Database indexing: yesterday and today

Harry Diakoff

This article surveys the history of database indexing and considers its current state and future prospects. It reviews the pros and cons of computerized searching, the rise and fall (and possible rise again) of private databases, and methods of framing and normalizing queries.

In the following attempt at an overview of the importance of database (DB) indexing today I have perhaps slighted some traditional distinctions dear to the indexing profession in the search for broader parallels and relationships. Whether the end has justified the means I must leave to the reader to judge.

A few definitions

Broadly speaking, if the primary objective of the back-of-the-book indexer is to help the reader locate specific topics within a single, usually lengthy, document, the primary objective of the database indexer is to help the reader find entire documents, typically journal articles, on specific topics within some large document collection.

While a back-of-the-book index is usually part of the document that it indexes, the index terms in the literature database are usually part of a separate database record, which functions as a sort of document surrogate, much like a library catalog card, for the purposes of sorting and searching. For a more comprehensive comparison of the techniques and cultures of these two indexing fraternities, see the recent overview in *The Indexer* (Klement, 2002). For a closer view of database indexing itself see Greenhouse (2000). I am speaking here of text databases, rather than numeric databases. The latter have quite different purposes and are optimally implemented in quite different software.[1] But even with these qualifications it must be admitted that database indexing is a protean activity, not easily defined, serving a wide range of purposes with a variety of approaches – an activity whose success often seems to depend on the degree to which the indexer recognizes and adapts to the special circumstances of the indexing task. I have accordingly chosen to emphasize the changing contexts within which database indexing has developed.

A little history

The origin of printed periodical indexes

Historically, the database index is of course a direct descendant of the periodical index, which emerged as a companion to the library book catalog at the beginning of the 19th century[2] and by the end of the century had developed into several formidable general indexes and even a few pioneering indexes for technical literature,[3] which ramified into more specific subject areas in the course of the first half of the 20th century.[4]

The classification of journal contents under subject headings gradually became more systematic and complex, and in technical subjects the controlled vocabularies began to approach what might be called ontologies. Although one person's thesaurus is another's ontology, usage generally confines the latter term to vocabularies that are meant to provide a more or less exhaustive set of categories for a given subject domain – especially when the categories are mutually exclusive at a given level of generality (Jurafsky and Martin, 2000: 601). I shall use the expression 'thesaurus' to refer to the controlled vocabulary from which index terms are chosen to characterize a document, and 'taxonomy' to refer to any hierarchical thesaurus.

Computerization of the periodical indexes

The introduction of interactive online searching of machine-readable databases in the 1960s transformed both searching and indexing profoundly, although systematic exploitation of the novel opportunities was far from immediate. Medical librarians led the way in exploring the use of computers, and the prominence of bioinformatics specialists in the development of indexing and information retrieval practices has persisted to this day. The chief medical periodical index, Medline, maintained by the US National Library of Medicine, certainly provided an enviable resource for illustrating the utility of computerization, with unusually reliable indexing, a carefully controlled and well-documented thesaurus, and well-structured database records with a variety of key information items located in separate fields.

The growth of online databases was extraordinarily rapid. By the end of the 1970s, Carlos Cuadra's directory showed some 400 online databases from 221 database producers hosted by 59 online services. By 1989 there were 4245 databases, 1870 producers and 622 online services (Cuadra, 1989). Online hosting was dominated by a few vendors who provided a common interface to large numbers of databases and partially standardized their formats, facilitating cross-database searches and, much less reliably, duplicate elimination.[5] Online professionals – that is, searchers, database producers, and information managers – began to form their own associations, subscribe to their own journals, and go to their own annual meetings.[6]

Advantages of computerization

Among the many obvious advantages of interactive online searching over the use of printed indexes was the ability to create much more comprehensive searches much more conveniently, simply by connecting several index terms with the special term 'or'. Another was the ability to gain much greater precision by requiring the search results to contain multiple index terms, by connecting them with 'and' in the query. A remarkable number of searches could be handled successfully with no more than the classical (a or b) and (c or d) model. Beyond the so-called 'boolean operators,' machine searching soon provided many more conveniences, such as the ability to search for all terms beginning with a given string of characters (the 'right-sided truncation' so useful in chemistry and medicine) and to allow some fuzziness in spelling with 'wild card' symbols, especially useful in name searching.

