text or algorithmic searching (Web search engines). This transition occurred without sufficient analysis of the weaknesses of full-text searching. Perhaps if searchers understood the number of resources they were missing because of full-text searching's reliance on word matching to generate retrieval, they would be less satisfied with it. Generally, books and articles on information retrieval often cite one or two examples of the weaknesses of full-text searching; few have been comprehensive in their analyses, as this one seeks to be.

Among those to write about the weaknesses of full-text searching is Thomas Mann, a reference librarian at the Library of Congress. He states "Keyword searching fails to map the taxonomies that alert researchers to unanticipated aspects of their subjects. It fails to retrieve literature that uses keywords other than those the researcher can specify; it misses not only synonyms and variant phrases but also all relevant works in foreign languages. Searching by keywords is not the same as searching by conceptual categories".[1] Here he makes reference to the synonym problem in full-text searching (and he prefers to use the term keyword searching rather than full-text searching, providing yet another example of the synonym problem). Mann also states,

> When all is said and done, keyword searching necessarily entails the problem of the unpredictability of the many variant ways the same subject can be expressed, within a single language ("capital punishment", "death penalty") and across multiple languages ("peine de mort", "pena capitale"). And no software algorithm will solve this problem when it is confined to dealing with only the actual words that it can retrieve from within the given documents (or citations or abstracts) themselves.[2]

Beall[3,4] presents two brief but more complete looks at the problems of full-text searching. The present paper aims for a more comprehensive analysis. Moreover, Beall[5] introduces the term "search fatigue" to describe the feelings of frustration searchers feel when they are unsuccessful in finding information due to the weaknesses of full-text searching. A recent study by Hemminger, Saelim, Sullivan, and Vision[6] compares full-text searching to metadata searching and finds that "it may be time to make the transition to direct full-text searching as the standard". However, later in the article the authors concede that their study may not be truly representative, for it compared the two searching modes using gene names, which are consistently used in the literature they studied.

The Weaknesses of Full-Text Searching

The Synonym Problem

Perhaps the biggest and most pervasive weakness of full-text searching is the synonym problem. This problem occurs because there is often more than one way to name or express a given concept, such as a person, place, or thing. There are several different aspects of the synonym problem.

"Perhaps the biggest and most pervasive weakness of full-text searching is the synonym problem."

True Synonyms

Synonyms are two words that mean the same thing in one language. In full-text searching, synonyms hinder effective information retrieval when a searcher enters a term in the search box and the system only returns results that match the term and does not return results that refer to the concept only by one of its synonyms. For example, if a searcher seeks information on leprosy, he would likely enter "leprosy" in the search box and expect complete results. However, some online documents refer to this disease as "Hansen's disease". While it's true that many documents will contain both terms, thus enabling access regardless of which term is searched, a certain percentage of the documents will only contain one term, thus providing an incomplete retrieval.

Variant Spellings

Words that mean the exact same thing can sometimes be spelled differently, as in variant British and American spellings. In full-text searching, a search for "harbour" will miss results that use the spelling "harbor". It is true that many full-text search engines have developed methods for overcoming this problem; searchers can use wild card or truncation operators to retrieve multiple spellings of a word. But there are also variant spellings within a single dialect of a language, and these differences are often beyond the scope of the truncation or wild card operators. For example, in American English the spellings "donut" and "doughnut" are both common. Unlike the case of synonyms, where in a single document both synonyms may appear, spelling tends to be consistent within a document. A document about harbors written in the United States is unlikely to also contain the spelling "harbour". This means that there is a smaller chance of retrieving documents with variant spellings than there is with true synonyms.

Shortened Forms of Terms

Abbreviations, acronyms, and initialisms can hinder recall in full-text search systems because a document may contain only the short form of the word or only the long form. When this occurs, someone searching on the short form (PETA) will miss in his retrieval documents that only use the long form (People for the Ethical Treatment of Animals). Alternately, searching on the long form of the term, like Magnetic Resonance Spectroscopy, will miss documents that only refer to the concept by its short form, MRS.

