PERGAMON

Information Processing and Management 38 (2002) 273–291

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

The challenge of commercial document retrieval, Part I: Major issues, and a framework based on search exhaustivity, determinacy of representation and document collection size

David C. Blair

Computer and Information Systems, Graduate School of Business, The University of Michigan, Ann Arbor, MI 48109-1234, USA

Received 4 February 2000; accepted 19 January 2001

Abstract

With the growing focus on what is collectively known as ''knowledge management'', a shift continues to take place in commercial information system development: a shift away from the well-understood data retrieval/database model, to the more complex and challenging development of commercial document/information retrieval models. While document retrieval has had a long and rich legacy of research, its impact on commercial applications has been modest. At the enterprise level most large organizations have little understanding of, or commitment to, high quality document access and management. Part of the reason for this is that we still do not have a good framework for understanding the major factors which affect the performance of large-scale corporate document retrieval systems. The thesis of this discussion is that document retrieval – specifically, access to intellectual content – is a complex process which is most strongly influenced by three factors: the size of the document collection; the type of search (exhaustive, existence or sample); and, the determinacy of document representation. Collectively, these factors can be used to provide a useful framework for, or taxonomy of, document retrieval, and highlight some of the fundamental issues facing the design and development of commercial document retrieval systems. This is the first of a series of three articles. Part II (D.C. Blair, The challenge of commercial document retrieval. Part II. A strategy for document searching based on identifiable document partitions, Information Processing and Management, 2001b, this issue) will discuss the implications of this framework for search strategy, and Part III (D.C. Blair, Some thoughts on the reported results of Text REtrieval Conference (TREC), Information Processing and Management, 2002, forthcoming) will consider the importance of the TREC results for our understanding of operating information retrieval systems.

E-mail address: dcblair@umich.edu (D.C. Blair).

0306-4573/01/$ - see front matter
PII: S0306-4573(01)00024-3

1. Introduction

A radical shift began to take place in the 1990s in the focus of commercial information processing software development: a shift away from the better-understood data retrieval/database model, and to the more complex and challenging development of commercial document retrieval models. This shift in focus came none too soon: management theorist Peter Drucker has stated that we have entered the third major evolution in the "... concept and structure of organizations... the shift from the command-and-control organization, the organization of departments and divisions, to the information-based organization, the organization of knowledge specialists." (Drucker, 1988). Much of this organizational information, perhaps most of it, takes the form of documents (e.g., reports, messages, letters, journal and magazine articles, memos, minutes of meetings, research bulletins, etc.), and it is easy to see why documents are important: they are often the organizing and interpretive medium that gives data, figures and other information meaning within an organizational context. In short, documents are the medium where organizational memory or intelligence resides. Poor access to document content means poor access to the knowledge that an organization creates or acquires.

Interleaf has estimated that more than 1 billion documents are being created each day in North America, and that executives spend 40% of their time dealing directly with documents. The Gartner Group has estimated that as much as 90% of a corporation's information is contained in its documents. This brings a new urgency to the task of providing access to this enormous, and growing, volume of document-based information: lawsuits are lost or settled out of court because supporting documents cannot be found; internal studies and analyses are redone because the documented results of the original work cannot be located; managers make decisions that are suboptimal or even incorrect because they are not aware of important relevant information that exists in other parts of the organization or is available from commercial document databases; scarce grant money is allocated to research that has already been done and published because neither the authors of the grant nor the referees are aware that similar work has already been completed. The consequences of poor document retrieval can be striking:

> The effects of a failure of "document control" can be dramatic. A major utility company was required to shut down four nuclear reactors because of lost repair instructions (at a loss of $2m per day). The US Department of Defense estimates that half of all military accidents result from missing or inaccurate technical information. A major airline was fined $10k per take-off because of out-of-date maintenance information. A major drug company lost its entire R&D investment owing to inability to provide timely documentation (Fleischer, 1990).

Clearly document, or text, retrieval has taken a prominent place in commercial information system development. But are we ready for this shift? Do we have the software tools to build large-scale commercial document retrieval systems, and, more importantly, do we have a good conceptual understanding of what factors influence document retrieval in the corporate context? A clear "yes" cannot be given. Certainly, there has been a rich history of theoretical work in document retrieval (Salton and McGill, 1983; van Rijsbergen, 1979) and a history of small-system tests. But tests of actual large-scale commercial applications of this theory are still rare, and the

market-place has been saturated with systems whose theoretical antecedents date back to the 1950s. [The periodic Text REtrieval Conferences (TRECs) have attempted to evaluate the retrieval effectiveness of comparatively large IR systems (though not in operational settings). But their conclusions must be taken with some caution. See Blair (2002, forthcoming).]

2. Data retrieval versus document retrieval

Why cannot we build commercial document retrieval systems based on the better-understood data retrieval model? Clearly, Data Base Management Systems (DBMSs) represent a relatively well-understood framework that should give us a firm foundation on which to build new commercial document retrieval systems. In addition, DBMSs have a wide selection of support software and utilities designed to facilitate database management (e.g., telecommunications interfaces, data loading programs, concurrency control systems, data dictionaries, report writing facilities, security and integrity control systems, etc.). But it has been argued that the data retrieval model is fundamentally different from the document retrieval model in a number of important ways (Blair, 1984a,b, 1999). For the purposes of this discussion, we will take document retrieval to be the retrieval of the *intellectual content* of documents. Retrieving documents which can be uniquely and unambiguously identified (such as an employment document which is uniquely identified by a social security number) is more like data retrieval than document retrieval, and is not considered here. Blair (1984a,b) identified four major distinctions between data retrieval and document retrieval (see the original article for more a more detailed discussion):

Data retrieval	Document retrieval
1 Direct ("I want to know X")	1 Indirect ("I want to know about X")
2 Necessary relation between a formal query and the representation of a satisfactory answer	2 Probabilistic relation between a formal query and the representation of a satisfactory answer
3 Criterion of success = correctness	3 Criterion of success = utility
4 Speed dependent on the time of physical access	4 Speed dependent on the number of logical decisions the searcher must make