Elementary numeric capabilities also made their appearance, enabling searches for publications that appeared before or after a certain date, or within two dates. For many users, one of the most attractive conveniences computerization provided was the automatic update of any previous search whenever the database was itself updated.

Early 'full text' searching failed to eliminate the need for indexing

Initially, most online database systems provided only the ability to search for combinations of index terms and basic bibliographic information such as author, title, and publication date. The full texts of abstracts, although they could be displayed after the search had retrieved the relevant records, could not initially be searched themselves, or could be searched only after the initial search had reduced the document set to a very small number of records that could be searched sequentially. The introduction of unlimited 'full text' searching across all the text in a database, including abstracts, was greeted with great enthusiasm, and it frequently permitted the retrieval of highly specific material, including topics referred to by neologisms that had not yet found their way into the indexing thesauri. However, contrary to some predictions, the introduction of full-text searching, initially of abstracts but soon of complete documents, in no way reduced the value of well-indexed databases such as Medline. If anything it helped illuminate the special utility that indexing can provide, simultaneously enhancing both the comprehensiveness and precision of searches. Indeed, recognition of the value of indexing to the new computerized databases was so widespread that companies such as Ovid Technologies, whose search software was able to suggest the index terms closest to the users' query terms ('mapping'), gained significant competitive advantage from this feature.

Constraints on the utility of indexing in computerized databases

This utility necessarily depends significantly on the reliability with which indexers collect the locations of synonymous expressions under a single index term, and judiciously apply the term only to occurrences above a certain minimal threshold of importance in a document. But the utility of indexed databases depends above all else on the quality of the controlled vocabulary from which the index terms are chosen. One of the most powerful features, present in Medline and most effective scientific thesauri, is a hierarchical structure, permitting the user to browse up and down a tree of related concepts or taxonomy, increasing generality and comprehensiveness in one direction, and specificity and precision in the other. Maintaining this structure so that it continues to reflect current knowledge and to ramify in an orderly way as new concepts are introduced and modified is an ongoing challenge. It illustrates the inescapably contingent or relative nature of database indexing, whose degree of success is always relative to a specific context.

Relativity of the utility of DB indexing

Since database indexes typically continue over a period of years, they must continually evolve to maintain the same degree of utility. One of the earliest ways of characterizing the adequacy of an indexing vocabulary was how well it partitioned the target documents into convenient groups for review. Terms were judged poor choices if the documents to which they applied were too numerous or too few. Obviously this is a variable that depends on the size of the database. A vocabulary that was quite adequate when a database was small becomes progressively less useful as the database expands.

Volume is not the only respect in which the adequacy of database indexing must constantly be re-evaluated. In the scientific and medical professions, which rely most heavily on indexed databases to keep up to date, the vocabulary is constantly changing and even the deeper relationships among concepts, including hierarchical relationships, can alter. Keeping track of such changes as they are introduced is laborious, but it is necessary if the user (or the search system) is to be able to adjust a search to the terminology available in each time period, e.g. falling back to a more general term for the period before a more specific term was introduced.

Recognition of the importance of indexing, and the ease with which its utility could be enhanced through computerization, also encouraged the development of private databases where the value of indexing could be maximized through various customizations even more closely tailored to specific contexts.