Different Languages or Dialects

When searching a term in one language, a searcher will not match documents that contain the foreign-language version for that concept, unless the two terms happen to be cognates. For example, if you search the term "butter", the search will miss documents that only refer to this by its Spanish equivalent "mantequilla". For many searchers, this exclusion is not a problem; they prefer their search results to be in just one language. However, scholarship supported by academic libraries, such as medical research, or research for a thesis or dissertation, needs to be comprehensive regardless of language. Additionally, variation occurs within a single language. The phrase "football coach" means different things in British and American English. In the United States, this term refers to a person who directs an American football team, that is, the coach. In British English, a "football coach" refers to a bus (motorcoach) for soccer players.

"The phrase "football coach" means different things in British and American English."

When the words are the same in two or more languages or dialects of a single language, however, such as the word "migration", which means the same in English and French, the different language problem does not occur.

Obsolete Terms

Linguistic change can also prevent complete information retrieval in full-text searching. For example, the phrase "French distemper" is one of many archaic ways of referring to syphilis (the term was also used metaphorically by the English to refer to the French Revolution). Someone researching the history of syphilis and using full-text searching would miss resources that only use the term "French distemper". It is possible in Google Books to find digitized academic library books that only use this term. While it is possible to search every possible variant term to generate a complete search result in full-text searching, this method is not very efficient and requires that one know all the variant terms, an unlikely possibility.

Humanities vs. STM

Overall, despite the above example, we should note that the synonym problem probably occurs more frequently in the humanities than it does in science, technology, and medicine. STM scholarship tends to be more consistent in its terminology, even across languages. For example, the scientific names of plants and animals (binominal nomenclature) are the same in most languages (*Tyrannosaurus rex*, for example). This tendency to use a standard terminology even across languages ameliorates the synonym problem in these fields. (Note, however, that *Tyrannosaurus rex* is often abbreviated to T-rex, creating an instance of the abbreviation problem described above.) This is not to say that STM fields always use consistent terminology. There are at least sixty different terms that all mean "Atlantic cod", for example.[7] The variation occurs in the common names and not in the scientific names, though. While scientific names tend to be applied consistently within the scientific domain, popular terms for natural things reflect a diverse terminology.

Unlike scientific terminology, humanities terminology varies significantly from one language to another and by time and dialect within a single language. Take the term "short stories" for example. In French it's "nouvelles", in Spanish it's "cuentos", and in German, "Erzählungen". The names for languages themselves differ from language to language too. The names for the German language include alemán, Deutsch, and allemand. Perhaps one area in the humanities where there is some cross-language consistency is music. Many languages share terms like "soprano". Also, as described earlier, regional differences within a single language can lead to problems in information retrieval when using full-text searching. In British English a "solicitor" is a lawyer; in American English, it is someone who goes door to door selling something or asking for contributions for charity.

The Homonym Problem

The homonym problem occurs in full-text searching when a single word or phrase has more than one meaning. Because full-text searching relies on word matching to generate results, a search for a term with several meanings will retrieve documents for all of the meanings, rather than just the one the searcher wants. Homonyms are perhaps the chief cause of low search precision.

True Homonyms

Without metadata, computers do not know the sense of each of a given pair of homonyms. That is, computers cannot effectively disambiguate two concepts when they are called by the same term. For example, a search on "cookies" will pull up documents both about the food and the little files stored on a computer. Searchers are aware of this problem, for it occurs frequently. Many have developed strategies to eliminate unwanted hits and increase the probability of search results matching the particular meaning of the homonym they seek. For example, someone looking for information on computer file cookies might add the word "computer" to the search terms (instead of only searching for "cookies"), because the documents about edible cookies are less likely to have this term in them. Alternatively, a sophisticated searcher might use the "not" operator to try to eliminate unwanted homonyms and increase a search's precision. The searcher might enter "cookies not recipes", for example. While these strategies help, they are not completely effective. Words can have many more meanings than just two, and one often does not anticipate that a search term has synonyms.

Disambiguation of Personal Names

This problem occurs in both full-text searching and in metadata-enabled search systems where the practice of name disambiguation is not employed. Name disambiguation is the process of making each person's name unique in a database. The more common a name in a database, the greater the problem. The problem is made worse by names that also function as other parts of speech, like bill, April, miller, and mike. Because name disambiguation necessarily involves adding metadata, virtually all full-text documents lack this value-added feature. This problem is significant in academic libraries because some style guides prescribe the use of initials instead of given names in citations, making a full-text search for an author's name more difficult.