To this list we can now add another major difference between document and data retrieval:

5. Data retrieval models scale up relatively easily while document retrieval models that provide access to the intellectual content of documents do not. The logical data model developed for a given database is not critically dependent on the size of the database. That is, once a logical model, such as a normalized relational schema, has been developed for a specific set of data, it will, with few exceptions, work just about as well on a small database as a large one. For example, if you submit a request for "The salary for the person with social security number 533 50 2857" it should not affect the search much if the desired record is on a database of 1000 records or 100,000. But the trial-and-error nature of document retrieval dictates that the uncertainty that exists in the retrieval process will increase as the document collection grows. For example, if one submits a query requesting documents on "distributed data processing" to a system of a 1000 documents, one might retrieve, say 100 documents, a few of which are relevant, but the majority of which, say

90%, are not. Now, looking through 100 documents to find the one(s) that you want is inconvenient, but tolerable. But let us suppose that the same query is submitted to a document system providing access to 100,000 documents. If the characteristics of the document collection remain the same, one might still expect to retrieve the same proportion of the database, 10%. But that 10% is now 10,000 documents rather than 100, and, more importantly, the retrieved set of 10,000 documents is now no longer a conveniently browsable size. To deal with the overload of non-relevant documents on large systems, the inquirer must submit a substantially different – often semantically different – set of queries than he would on the smaller system. This can introduce specific biases in the searching process that are well known in the decision making under uncertainty literature, and can make the entire searching process less certain than it would be on a substantially smaller system (Blair, 1980). This explains why a well-designed document retrieval system may become less effective merely because the collection of documents to which it provides access grows. Data retrieval systems do not typically suffer from such a severe dependence on database size. This is not to say that there are no problems endemic to very large databases. There are. But the principal scaling problem with document retrieval affects the search process more severely. Such an increase in size does not affect the search process in data retrieval to any comparable extent, though for some exceptions see (Blair, 1999). [1]

3. Document retrieval and the problem of scale

This scaling problem is central to the problem of document retrieval, so it is important that we examine some of the factors that influence it more closely. Document retrieval is critically dependent on how the documents are represented on a particular system, and this system of representation comprises a kind of "language" in which document content or context can be described and searchers' requests can be expressed. Consequently, the properties of this "document language" can influence the effectiveness of document descriptions and searchers' requests. As it turns out, natural languages (such as English or French) have some well-known statistical regularities that document representation "languages" can also share (van Rijsbergen, 1979). Specifically, natural languages conform to what has become known as the "Zipf distribution" [Zipf: Hill (1970, 1974); see Brookes (1968, 1984), for discussions of the Zipfian distribution in IR]. If you tabulate the occurrences of all words in a sufficiently large sample of natural language text and rank them according to their frequencies (from highest to lowest), then the product of the rank of any given word and its frequency will approximate a constant. If you then plot this distribution of word frequencies on logarithmic scales, where the x-axis = rank and the y-axis = frequency, the plot will be linear with a slope of -1 (see Fig. 1). Further, because the slope is -1, the x and y intercepts will have the same values, meaning that the total number of different

[1] This discussion of the distinction between data retrieval and document retrieval is not meant to imply that document system development has nothing to learn from the rich and impressive history of data retrieval system development; nor should it imply that data retrieval systems *never* have scaling problems. There are many lessons to learn from data retrieval (see Blair, 1988), but there are real differences that should not be lost sight of. It is also the case that these five differences between data retrieval and document retrieval are not the only differences, but they are the major ones of concern here. Eight other differences are discussed in (Blair, 1999).

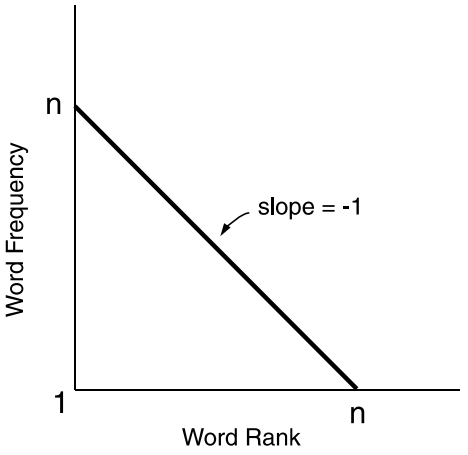

Fig. 1. Zipf distribution.

words in the vocabulary (the x intercept) will be equal to the frequency of the most frequently occurring word (the y intercept: the frequency of the word of rank number 1). In document retrieval systems, the words in the "language" are those terms or phrases which are used to represent documents, while their frequencies are merely the number of times they appear in the system representing documents. We can model the document representation language as follows:

$$N_a = \sum_{i=1}^{N_T} F_i = F_1(\ln(2N_T)),$$

where N_a are the total word occurrences in the database, N_T the total number of unique terms in the document representation language, F_i the occurrence frequency of term i, and F_1 is the occurrence frequency of the most frequently occurring word in the database – i.e., the word of frequency rank 1 (Blair, 1988).

With this equation we are in a position to say something specific about how the frequency of occurrence of individual words increases as a function of the increase in the number of documents in the collection, and this, in turn, will give us a more accurate picture of how searching degrades when the document collection increases. For example, if we have a database of 1000 documents, and each document is represented with, on average, 100 different words, then the total word occurrences, N_a will be equal to the number of documents multiplied by the average number of words used to represent a document; here, 1000×100, or, 100,000. By substitution

$$N_a = 100,000 = F_1(\ln 2N_T)).$$

Since the rank-frequency distribution is hyperbolic, then if the distribution were perfect, F_1 (the y-intercept) would be equal to N_T (the x-intercept). (In empirical studies it has been found that F_1 is somewhat less than a perfect distribution would predict, and N_T is somewhat greater. But for present purposes we will assume that the distribution is ideal.) Solving for values of F_1 and N_T we find that

$$N_T = 10,000, \qquad F_1 = 10,000.$$

Which means that there are approximately 10,000 unique words in the document representation language, and the frequency of occurrence of the most frequent word is also 10,000. Using the same methods, we find that if the document collection increases from 1000 to 100,000, F_1 and N_T will increase to 710,000. While 710,000 seems like an unusually large vocabulary, the reader must remember that it includes all the unique "words" in the running text: that is, proper nouns, neologisms, acronyms, etc., in addition to common words and expressions (see Appendix A for a brief discussion of this phenomenon).