Private databases and customized indexing

Although the public databases were proving invaluable for many purposes, including illustrating the benefits of computerization in general, and the value of interactive rather than batch searching in particular, it was not difficult to recognize ways in which customized indexing could render a private database much more useful. In many cases, companies were already maintaining much more detailed indexing systems on critical aspects of their business such as safety data, including reports that had appeared in the published literature. But this was typically stored either in

manual files or in cumbersome computer systems that handled text searching as a very secondary function, and permitted retrieval only in batch mode, with all its inevitable delays and missed opportunities.

The depth and reliability of the indexing used to organize these files varied widely, but in many cases, the indexing used in these batch systems was much more precise than that found in the public online databases, which were designed to balance cost with the broadest possible utility to the greatest number of different kinds of users. Once again many of the best examples of highly developed indexing systems come from the pharmaceutical industry, where ambiguities tolerated in the public databases were not acceptable. In Medline database records, for example, the medications used in a study are listed separately from any side-effects observed, so that in any paper with multiple drugs and side-effects, it is necessary to turn to the original full-text document to determine which drug is associated with which side effect.

Similarly, Medline made little attempt to keep track of the actual results of the study, leaving ample opportunity for customized indexing to add value, with enhancements such as:

- normalized numeric values such as those for dosages and duration that often appeared in different units in different studies
- which drugs performed better or worse than other drugs with respect to which endpoints
- linking specific subject characteristics, such as age or concomitant morbidity or medications, with specific outcomes
- recording the absolute number of subjects in the study and those experiencing given outcomes.

The most ambitious indexing systems even kept track of outcome data in quantitative terms, so that a variety of reports, including calculations such as the incidence of side-effects across studies, could be generated directly from the database records.

Indexing at this latter extreme of precision was rarely attempted without the impetus of some special need, such as liability litigation or regulatory concerns, when an ad hoc, and usually relatively narrowly focussed, database might be needed. But, in general, the need for highly detailed indexing was widely appreciated.

Private databases: from the 1960s to the 21st century

The golden age: databases by and for information professionals Several industries, including petroleum and pharmaceuticals, made collective efforts at the development of standard indexing thesauri (e.g. RINGDOC) which often also became the basis for private, in-house databases under the pressure of competitive intelligence concerns. These databases were typically maintained by an in-house staff who collected much of the input themselves, applied their own carefully developed indexing protocols, and distributed current awareness bulletins, later in the form of emails, throughout the company. Individual librarians within each company became expert in extracting information from the customized system as well as the chief public databases.

Largely independent of the company literature databases, which were usually managed from the library, numeric data relevant to a company's activities was also computerized in many other departments. In the case of pharmaceutical companies, extensive data on drug development and drug safety, often including data extracted from publications, was captured in such departmental databases. However, since the numeric data and its manipulation had priority, relatively little text was typically entered, and the need for extensive concept taxonomies was limited and poorly supported. Despite the very different purposes and approaches of these computerization efforts, information from the same documents or publications often appeared in more than one database. This was entirely justifiable on the basis of the distinct needs of different user communities, but it sowed the seed of future conflicts, particularly where an attempt was made to expand the indexing of the general literature database into increasingly technical, and quantitative, detail.

In many ways, the 1960s and 1970s were the golden years of private database indexing. Customized company indexes abounded and their value was widely assumed. Their absence was considered as suspicious a sign as lack of a company web address would be today.

Loss of private database support and the turn to generic solutions However, beginning in the late 1980s, the costs of processing an ever-increasing volume of relevant literature, and competitive pressures on profitability, led a number of companies to scale back internal databases, and rely more on public databases and external indexers.

In many cases, and in many respects, this proved short-sighted and counter-productive. Literature still had to be managed, and public databases without custom indexing proved time-consuming to use, especially for the end-users who were now encouraged to do their own searching. One consequence was the proliferation of fragmentary databases springing up in individual departments throughout large companies – each tailored to its own user base but inaccessible (and usually unknown) outside that department. Another was the need for continual ad hoc searches as specific problems arose. Yet another was the vulnerability to the panacea of all-purpose 'knowledge management' systems within which access to journal literature was merely one (usually quite secondary) function.