False Cognates

These are two words that are spelled the same (or almost the same) in two languages but, deceptively, do not mean the same thing. In full-text searching, false cognates are only a problem when they are spelled exactly the same. The problem occurs when a word entered into a search box happens to match a word in a different language that has no semantic relationship to the original search term. For example, the word "location" in French doesn't mean "location" in English; it means a rental or a lease.

Inability to Search by Facets

Sometimes searchers have a need to search by only a specific characteristic or attribute of an online resource, such as author, title, subject, date of creation, etc. These attributes, or facets, help to cluster resources by specific shared characteristics. Clustering, or collocating, is helpful in information retrieval because it helps exclude unwanted resources from a search. Also, clustering matches typical searcher queries, such as "I want all DVDs on agriculture", or "I want all PDF files on land use planning in Utah published before 2000". Pure, full-text searching fails at these tasks, because the search engine doesn't know the format (DVD's) or the subject (agriculture) or the publication date (2000) of the documents it searches. If a search engine does know these dates, then it's not a pure, full-text search engine. Instead, it is a metadata-enhanced search engine and draws its ability to sort by facets from metadata assigned to each resource it indexes.

Clustering

Clustering is most helpful when it attempts to solve the homonym problem in subject searches. Here, clustering is the process of grouping and separating out resources by subject. For example someone searching for information on ocean banks might just enter "banks" as the search term. A search engine with the ability to cluster would then separate out the results that refer to ocean banks from those that refer to banks, the financial institutions. It's probably not uncommon for users who stumble on the homonym problem in a full-text database to do a revised search that includes a second search term, as a strategy for eliminating unwanted documents. For example, a searcher could enter "banks ocean" to eliminate documents in the retrieval that are about banks the financial institutions. This stratagem is not foolproof, however, for there are many resources about financial institutions that contain the words "banks" and "ocean". Increasingly, proprietary databases are performing this type of cluster analysis algorithmically, but with limited success.

Inability to Sort

Just as full-text search engines lack the ability to cluster search results, they also lack the ability to sort results by facets. Sorting plays an important role in and can increase the value of information retrieval because it helps arrange search results in a meaningful and predictable order. For example, sorting search results by publication date (oldest first or most recent first) is helpful to searchers looking for recent or old publications. Traditional full-text search engines cannot perform this type of sort because they don't know the publication dates of the documents in the database. Search engines that do have the ability to sort by publication date are actually using metadata to do the sort and are not true, full-text search engines. The alphanumeric sorting of information resources' other main facets, author, title, and subject, also adds great value to information discovery and retrieval, but a true, full-text search engine cannot perform this function. Relevance ranking, a sorting system based on stochastic analysis, works well when the resource a searcher desires falls on the first screen of search retrieval display, but, increasingly, this is seldom the case in full-text search engines.

Spamming

This problem is limited to open databases, such as the Internet, where anyone can upload data that becomes part of the database. In this context, spamming refers to the addition of text to cause documents, such as Web pages, to appear in search results gratuitously. This is sometimes called "keyword stuffing".[8] The result is that irrelevant search results appear. Most of the major Web search engines are sufficiently able to deal with this problem algorithmically and strategically, but at a cost. Most big search engines ignore subject metadata (often referred to here as "keywords") added into a document's meta tags for fear that it is spam. Brooks summarizes:

> We are now just learning that the Web has a different social dynamic. The Web is not a benign, socially cooperative environment, but an aggressive, competitive arena where authors seek to promote their Web content, even by abusing topical metadata. As a result, Web crawlers must act in self defense and regard all keywords and topical metadata as spam.[9]

Thus this potential added value, that is, the value of rich subject metadata, is often lost in the jungle of the World Wide Web.

"Just because a Web site contains a word doesn't mean it's about whatever concept that word names."

The Aboutness Problem

Language and words do not always convey what a resource is about. Just because a Web site contains a word doesn't mean it's about whatever concept that word names. But because full-text search engines rely on word matching to guess at aboutness, searches for information on a topic often fail. Online documents do an inadequate job of providing their own subject metadata. "A classical problem for document retrieval systems is the failure of keywords to identify the conceptual content of documents".[10] Searchers have an idea of what information they want to find in their minds. They express this idea as search terms. The problem is that language is often an ineffective means of unambiguously and clearly stating an information need. Garrett summarizes that, "an extraordinarily subtle and intricate process relates speaker meanings to language output in all natural (i.e., human) language. Individual words and even complete sentences therefore do not necessarily map one-to-one to phenomena of the world".[11] Further, systems that rely on resources' titles for subject analysis are often unsuccessful, for titles frequently fail to describe a resource's content. For example, some books and articles have clever titles that are designed more to attract attention than to describe the content. Indeed, some titles fail to describe the content at all.