Given the above statistics, we can see that a 100-fold increase in the size of the document collection causes a 71-fold increase in the frequency of occurrence of the document description vocabulary terms. For example, suppose we submit to the 1000 document system a simple search request consisting of the subject description "computer", and we find that 100 documents contain that word as a description (that is, 100 documents are retrieved). If that same system increased to 100,000 documents and covered the same kind of material, then the word "computer" would now retrieve 7100 documents, not 100. [One of the regularities of the Zipfian description of language is that high frequency words typically do not change their rank much (new terms naturally appear at the tail end of the distribution, then gradually move up as their frequencies increase and new terms are added to the vocabulary). Therefore, the frequencies of most of the words in the distribution will increase by the same proportion as more documents are added to the collection.] While the searcher may be able to examine the 100 documents from the smaller retrieval system, it would be hard for him to examine all 7100 documents retrieved from the larger system. The searcher is now presented with not only a *quantitatively larger* system, but with a *qualitatively different* search space. That is, he can no longer retrieve a "browsable" set of documents using the single search term "computer". In order to retrieve a smaller, "browsable", set of documents he must construct a *semantically* different search request – a search request that uses either a term other than "computer" or one that intersects another term with "computer". [See Blair (1980) for a discussion of the bias that this method of searching introduces into the search process.] This scaling difficulty poses system design problems on document retreival systems which traditional data retrieval/database management systems do not need to address directly. In short, design solutions and searching techniques which work on smaller document retrieval systems or data retrieval systems will not necessarily work on larger systems. [2]

4. Zipf and the trial-and-error nature of document retrieval

One might ask, reasonably, why the mere increase in the number of documents that are represented by the word "computer" increases the indeterminacy of retrieving documents "about computers". Is it not possible that the increase in the number of documents represented by

[2] It is not necessary for the word frequencies to follow the Zipfian distribution exactly. The argument presented here holds even if the rank:frequency relationship is somewhat different than the model presented here. It is also the case that the 7100 occurrences of the word "computers" may not occur in 7100 different documents. Some documents will have more than one occurrence of the word so the total documents retrieved will be less than 7100. Even if the growth of the retrieved documents does not go from 100 to 7100, the increase will be substantial, and enough to make the point I wish to convey.

"computer" simply increases the number of documents that are useful to the searcher? This does not necessarily happen. To understand why this is the case, we need to look again at Zipf's model of language. We had mentioned earlier that one of the essential characteristics of document retrieval is that the search process is intrinsically trial-and-error (Swanson, 1977). It turns out that Zipf has a statistical model for this, too. In short, words have multiple meanings or uses so the words which represent documents in retrieval systems are, by themselves, inherently ambiguous. Zipf showed in empirical studies that the number of different meanings or uses that a word has varies according to the square root of its frequency of occurrence in a body of text (Zipf, 1949). For example, suppose the word "computing" occurs 100 times as a document description in the smaller system. According to Zipf it is likely that "computing" is being used in about 10 different semantic ways (e.g., one document is about "distributed computing", another is about "computing charge-back rates", another is about "control of computing resources", etc.). Clearly, as the frequency of word occurrence goes up in a document retrieval language, so does the number of ways in which those words are used – the number of different meanings that they have. When our example system increased from 1000 documents to 100,000 and the occurrence of the word "computer" went from 100 to 7100, according to Zipf we can expect the number of different ways in which "computer" is used will increase from 10 to 84. So the quantitative increase in the frequency of occurrence of "computer" has increased the semantic ambiguity of the search space 8-fold. More importantly, the numeric increase in documents that are represented by "computer" will not necessarily add a comparable proportion of new relevant documents to the database, since we can assume that in most situations the inquirer would be looking for documents which related to only one of the possible meanings. The new documents represented by "computer" may be discussing subjects that are related to, but substantially different from those that the searcher would find useful. Consequently, the increase in the size of the document retrieval system may not be increasing the number of documents relevant to the searcher's request by very much, it may only be increasing the size of the retrieved set the searcher must browse through to find what he wants – the haystack in which the searcher must find the needle is simply getting larger. [3]

5. Vocabulary balance and the competing forces of language

Zipf postulated that the statistical regularities which he noticed in language were the result of the competition of two forces in language: unification and diversification (Cherry calls these forces "Personal" and "Social" in his informed application of Zipf's work to communication theory (Cherry, 1971)). Basically, language needs two kinds of words to work efficiently: *general words*

[3] Clearly, the "number of meanings" of a word is a slippery entity to measure empirically, and Zipf himself qualified his measurement by pointing out that he did not have data for the "meanings" of the 500 most frequent English words. He did have semantic data for 20,000 other words, defining the "meaning" of a word empirically as that which is listed as a meaning in the *Thorndike-Century Senior Dictionary*. Nevertheless, Zipf's observation that there is some relationship between the square root of the frequency of a word's use and the number of different meanings it has, is important. Zipf also used Lorge's *The English Semantic Count*, which he mentions without giving a citation. A more detailed presentation of these calculations occurs in Zipf (1945). From the point of view of an indexing or document retrieval vocabulary, there may be some empirical relation between the different meanings of an index term and the number of different index terms it co-occurs with.

which can be used in a variety of contexts (and have a variety of related meanings), and *specific words* with more precise meanings that have very few relevant contexts. Since you can use the same *general word* in a variety of contexts, this "unifies" language, while a *specific word*, which has a unique context, "diversifies" language. A preponderance of either general or specific words makes language inefficient, so a "vocabulary balance" occurs where language contains a spectrum of words from the very general words of high frequency to the very specific words of low frequency, and a middle range of words that balance generality and specificity in varying levels. This spectrum from generality to specificity maps right onto Zipf's distribution of word occurrences from high to low frequency. For document retrieval languages, the competing forces of unification and diversification can be more accurately interpreted as the forces of description and discrimination. [4]