Frequently the information professionals responsible for maintaining in-house databases failed to obtain a secure 'buy in' from the users they believed they were serving. In other cases, user groups that had unmet needs failed to realize that the in-house system could be modified to provide what they wanted, and quietly went ahead with the creation of departmental databases or turned to outside vendors. The lack of a constituency who could sufficiently appreciate the advantages of customized indexing over what was available in the public databases rendered the in-house systems vulnerable to downsizing or elimination as budgets became more competitive.

Limitations of generic solutions The notion that public literature databases such as Medline, Biosis, CHEM, or Embase

could fully replace customized in-house databases soon began to run foul of reality. The public databases typically captured only a small fraction of the relevant literature, and even the journals that were ostensibly included in them were often not indexed cover to cover, so crucial letters to the editor or editorials were often omitted. Although private commercial databases were much more narrowly focussed, usually on literature relating to their own or competitive products, they typically contained several times as much relevant material as the public databases on topics of interest. While clinicians may feel they only need to know what makes its way into the major journals, and researchers often feel they know everything happening in their specialty that is available in the public domain, these are not the only consumers of indexed databases. In particular, safety issues associated with the use of a company's products often appear initially in publications that are not included in the public databases.

The public databases also have their own priorities for the speed with which given journals are indexed, which may or may not correspond to those of the company, and in any case can result in lag times of from several months to over a year for any but the most popular journals. But by far the most serious problem arising when companies attempted to rely on public databases was the loss of company-oriented, customized indexing that could continually save some search time and money by reducing the size of the set of documents that the user would have to review, and occasionally save a great deal of time and money when a liability or regulatory issue required greater precision in a critically short time frame. In many cases customized indexing would have provided the information for an immediate response.

Realism: cost-effective, modular databases defined by users' needs

Today we are seeing a reversal of the pendulum as many companies attempt to recentralize their use of database resources, but on a more explicit cost/benefit basis, using standardized software and a flexible mix of external vendors and in-house capabilities. The modularization of information retrieval resources, and the commoditization of much of the functionality, have driven some costs down. Management has also become much more alert to the hidden costs of lost productivity that may not show up within individual departments, but tax the company as a whole. For example, the library budget for searching might be down, but end-users throughout the company, each spending five minutes more to find what they need, cost the company far more than is saved.

The specific level of precision in indexing that is really cost-effective for specific users is being examined, often for the first time. If it is known how much time is spent by users in a given salary range searching for information (this is usually readily derivable from the logs of public database providers) a few simple pilot tests on dummy databases with a few typical users can provide a rough indication of the value to the company of a given level of indexing specificity – that is, the cost of the time users spend finding what they need minus the cost of indexing at that level. The real-world situation is usually significantly more complicated because of the different needs of different users whose time is variously valued, and the difficulty of predicting how much more use a system might find if the indexing corresponded more closely to the users' needs. But the basic point is not difficult to make.

In some respects, the development of electronic new drug applications (or computer assisted new drug applications – CANDAs) has helped define a model for private databases that is at once more powerful and practical. CANDAs require the management of a large volume of text, large quantities of numeric data, and the ability to transfer both kinds of information back and forth among a variety of software for analysis. The freeing of the data from any specific software with the development of data transfer standards has encouraged the emergence of today's highly modular and customizable database environment – where tools for text management, numerical analysis, and report generation are often quite loosely coupled in systems that provide maximum flexibility rather than the maximum data transfer rates with the maximum numbers of concurrent users. (For CANDAs, see the FDA's website on Electronic Regulatory Submissions and Review at `http://www.fda.gov/cder/regulatory/ersr/default.htm`.)

Thus the user may be happily unaware that a request has been processed by several different kinds of software, each optimized for a different functionality, e.g. text searching, numeric analysis, or report generation.