Figurative Language

Figurative language also impedes effective information retrieval in full-text searching. Figurative language is language that is not used according to the literal, dictionary definition of the words used. For example, the sentence, "She's drowning in birthday presents" invokes figurative language, in this case a metaphor. The word "drowning" is not used in a literal sense. But a searcher looking for information on drowning in a full-text database would retrieve the document with this sentence among the search results. In the looking-glass world of full-text search engines, all metaphors become real.

Word Lists

Individual entries in online dictionaries, glossaries, and word lists also often match search terms in full-text databases and appear in search results. Such lists are the source of many "false drops" or irrelevant hits in full-text searches. Mann says,

> The Google Print project will be hampered by a further problem: its scanned 15,000,000 books will include tens of thousands of dictionaries. Any keyword searched will thus retrieve all dictionaries in which the word appears—nor could results be "progressively refined" by adding more words because those words, too, will "hit" in the same dictionaries. (This is already a problem for researchers using a much smaller full-text database, the Evans Early American Imprints.).[12]

Of course, for some searches, the word list may be exactly what the searcher desires. But for the tens of thousands of times the list is not what the searcher desires, the search results will amount to little more than noise.

Abstract Topics

It's difficult to search successfully for documents on abstract topics in full-text databases. Subjects such as "health", "free will", and "ethics" generate large retrievals in Web search engines, decreasing the probability that the first few screens of search results will contain documents useful to the searcher. First-year university and college students, who are sometimes unable to narrow their searches, often encounter this problem in academic library databases.

The Incognito Problem

This problem refers to a person, place or thing not being called by its standard name in at least one step of the search process. Specifically, in order to retrieve information in a full-text database, both the terms the searcher enters in the search box and the terms in the best resources have to match. To understand this, it's important to understand that searching is a process that involves several steps. The specious statement "Only librarians like to search; everyone else likes to find"[13] displays an ignorance of information discovery and retrieval as a process. That is to say, there's more to the process than just the last step, finding.

"The specious statement "Only librarians like to search; everyone else likes to find" displays an ignorance of information discovery and retrieval as a process."

Search Terms Not in Resource

It's not uncommon for a document to describe something and fail to name it. Thus, searches for the concept will not retrieve the resources that do not match the term. Garrett shares this example: "Michel Foucault's foundational work on meaning and signifying, The Order of Things, can be said to be all about the French Revolution, and yet it's possible–and I haven't checked–that the word string "French Revolution" does not occur a single time in the entire book".[14] Also, Batty reports, "the golfer Arnold Palmer hit two holes-in-one on the same hole on two successive days in a major tournament, and the article describing this unprecedented feat never mentioned the word GOLF".[15]

Searcher Doesn't Know Term

Frequently the searcher is the source of the problem in full-text searching. When a searcher does not know the correct term for a concept, it can be very difficult for the searcher to find desired information. For example, chondromalacia is a painful medical problem involving the cartilage of the knee joint. A person with a sore knee seeking information about the problem might have this precise condition but not know the name of it. In this case, the person will likely look for information using a Web search engine but will only search with terms such as "pain" and "knee". It's true that the Web sites themselves will probably help deal with this problem. There are Web sites that will be retrieved that will describe one form of knee pain as chondromalacia. Then the searcher, provided he can make the connection, is able to do a second, more precise search on the exact term.

"If a searcher does not know a resource exists, he will not know when a full-text search fails to include it in the search results. This often leads searchers into an endless and exhausting search for information."

Searcher ignorance brings up another point: the insidious nature of full-text search engines. If a searcher does not know a resource exists, he will not know when a full-text search fails to include it in the search results. This often leads searchers into an endless and exhausting search for information.

Non-Textual Resources

Full-text search engines are only able to read and index textual information. In the absence of metadata, objects such as pictures, sound files, video files, etc. are not indexed and therefore not searchable, even though they might contain valuable information.