The "force of description" in document representation dictates that documents, especially the intellectual content, should be described as completely as possible, so that all of the information in the document is faithfully represented. This makes the representation of any given document easier to predict for the searcher (one could say that when the document is described as completely as possible it has a larger semantic "target area" in the document search space). The "force of discrimination", though, dictates that documents should be distinguished from other documents in the search space of the system. The force of description applies to individual documents, while the force of discrimination applies to the relationship between a particular document and other documents in the collection. When a document retrieval language is biased towards *description* – as it is in simple full-text retrieval systems – the documents may be "over-described". "Over-description" means that some, or many, of the terms that are used to describe the document may mis-represent the intellectual content of that document. For example, a document that analyzes demographic data may begin with a statement "The data in this study were processed by SUN computers running SPSS under the LINUX operating system.". If this document were accessible on a simple full-text retrieval system, then it would be retrievable by queries that had any of the terms "SUN", "SPSS", "LINUX", or "operating system" in them. It is unlikely, though, that a searcher who used the search terms "SUN", "SPSS", "LINUX", or "operating system" in his/her search query would find such a document useful. When documents are retrievable by any content-bearing word that occurs in them the searcher will have an easier job of predicting the words or phrases contained in useful documents, but will have greater difficulty discriminating needed documents from unwanted ones (that is, both useful and useless documents will have some or many of the same words appearing in them or assigned to them so the words will not discriminate useful from useless documents well; Blair & Maron, 1985). Such systems have document descriptions of high *predictability*, but low *discrimination*, and the searches conducted on these systems are characterized by what is known as "output overload" – the retrieval of large sets of predominantly useless or non-relevant documents (we might call this problem a "Failure of Discrimination"). When a document retrieval language is biased towards *discrimination*, as it is in the author-title catalogue for a large research library, the documents may be described more or less uniquely. But these specific descriptions may be hard for the searcher to anticipate if he does

[4] van Rijsbergen makes a similar distinction in indexing languages and calls the competing forces representation and discrimination. He does not base this distinction on Zipf's work, though. I prefer to keep the term "representation" as a more general term, which includes both description and discrimination.

not know the exact title or author of the book he wants – that is, he wants a book with a particular intellectual content, not a book with a specific author and title. The document descriptions of such systems have high *discrimination*, but low *predictability* (we might call this problem a "Failure of Description"). The most extreme example of the *failure of description* would occur on a system that merely listed books by their unique ISBN's.

If the document descriptions on a system are biased towards *description*, searches will tend to have high *recall* but low *precision*. But if descriptions are biased towards *discrimination*, searches that retrieve relevant documents will tend to have high *precision* but low *recall* (in some cases, though, the inquirer will not be able to anticipate *any* representations of relevant documents and *recall* and *precision* will both be zero). Just as the Zipfian distribution is due to the competing forces of unification and diversification, the classic trade-off between recall and precision may be due in large part to the competing forces of *description* and *discrimination*. Recall and precision, though, do not give us any real sense of how well a query *discriminates* between useful and useless documents. "Fallout," an old, but seldom used measure of retrieval effectiveness gives us a better sense of *discrimination* than either recall or precision does.

6. Document description and the small-system effect

Ironically, biasing the document representation vocabulary towards *description* tends to make small document retrieval systems of several hundred documents work better. Since many of the older tests of document retrieval effectiveness were done on such small systems, the results tended to confirm the effectiveness of "over-description". A prominent study of document retrieval language used on a small system came precisely to this conclusion:

> Many, many alternative access words are needed for users to get what they want from large and complex [document retrieval systems]... A more familiar approach for textual objects is to extract multiple access terms automatically from the content... Gomez, Lochbaum and Landauer found full-text indexing to be roughly equivalent to extensive alias 'harvesting' from experts for the recipe database. (This strategy is referred to as "unlimited aliasing".) (Furnas, Landauer, Gomez, & Dumais, 1987)

This study was done on a database of a few hundred documents (recipes) so it is not surprising that its conclusions were dominated by the force of description. Such conclusions, though, while valid on small systems, fail to recognize the scaling problem of document retrieval, and make the mistake of assuming that document description and search strategies that are successful on small systems will work on large systems. Such is not the case. The rationale behind the unlimited aliasing of document representations is to increase the semantic "target area" of the desired documents – i.e., to increase the probability that a searcher can guess a term used to represent a desired document. What proponents of this strategy do not realize is that the more descriptions you add to the representation of a document, the more that document's descriptions begin to look, in part, like the descriptions of other documents; and there is both empirical (Swanson, 1966) as well as theoretical (Blair, 1990; Langendoen & Paul, 1984) evidence that there may be no practical limit to the number of ways a single document can be described. As a document

collection grows, the more likely it becomes that there will be documents with similar words in their text; and given the multiplicity of meanings for a particular word, documents with the same or similar descriptions may have significantly different intellectual content. The advocacy of unlimited aliasing is rather like claiming that ambiguity in natural language could be lessened simply by increasing the number of words in the vocabulary without regard for how they are used. But the lesson of Zipf's work is that languages are most effective when the vocabulary size maintains a kind of between general and specific terms, that is, to the size of the text. Insofar as a document representation language is similar to natural language, one might expect it to have similar characteristics. In short, the key to document representation is not unlimited aliasing – just adding indexing terms – the key is to add just the *right* indexing terms, ones that simultaneously *describe* the documents well and *discriminate* them from other documents in the collection. (In an empirical study, Brooks could not find evidence to support the unlimited aliasing of Furnas et al. (1987): "This experiment found no evidence to support the Strategy of Unlimited Aliasing... some index terms are simply better than others." (Brooks, 1993, p. 146).)

Thus, while *description* is the more important factor on small retrieval systems, *discrimination* becomes increasingly important as systems grow. This growth has two major consequences. First, there may never be a single "best" representation of a particular document since the requirements for discrimination will change as new documents are added or taken away from the collection. Second, since discrimination grows in importance as the document collection grows, a document representation cannot be based soley on the document itself but must take into consideration how other documents in the collection are represented.