Even if company-wide literature indexing that is sufficiently comprehensive and detailed to satisfy the needs of every constituency remains too expensive, there is in principle no reason that a single system cannot begin processing the literature, provide basic indexing, and reroute the output to whichever departments feel they need a customized view of the result. This could involve further indexing that should not be shared throughout the company, a customized interface, or even specialized search and report generating software. Such systems are now in development, and it is expected that the balance between centralization of literature resources and distributed customization of literature characterization will provide less redundancy with a greater potential for use by different constituencies and reuse over time. It should also create a vehicle for distributing the costs of the database more equitably.

For the foreseeable future, customized indexing of private databases, including an ever-diminishing measure of manual indexing, will continue to survive where it can clearly be demonstrated that the advantage in comprehensiveness, currency, reliability, or speed of retrieval is cost-effective.

Indexing and the internet

The importance of indexing today goes far beyond the needs of individual companies or even industries. Some supposed that the introduction of vast online repositories of full-text articles on the World Wide Web, together with a sophisticated new generation of web search engines, would finally eliminate the need for database indexing. Of course, nothing has been further from the truth. As any novice Google searcher will testify, searching full text directly can be a frustrating exercise in unexpected ambiguities, and involve browsing through a great deal of irrelevant material.

Nonetheless, if success is measured by rapid retrieval of something approximately relevant to a query, perhaps requiring a bit of browsing through a result set, a Google search is usually successful. But if success is measured by either of the traditional metrics of information retrieval, recall (the percentage of existing relevant documents found) or precision (the percentage of found documents that are relevant), its performance is more modest.

The problem of missing relevant documents is however an invisible one to the user, who sees only what the system finds. Furthermore, if the search is simply for a piece of information, say the birth date of Mozart, one document will do as well as another and the number of relevant documents found becomes irrelevant. Most searches in the public databases are of course not for the birth date of Mozart, but rather consist of one or two words relating to pornography or current celebrities, and recall is widely perceived to be quite adequate on these topics. The problem of lack of precision, often arising from an unforeseen meaning of the query terms, or a skewing of the relevance ranking by commercial websites, with or without the search engine's cooperation, is more readily perceived. Search-engine sites go to considerable lengths to defeat the more intrusive commercial attempts to distort the sequence in which results are presented, but many of the problems with precision are very difficult to deal with. An example of a situation where it is very difficult for a single search-engine algorithm to satisfy all users equally is the common problem of ranking documents by more than one criterion, e.g. topicality, currency, authority. Those searching for medical information, for example, often want the most authoritative information, whether or not it is precisely the most current, and even if it is slightly off-topic.

One of the founders of Google recently proposed a third metric for evaluating the success of a search engine at a meeting of the Association for Computing Machinery's Special Interest Group on Information Retrieval. He had suggested that the traditional measures of search engine success such as recall and precision were not appropriate for the web, and was challenged from the floor by Chris Buckley, a pupil of the IR guru Gerald Salton, to provide a more appropriate metric. He proposed simply 'time required for a user to find what he wants'. This is of course an excellent metric for the information-seeking query mentioned above, but even here it must be acknowledged that if the database were pre-indexed, particularly if the indexing could anticipate common types of queries, the time for a successful search could be greatly reduced.

Searching a text database typically involves the attempt to match an incompletely characterized document with an imperfectly expressed, and often imperfectly conceived, query. Accordingly, anything that can be done to normalize either the target text (which might be a database record, full text or a web page) or the query can improve recall, precision, and, yes, the time required for finding any specific piece of information.

I will consider attempts to enhance queries shortly (discovering in the process that the most successful efforts require the same tools that indexers use), but I will start with attempts to characterize or otherwise preprocess the target text. No matter how clever a search engine can be with query processing at run time, it is always possible to do more beforehand, when the related constraints of elapsed time and machine resources are not so limiting.