Difficult-to-Search Paired Topics

Because full-text searching lacks pre-coordination, it is frequently difficult to search paired topics in full-text search engines. Pre-coordination is the assigning of subject metadata into strings or phrases that reflect a summary of a resource's content. Often, resources describe or present information on two topics in relation to each other. Here are some examples of paired topics:

Art and mental illness

Architecture and philosophy

Movies and fascism

Libraries and German Americans

Searching paired topics such as these in a full-text database is problematic however. A full-text search matches documents that happen to contain both terms. Frequently a search on two topics will retrieve resources that do not in fact discuss the two concepts in relation to each other; the resources merely happen to contain both terms. The ability to sort out only the resources that describe the relationship between the two topics is a valuable one, but full-text searching performs poorly at this task.

Variability Among Different Search Engines

Web search engines are big business, and like makes of automobiles, each one is a little different. Moreover, there are hundreds of proprietary databases, each with its own full-text search methods. Individual Web pages also often offer a full-text search box, and these also differ greatly from one resource to another.

Lack of Standardization in Searching

This variability means that searchers have to learn the best way to search for each database that they search. For example, is the default Boolean operator in a given database "and" or "or"? Or does the search engine search all the terms as a phrase? In each case, in order to ensure effective retrieval, the searcher has to know or learn the particular search rules for the database. Probably many searchers assume that all simple search boxes work the same way as the Google search works, but this is not always the case.

Variability in Result Ranking

Searchers also have to adjust to the different ways that full-text search engines rank search results. The so-called relevancy rankings that are frequently found in Web search engines are created according to proprietary algorithms. Thus ranking differs from one system to another. Some less sophisticated full-text search engines might sort by some other aspect rather than a probabilistic calculation of relevancy, such as date the resource was first added to the database, or date the file was last updated.

Indexing Differences

Search engines index text differently. One example is hyphenated words. Different search engines might index the hyphen in "full-text" as a space, or they might ignore the hyphen and index the phrase as "fulltext". These differences require searchers to be aware of each search engine's indexing and searching policies to ensure complete search results. Some search engines might get around this problem by indexing hyphenated text both ways, that is, both with and without a space. But this would not resolve the problem of a resource that does not use a hyphen in a compound word (as in "nonstandard") and a searcher who searches it as two words (as in "non standard").

Brooks describes the problem of digitized text that includes words originally broken at the end of a line of print with a hyphen. He cites such examples as "'Europeans' broken into *Europe* and *ans*, 'distinguishing' broken into *distingu* and *ishing*, 'occurred' broken into *occ* and *urred* ... ".[16] Each of these cases could represent a failed search in a full-text search system. This problem will only worsen as more print works are digitized.

Brooks[17] also describes the problem of stopwords in full-text indexing and searching. Stopwords are short and common words that normally carry little meaning in a document. Many full-text databases contain a list of stopwords that are not indexed in their systems. One problem is that each database generally has its own unique list of stopwords, and another problem is that these words do occasionally carry substantive information. Two examples include the word "in" in the phrase "mother-in-law", and the word "a" as in vitamin A. When these words are not indexed in a given system, retrieval on "mother-in-law" and "vitamin A" would be made much more difficult.

The Opaque Web

There is a great amount of information available on the Internet that is hidden behind search interfaces, including those in many academic library databases. This information is opaque, or invisible to search engines. Henzinger states, "A plethora of content is stored in databases rather than in typical Web pages. The pages as well as their URLs are created in response to a user filling out a form on the Web. Because search engines are unable to emulate this behavior, such dynamically generated pages cannot be indexed. There has been some research on trying to make form-filling automatic, but the problem remains largely unsolved".[18]

A recent report written by two government watchdog groups, OMB Watch and the Center for Democracy and Technology, bemoans the amount of valuable United States government information that is hidden behind search interfaces and therefore not indexed in the popular search engines. The report states, "Unfortunately, many important information sources within the federal government are essentially hidden from the very search engines that the public is most likely to use".[19] The report also states, "Often the agencies mentioned operate tens or hundreds of dynamic databases that cannot be indexed and searched".[20] One example of such a database familiar to librarians is the Library of Congress Authorities Web site. This Web site contains a search interface that leads to hundreds of thousands of name, title, and subject authority records. These records contain a great deal of valuable biographical, geographical, and subject information, but because Web search engines can't see them, the information they contain is not accessible without knowing in advance about the Web site and accessing it directly.