7. What do we know about large-scale commercial document retrieval systems?

Since document retrieval systems, when used to access intellectual content, do not scale up as easily as data retrieval systems, it becomes imperative to treat large-scale systems as fundamentally different from small-scale systems. That is, large-scale document retrieval systems may require not only a different design model, but may also require a substantially different theoretical foundation than small systems (Blair, 1990). It also means that usually we cannot infer the reliability of large-scale systems based on tests done on smaller systems (Blair & Maron, 1985, 1990). The exception to this is Swanson's early retrieval experiment:

> The restriction of the collection to homogeneous subject matter was intended to permit some degree of extrapolation to the behavior of larger collections. Computer search techniques must necessarily be based on the language content of the documents searched, and it is clear that the more homogeneous the language the more difficult the problem of discriminating relevant material from irrelevant for any given request. Thus, homogeneous subject matter in a small collection would tend to present about the same level of retrieval difficulty as diverse material in a larger collection. (Swanson, 1960, p. 1100)

This is an important insight and suggests an interesting avenue of investigation, namely, to attempt to measure the "homogeneity" of text and to relate that to the comparative difficulty of searching document collections of different sizes. Unfortunately, subsequent researchers did not

do this, leaving us without any clear way to extrapolate searching from small collections to larger ones. As a consequence, we must be resolved to the fact that to understand the effectiveness of large, commercial document retrieval systems, we must test large systems in operational environments. But the testing of large-scale document retrieval systems in a commercial environment is a costly and time-consuming endeavor, so detailed tests of document retrieval effectiveness such as those of the Blair and Maron (1985) study which cost over $500,000 in today's dollars, will be rare. As a result, few precise, general conclusions can be drawn about the performance of large document systems. [5] (For a discussion of the relevance of TREC data to our understanding of the effectiveness of large document retrieval systems see Blair, 2002, forthcoming.)

It is precisely the difficulty getting accurate values of retrieval effectiveness, particularly recall, that motivates the efforts in this discussion to identify some of the factors which may affect retrieval effectiveness. What we need, at least in part, is a framework for investigating document retrieval systems – a framework which identifies the major types of possible retrieval situations. Without such a framework, the document retrieval problem is simply too large and the factors influencing good design are too varied to permit us to investigate these problems systematically. The following framework, it is argued, will help identify some of the fundamental issues of document retrieval system design that influence search effectiveness. In short, there is no single type of document retrieval situation, but a family of similar types of searches which can be classified along three dimensions: search exhaustivity, database size, and the determinacy of document representation. The description of this basic taxonomy will give us a framework in which to identify some of the major issues and processes of document retrieval. This, in turn, should help us to focus future research on the crucial areas of development in this important field. This framework is an extension of the framework first proposed in Blair (1984a,b).

8. Search exhaustivity and database size

We have already discussed how database size can significantly affect the retrieval of intellectual content. Search exhaustivity has a similar important influence. An *exhaustive search* on a document retrieval system is one in which the inquirer needs to see all, or nearly all, of the documents which are useful to him. This can be contrasted with what might be called a *sample search*, where the inquirer does not need all of the useful documents on the system. For example, a lawyer preparing to defend a client, must have exhaustive access to all documents germane to the lawsuit. To miss useful documents might put the successful conduct of the case in jeopardy. On the other hand, an individual doing market research for a manufacturing company might only need to get a

[5] Information retrieval is not unique in its need for costly empirical validations of theory; astronomy and sub-atomic physics, among other disciplines, have similar requirements. But astronomy and sub-atomic physics do not require an empirical validation of every advance they make, because each of these disciplines has a fairly well-developed theoretical model which can be used to advance the understanding of these disciplines for brief periods independently of empirical verification. Information/document retrieval does not have such a rigorous model, so, at present, it is more dependent on the empirical validation. Since such empirical validation is rarely carried out on large, operational systems, the design of successful, large-scale document retrieval systems can be somewhat hit-or-miss. Moreover, without routine, detailed tests of retrieval effectiveness we have scant hard evidence on how well these large commercial systems can be used to access document content.

sampling of documents containing information about, for example, consumers' views. Because document retrieval is inherently a trial and error process, an exhaustive search involves "more trials and more errors" than a sample search requires, *ceteris paribus*. On a large document retrieval system with the inquirer looking for a sufficiently obscure set of documents, an exhaustive search may not even be possible (many searches on the World Wide Web are like this).

There is a further distinction that can be made about *exhaustive searches* that is not reflected in the framework we will discuss shortly, but is nevertheless important and should be mentioned, if only in passing. There are really two kinds of *exhaustive searches*: those that look for documents that are *known to exist* (or *probably* exist), and those that look for documents that are *not known to exist*. The latter kind of search can be called an "existence search" – the searcher is trying to determine whether or not a desired document exists (Cooper, 1973; Cooper & Maron, 1978). Paradigm *existence searches* are patent searches and legal precedent searches. *Existence searches* differ from *exhaustive searches* in their criteria for ending a given search. *Existence searches* end when the desired document (or documents) is found – for example, a patent search ends when you find a patent for an invention that is sufficiently similar to the proposed one – or, you find evidence that such a patent exists. *Exhaustive searches* often need to continue searching even after a significant number of relevant documents are found, to insure that all the necessary documents have been found. By and large, *existence searches* tend to follow the law of the "excluded middle": the desired documents either exist or they do not, there is no notion of partial existence. *Exhaustive searches*, though, can find all, none, or some of the desired documents, and some of those documents will be more useful than others. [For example, in the study reported by Blair and Maron (1985), the *exhaustive search* was defined as retrieving a minimum of 75% of the desired documents]. In short, an *existence search* can end after the retrieval of a single, relevant document; an *exhaustive search* rarely ends so quickly and decisively. For the purposes of this discussion, though, we will not distinguish *exhaustive* from *existence* searches.