Characterizing the target text – documents and web pages

The explosive growth in the volume of machine-readable text on the web has only made the need for indexing more acute. Any user of Google quickly discovers that the lack of standardized characterization of online contents is one of the most frustrating sources of irrelevant hits.

The obvious need for better characterized web pages has given rise to a great variety of (largely) complementary efforts.

Standards for indexing light metadata

In the attempt to standardize at least some minimal bibliographic indexing of web pages to reduce the ambiguity that escalates along with the size of the web, a variety of so-called 'metadata' initiatives are underway, such as:

- `http://dublincore.org`
- `http://www.ifla.org/II/metadata.htm`

Standards for combining metadata with locators such as URLs

- `http://doi.org`

Manually created directories

Manually created taxonomies, or hierarchical thesauri that are applied to web pages instead of documents, are used either to browse (as is done in any indexed database) or to help rank search results from major web search engines, providing what is often an initial layer of very high quality hits. They include:

- Open Directory Project `http://dmoz.org/about.html` (used in Google, Lycos, Hotbot etc.)
- Yahoo `http://www.yahoo.com`

For the current relationships between search engines and directories see `http://searchenginewatch.com/links/article.php/2156221`. Few web efforts have better illustrated the enthusiasm that a manually constructed browsable index (i.e. 'directory') can still generate in the age of the Internet.

Development of specific subject thesauri/taxonomies/ontologies/knowledge representations

Some of the most detailed and comprehensive have been developed for biomedicine. The extraordinary explosion of controlled vocabularies to help organize practically every conceivable domain of human knowledge has necessitated the appearance of many guides to the available thesauri, and the development of meta-thesauri that try to facilitate the

translation from one controlled vocabulary to another. For example the Unified Medical Language System developed by the US National Library of Medicine now contains over 1 million distinct concepts and nearly 3 million terms from over 100 separate controlled vocabularies (http://www.nlm.nih.gov.research.umls.umlsmain.html).

Support of ontologies on the web

- semantic web http://www.w3.org/2001/sw/
- OWL http://www.w3.org/2004/OWL/
- DAML/OIL http://www.daml.org/

Examples of major ontological initiatives are:

- guides: http://daml.org/ontologies
- general linguistic: http://wordnet.princeton.edu

Theory of knowledge representation

A great deal of attention has also been given to the theory of classification and knowledge representation that underlies indexing practice. For a recent review see Sowa (1999). A history of knowledge representation could plausibly begin with the Sanskrit grammarians of the first few centuries BC (Briggs, 1985) but to keep up with more recent developments it could be useful to follow the bi-annual conferences on Theoretical Aspects of Rationality and Knowledge (http://www.tark.org) or the annual conferences on Principles of Knowledge Representation and Reasoning (http://kr.org).

Systems to support the manual development or enhancement of controlled vocabularies, ontologies, taxonomies, thesauri, and their automatic application in searching

Virtually all the major search software vendors have developed modules for the development (manual, automatic, and semiautomatic), and maintenance and application of controlled vocabularies, e.g.:

- Verity: http://www.verity.com/products/vcc.index.html
- Fulcrum/Hummingbird: http://www.hummingbird.com
- Convera: http://convera.com/Products/rw_categorization.asp

Many further examples can be found at http://www.searchtools.com/info/classifiers-tools.html.

Most of the traditional specialists in developing controlled vocabularies for customized applications have now been swallowed up by general knowledge management systems, which tend to emphasize their automatic application, e.g.:

- Lexiquest's acquisition by SPSS (http://www.spss.com/lexiquest/lexiquest_categorize.htm)
- Semio's acquisition by Entrieva (http://www.entrieva.com/entrieva/products/semiotagger.asp?Hdr=semiotagger)

Software to assist in the construction and application of controlled vocabularies is also provided by modular standalone software, e.g. DataHarmony's ThesaurusMaster to help build the vocabulary and its MachineAidedIndexer to help apply it as efficiently as possible (http://www.dataharmony.com/products.tm.htm).