> **"A recent report written by two government watchdog groups, OMB Watch and the Center for Democracy and Technology, bemoans the amount of valuable United States government information that is hidden behind search interfaces and therefore not indexed in the popular search engines."**

Further Research

Research that measures the deficiencies of full-text searching would provide valuable information. For example, research that studies the synonym problem could measure the proportion of resources missed when a library patron searches a word and fails to retrieve in the search resources that only refer to the concept being searched by its synonyms. In addition, research that compares the weaknesses of full-text searching in the humanities versus STM would prove valuable, especially if it could quantify in which of these two areas of study full-text searching is a greater hindrance to information access.

Conclusion

Linguistic problems, the limitations of full-text search engines, and missing data combine to make full-text searching unreliable, incomplete, and insidiously imprecise, especially for serious information seeking, such as scholarly research. Many Web-based applications still use basic full-text searching as their chief information retrieval mechanism. Over the past fifteen years, most information retrieval has transitioned from searching based on rich metadata to full-text searching. The result of this transition is an overall decrease in the quality of information retrieval. Academic librarians need to understand the weaknesses of full-text searching to better evaluate the search engines in databases that libraries purchase and create.

Notes and References

1. Thomas Mann Will Google's Keyword Searching Eliminate the Need for LC Cataloging and Classification? (2005). Available: http://www.guild2910.org/searching.htm (Jan. 21, 2008).
2. Thomas Mann, The Oxford Guide to Library Research (Oxford: Oxford University Press, 2005) p. 102.
3. Jeffrey Beall, "The death of metadata," *The Serials Librarian,* 51 (2006): 55–74.
4. Jeffrey Beall, "The death of full-text searching," *PNLA quarterly,* 70 (Winter, 2006): 5–6.
5. Jeffrey Beall, "Search fatigue: finding a cure for the database blues," *American Libraries,* 38 (March, 2007): 46–50.
6. Bredley M. Hemminger, Billy Saelim, Patrick F. Sullivan, & Todd J. Vision, "Comparison of full-text searching to metadata searching for genes in two biomedical literature cohorts," *Journal of the American Society for Information Science & Technology* 58 (2007): 2341–2352.
7. Beall, Search fatigue.
8. Monika Henzinger, "Search Technologies for the Internet," *Science,* 27 (July 2007): 468–471.
9. Terrence Brooks, "Web Search: how the Web has changed information retrieval," *Information Research,* 8.3 (2003). Online. Available: http://InformationR.net/ir/8-3/paper154.html (Jan. 21, 2008).
10. Kai A. Olsen, Kenneth M. Sochats, & James G. Williams, "Full Text Searching and Information Overload," *International Information & Library Review* 30 (June, 1998): 105–122.
11. Jeffrey Garrett, "KWIC and Dirty? Human Cognition and the Claims of Full-Text Searching," *The Journal of Electronic Publishing* 9 (Winter 2006). Online. Available: http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0009.106 (Jan. 21, 2008).
12. Mann, "Will Google's Keyword Searching..."
13. Roy Tennant (2001). "Avoiding Unintended Consequences," *Library Journal* 126 (Jan. 1, 2001): 38. Online. Available: http://www.libraryjournal.com/article/CA156524.html (Jan 21, 2008).
14. Garrett, "KWIC and Dirty?"
15. David Batty "WWW—Wealth, weariness or waste: Controlled vocabulary and thesauri in support of online information access," *D-Lib Magazine* (November, 1998). Online. Available: http://www.dlib.org/dlib/november98/11batty.html (Jan. 21, 2008).
16. Terrence Brooks, 1998. "Orthography as a fundamental impediment to online information retrieval," *Journal of the American Society for Information Science* 49 (1998): 731–741.
17. Ibid.
18. Henzinger, "Search Technologies for the Internet," p. 469.
19. OMB Watch and Center for Democracy & Technology, Hiding in plain sight: Why Important Government Information Cannot Be Found through Commercial Search Engines (2007). Available: http://www.cdt.org/righttoknow/search/Searchability.pdf (Jan. 21, 2008).
20. Ibid.