The reason why *exhaustive* and *existence searches* are so difficult is that the searcher cannot easily distinguish between the situation where the desired documents do not exist, and the situation where the desired documents *do* exist, but the searcher has missed them. Simply not finding the desired documents may not be evidence, by itself, that they do not exist. This makes such searches doubly difficult: they must not only employ good search strategies if they are likely to be successful, but they must also have justifiable "stopping criteria" for ending the search when the desired documents are not found. This is especially important on systems with large numbers of documents since these systems cannot be searched in their entirety. Of course the large size of a document collection and the ease of physical access add further complications to the search process. *Exhaustive searches* on large databases are typically the most difficult searches of all.

9. The determinacy of representation: content and context

The third and final component in this framework for document retrieval is based on how precisely the documents can be represented on any given system; this is the system of representation through which we can provide access to the intellectual content of a given set of documents by including them in specific logical or intellectual categories. In more formal terms, it is the

system of representation that, for any search, can provide an ordering of the documents in the database from those most likely to be useful to those least likely to be useful. Document retrieval systems typically match search queries to document "representations" rather than to the documents themselves (even full-text document retrieval systems usually represent documents by a subset of a given document's text which excludes specific "stop" words). These "document representations" can vary greatly in their level of determinacy – their ability to precisely identify a given document or set of documents. For example, a document's title can often identify a document fairly precisely; that is, there are very few duplicate titles even in large document collections. For those titles that are duplicate, the combination of the title and the author's name are almost always unique. Here we would say that a document's title, or its title and author, are highly determinate representations, that is, there would be little or no disagreement among reasonable individuals about what the exact title or author of a typical document is.

On the other hand, keywords, which represent the intellectual content of a document, usually identify that content far less precisely than an author or title can since many documents can be represented by the same keywords, and there can be unresolvable disagreements among individuals about which keywords represent a particular document best. Even professional indexers can honestly disagree about keyword assignments (Zunde & Margaret, 1969). In formal logic the degree of precision by which some intellectual content can be represented is known as the "determinacy of sense". The determinacy of sense has been an important issue in formal logic since the pioneering work of Gottlob Frege, Bertrand Russell and the early Wittgenstein (Baker & Hacker, 1985). To distinguish the *logical* view of determinacy from the *information retrieval* view, I will call the latter the *determinacy of representation.*

To see the relationship between document representations of high and low determinacy it is useful to consider some relevant categories of document representation. Documents are often represented by descriptions of their *content* and their *context*. The *content* (more specifically, *intellectual content*) of a document concerns the subject of that document – loosely, what it is about. The *context* of the document describes the internal or external "framework" of the document. Internal contextual information is comprised of such things as the name(s) of the author(s) of the document, its title, the type of document (memo, directive, correspondence, minutes of a meeting, etc.), date, author's affiliation, etc. External contextual information is that which cannot be gotten from the text of the document itself, such as: its place of origin, its present physical position, where it was published (if appropriate), routing record (who has seen it), what activities it has been used in, who is responsible for it, its level of confidentiality, etc.

It is important to note that the distinction between *content* and *context* descriptions is very rough, and often not at all simple and straightforward. Sometimes a description can refer to *content* or *context*. For example, when a book in the library is described as being about "American History" such a description can be taken as a reference to the book's intellectual *content*. But such a description could also mean that the book is physically located in the "American History" section of the library. In the latter case, it is a *context* description. [6] It is also the case that *content descriptions* can vary in their determinacy: for example, certain normative

[6] I am indebted to Scott Serich of the George Washington University for this example.

scientific nomenclatures, such as the names of biological flora or fauna, can identify document content more precisely than, say, the descriptions of the subject content of a series of political essays. Nevertheless, there is enough difference in the determinacy of *content* and *context descriptions*, in general, to illustrate the range of determinacy in representing intellectual content. But for the purpose of making some basic distinctions in document retrieval even a rough distinction between *context* and *content* descriptions will be useful. Those who are interested in the complexity of intellectual content would do well to consult Wilson's Two kinds of power: An essay on bibliographical control (Wilson, 1968), as a first source.

Why do we need to distinguish between *content* and *context* descriptions – between our examples of *highly determinate* and *less determinate* representations? By definition, all document retrieval searches are "content searches" – their goal is to find documents which contain certain specific information. But the process of describing the *content* of a document (usually called "subject indexing") is, at best, an imprecise process (Blair, 1986), and empirical evidence for this imprecision has been around a long time (Zunde & Margaret, 1969). Hence, *exhaustive document searches* based on *content* descriptions are often difficult, while *sample* searches based on *content* descriptions have a better chance of success, all other things being equal. As a result of this indeterminacy of *content* representation (or when *content* descriptions are not available for searching), many document searches are based on *context* descriptions. "Context searches" are inferential in nature. For example, an inquirer might infer that the information he wants is likely to be in a letter written by "Steven Dedalus" in June or July of 1990, so he retrieves the set of letters with this author and these dates and looks through them. This a context-based search. The advantage of *context* descriptions is that they do not suffer from the same high level of indeterminacy that *content* descriptions do – for example, one can usually identify the author, date, or type of a document with a higher certainty than one can identify its *intellectual content* or subject. The trade-off for *context* descriptions is that there is usually no direct connection between a *context* description and the *intellectual content* of the described document – all such connections are inferential in nature and usually must be made by the inquirer. An individual may be able to make such inferences for a document collection that he is intimately familiar with (such as his own message file), but not for one with which he is not.

One might infer that because *context* descriptions identify a document more precisely than *content* descriptions we would want to use "*only* context" descriptions to represent documents. This would be a valid inference if determinacy of representation were always a virtue. Unfortunately, it is not. An ISBN certainly identifies a published book uniquely, but it would not usually be a good starting point for a book search for the obvious reason that most people do not know the ISBNs of books that they want. ISBNs are useful when someone wants to order a specific book, but the individual usually starts with the title and author of a specific book and uses that to find out its ISBN.

The perceptive reader will see that the determinacy of representation is the most important factor in the data:document distinction, which we discussed earlier (*supra*). All the differences between data retrieval and document retrieval enumerated above are caused, in whole or in part, by contrasting levels of determinacy of representation (Blair, 1999). Data retrieval systems usually have highly determinate representations, while document retrieval systems have less determinate representations for intellectual content.

10. A framework for document/text retrieval

Fig. 2 gives the basic framework for document retrieval broken down by database size, search type (exhaustivity/sample), and the level of representational determinacy (content/context). This framework breaks down document retrieval into eight classes. An example of the type of retrieval for each class would be:

1. (*Large DB*: *exhaustive*: *content*) Corporate Litigation Support. (''Did the defendants write anything objecting to the contract changes?'')
   Research and Development. (''What work has been done in the industry on Widget development?'')
   Patent Searching. (''Is there already a patent existing for the type of product we are proposing?'')
2. (*Large DB*: *exhaustive*: *context*) Competitive Strategy. (''How is Ajax, Inc. perceived publicly? Give me all NYTimes/WSJournal articles since January 1997 that mention Ajax, Inc.'')
3. (*Large DB*: *sample*: *content*) Marketing Survey. (''I'd like to know how consumers perceive Ford's new line of cars. Give me some articles in a variety of publications in which these consumers are interviewed, or their views are tabulated.'')
4. (*Large DB*: *sample*: *context*) Business Intelligence. (''What have the Wall Street Journal editorials been primarily concerned with, lately? Get me copies of some recent ones.'')
5. (*Small DB*: *exhaustive*: *content*) Personal Word Processing Search. (''Give me copies of all the documents which Madeline has written in which she discusses new product development.'')
6. (*Small DB*: *exhaustive*: *context*) Personal Word Processing Search. (''Give me all the letters which Bill wrote to Blazes Boylan, Assoc.'')
7. (*Small DB*: *sample*: *content*) Electronic Message Search. (''What kinds of issues was Deasy dealing with before he left the company? Get me copies of some of his recent messages.'')

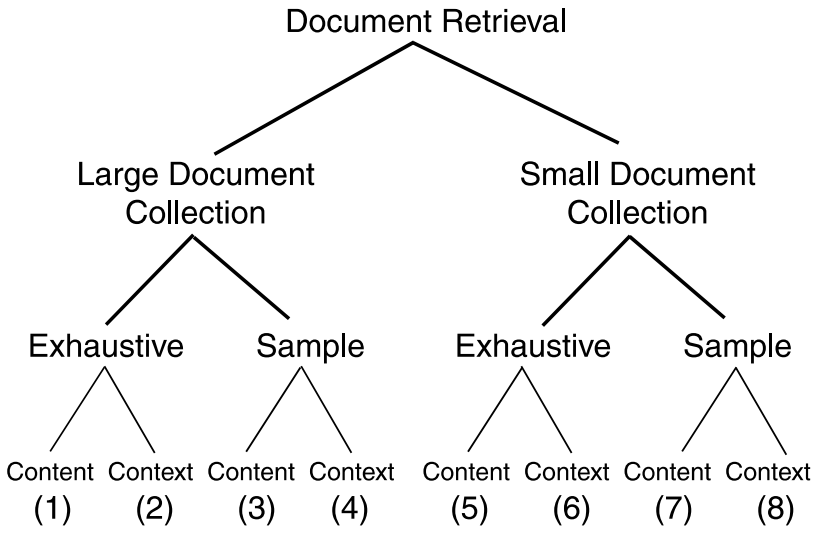

Fig. 2. The document retrieval framework.

8. (*Small DB*: *sample*: *context*) Correspondence Search. (“Whom did we write to at Blazes Boylan, Assoc.? Get his/her name and address.” (That is, find any recent letter to BB, A.))

11. Degrees of difficulty

What can we say comparatively about these eight classes of document searching? In general:

A Exhaustive content searches are more difficult than sample searches
B Searches for precise intellectual content on large document collections are usually more difficult than searches for the same material on small document collections, *ceteris paribus*
C Content searches based on descriptions of low determinacy are usually less precise than those based on descriptions of high determinacy (e.g., context), *ceteris paribus*

The comparative influence of these three factors is that database size probably influences the success of a search the most, while exhaustivity is the next most influential factor, followed by the determinacy of representation, though this may vary from system to system. By inference, then, we can say that the most difficult searches would probably be searches 1 and 2 (above), while the easiest would be 7 and 8. These are rough classifications, of course, and certainly there are other factors that can affect the success of searching in individual cases – a large system with a well-designed logical structure, such as Medline, may be easier to search for intellectual content than a small one that is sloppily or idiosyncratically designed; and a clever, experienced searcher may be able to overcome some of these problems and conduct successful searches on even very large systems. Still, they will probably have a harder time with searches of type 1 than those of type 8. But, all things being equal, this framework gives us a sense of at least some of the major factors other than recall that can affect the ease of access to intellectual content.

12. Conclusion

The thesis of the first part of this two-part article has been that document retrieval is a complex process which is strongly influenced by at least three major factors: the size of the document collection; the type of search (exhaustive, existence or sample); and, the determinacy of document representation. Collectively, these factors can be used to provide a useful framework or taxonomy of the major kinds of document searches. Such a framework helps to highlight the fundamental issues facing the design of commercial document retrieval/management systems, and may be particularly useful when thorough tests of retrieval effectiveness – such as *recall* measures – cannot be performed.The consequences which these factors can exert on the system design process are discussed in the second part of this series.

Acknowledgements

The author wishes to thank M.E. (Bill) Maron of the University of California, Berkeley, Don Swanson of the University of Chicago, Scott Serich of George Washington University, Bruce Hill

of the University of Michigan and Steven Kimbrough of the University of Pennsylvania, for their comments on earlier versions of this paper, and their discussions of the issues raised in this article.

Appendix A. How many words are in natural language?

There is no theoretical upper limit to the number of distinct "words" that can occur, and no tests of sufficiently large bodies of text have been conducted to give evidence for a practical limit. Still, it may be hard to believe that the number of distinct words will continue to grow as long as the body of text grows. Some evidence that the upper limit of distinct words, if it exists, is at least above 600,000 is provided by the ORACLE "ConText" system. ConText is an automatic document representation/indexing system that has a built in lexicon of over 600,000 words and phrases. In the first version of ConText (ca. 1996) no provisions were made to expand the size of the vocabulary beyond 600,000, but it quickly became clear that as extensive as a 600,000 item vocabulary is, it is still not large enough to capture all the different words that might be useful in retrieval. To this end, subsequent versions of ConText have the capability for the users to add additional vocabulary to the system as they see fit, thus extending the vocabulary of document representation beyond 600,000. [7]

The rate in which new words appear in everyday speech is often underestimated, though we all have some familiarity with the introduction of slang into language. Nevertheless, the addition of new words to everyday speech when considering all the topics a large document retrieval system might cover, may be substantial. Certainly, new proper nouns, and acronyms made out of proper nouns, will be added to a vocabulary regularly. For example, while the words in the company name Advanced Micro Devices, are not new, the name of the company and its acronym, AMD, represent an addition to our vocabulary. Even the number of entirely new words, or slightly changed versions of old words, may grow more rapidly than anticipated. For example, consider the neologisms which corporate America has coined to talk about "downsizing" companies: instead of being "fired", layed-off workers have been variously referred to as having been:

decruited
de-hired
deselected
destaffed
disemployed
downsized
nonretained
nonrenewed
surplussed
transitioned

In the same manner, "layoffs" have been variously referred to as:

degrowing
executive culling

[7] ORACLE: http://www.oracle.co/products/context/html/context.html

riffing ("rif" = "reduction in force")
refocusing the skill mix
right-sizing
payroll adjustment [New York Times]

The creativity and complexity of language indicates that if there is an upper limit to the size of the vocabulary used in a large document collection covering a range of topics, that upper limit may be very high, certainly higher than we have measured so far.

References

Baker, G. P., & Hacker, P. M. S. (1985). Vagueness and determinacy of sense. In *Wittgenstein: Meaning and understanding.* Chicago: University of Chicago Press.

Blair, D. C. (1980). Searching biases in large interactive document retrieval systems. *Journal of the American Society for Information Science*, *31*(4), 271–277.

Blair, D. C. (1984a). The data-document distinction in information retrieval. *Communications of the ACM*, *27*(4), 369–374.

Blair, D. C. (1984b). The management of information: basic distinctions. *Sloan Management Review*, *26*(1), 13–24.

Blair, D. C. (1986). Indeterminacy in the subject access to documents. *Information Processing and Management*, *22*(2), 229–241.

Blair, D. C. (1988). An extended relational document retrieval model. *Information Processing and Management*, *24*(3), 349–371.

Blair, D. C. (1990). *Language and representation in information retrieval*. Amsterdam: Elsevier.

Blair, D. C. (1999). *The data-document distinction revisited.* Working Paper, University of Michigan, Ann Arbor.

Blair, D. C. (2001b). The challenge of commercial document retrieval, Part II: A strategy for document searching based on identifiable document partitions. *Information Processing and Management*, *38*(2), 293–304.

Blair, D. C. (2002). Some thoughts on the reported results of TREC. *Information Processing and Management* (forthcoming).

Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document retrieval system. *Communications of the ACM*, *28*(3), 289–299.

Blair, D. C., & Maron, M. E. (1990). Full-text information retrieval: further analysis and clarification. *Information processing and management*, *26*(3), 437–447.

Brookes, B. C. (1968). The Derivation and Application of the Bradford-Zipf Distribution. *Journal of Documentation*, *24*(4), 247–265.

Brookes, B. C. (1984). Ranking techniques and the empirical log law. *Information Processing and Management*, *20*(1), 37–46.

Brooks, T. A. (1993). All the right descriptors: a test of the strategy of unlimited aliasing. *Journal of the American Society for Information Science*, *44*(3), 137–147.

Cherry, C. (1971). *On human communication: A review, a survey, and a criticism* (2nd ed.). Cambridge, MA: M.I.T. Press.

Cooper, W. S. (1973). On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, *24*, 87–100.

Cooper, W. S., & Maron, M. E. (1978). Foundations of probabilistic and utility-theoretic indexing. *Journal of the ACM*, *25*, 67–80.

Drucker, P. F. (1988). The coming of the new organization. *The Harvard Business Review*, (January–February), 45–53.

Fleischer, R. (1990). Total document control: a text-retrieval perspective. In P. Gillman (Ed.), *Proceedings of the Institute of Information Scientists 1990 text retrieval conference on Text retrieval*: *Information first* (October, 1990, London). London: Taylor Graham.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human–system communication. *Communications of the ACM*, *30*(11), 964–971.

Hill, B. (1970). Zipf's law and prior distributions for composition of population. *Journal of the American Statistical Association*, *65*, 1220–1232.
Hill, B. (1974). The rank:frequency form of Zipf's law. *Journal of the American Statistical Association*, *69*, 212–219.
Langendoen, D. T., & Paul, P. (1984). *The vastness of natural languages*. Oxford, UK: Basil Blackwell.
Salton, G., & McGill, M. (1983). *Introduction to modern retrieval*. New York: McGraw-Hill.
Swanson, D. R. (1960). Searching natural language text by computer. *Science*, *132*, 1099–1104.
Swanson, D. R. (1966). Studies of indexing depth and retrieval effectiveness. Unpublished report, National Science Foundation Grant GN 380, February 1966 [Discussed in Blair (1990), pp. 170–171].
Swanson, D. R. (1977). Information retrieval as a trial-and-error process. *Library Quarterly*, *47*(2), 128–148.
van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). London: Butterworths.
Wilson, P. (1968). *Two kinds of power: An essay on bibliographic control*. Berkeley, CA: University of California Press.
Zipf, G. K. (1945). The meaning–frequency relationship of words. *Journal of General Psychology*, *33*, 251–256.
Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.
Zunde, P., & Margaret, E. D. (1969). Indexing consistency and quality. *American Documentation*, (July), 259–267.