Several companies specialize in customizing and applying existing medical taxonomies, e.g.:

- http://www.healthlanguage.com
- http://www.apelon.com
- http://www.medrahelp.com (a division of PSI which provides assistance to pharmaceutical companies in the preparation of new drug applications (http://www.psiint.com).

Standards

As important as the ontologies themselves are the standards for using them. The main standards for expressing the structure of documents that allow one to specify the meaning of structural elements in ways that other software programs can recognize are SGML and XML. See:

- http://www.oasis-open.org/cover/
- http://www.xml.org/xml/xmldev.shtml
- http://www.isgmlug.org

An excellent annual conference is http://www.extrememarkup.com/extreme/

Standards for expressing annotations to documents, links among documents, and links among links are:

- RDF: http://www.w3.org/RDF/
- Xlink and Xpointer: http://www.w3.org/XML/Linking/
- Hytime: http://www.sgmlsource.com/history/hthist.htm
- http://www.y12.doe.gov/sgml/wg8/docs/n1920/html/n1920.html
- Topic Maps: http://www.infoloom.com/tmstands.htm
- http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html

It should be remembered that any software system that supports searchable annotations to documents can in principle become the nucleus of an indexed database. When Tim Berners-Lee originally conceived what became the World Wide Web, he intended the web browser to be an annotation tool as well as a passive search tool, and the W3C has in fact devoted considerable effort to developing a free annotating browser, Amaya (http://www.W3.org.Amaya/).

The relationship of purely manual indexing to more automated approaches continues to evolve. Needless to say, most systems in use by database indexers today are able to import the sorts of standard bibliographic data that appear in the public databases and which citation manager software can accept and reformat, but a variety of efforts have been made to go much further.

Some representative approaches to automating indexing

'Automatic indexing' is an expression that can quite legitimately be used to refer to any one of a very broad spectrum of activities, ranging from the trivially easy to the impossibly difficult. At the easy end of the spectrum is what might be

considered as simply standardizing vocabulary. All this requires is a synonym thesaurus and software capable of recognizing nouns and noun phrases in documents. The software reads through the documents looking up all the concepts it recognizes in the thesaurus, then creates a new field for index terms into which it places either all the synonyms for the concept that it has found, or one preferred term that will represent the concept. Adding all known synonyms makes no assumptions about the query processor that will subsequently be used, while adding only a preferred term assumes that any query processor subsequently used will be capable of the same translation in reverse – from synonyms to preferred term. This is not a machine-intensive process, and in either case it is a much more efficient search than attempting to expand queries into all their synonyms at run time and search each separately while the user stares at the terminal.

A variety of more ambitious systems claim to extract data directly from unstructured full text, but while many of them are very useful for exploratory text mining, none would be sufficiently reliable for 'mission critical' applications without a great deal of attention to query types and target documents. 'Entity recognition' based on the standard formats for dates, prices, company names, and the like is quite straightforward. Much more demanding is the attempt to capture relationships among concepts (such as distinguishing between cause and effect). Such relationships can be expressed in a bewildering variety of locutions, many of which require parsing entire sentences for their reliable discovery, and some of which require the interpretation of even larger contexts. This kind of parsing is laborious and slow even as a pre-processing option. For run-time query interpretation (see NLP below) it is still usually considered too slow to be of much practical use unless the target documents are well known and appropriately characterized in advance.

Many probabilistic and correlational systems can identify topical clusters of documents within a database automatically, and can be useful for browsing unfamiliar collections, or for finding more documents that are similar to those the user has already identified as relevant. Similarly automatic systems, such as neural networks, can be trained to route similar documents to one or more destinations based on typical correlations of terms, or apply a small number of index terms to them. But systems of this sort typically do not perform well with very large numbers of categories. Medline now has nearly 23,000 index terms, most of which can take several of the 80-plus subheading qualifiers.

An important challenge of particular importance on the web is automatically assigning a document some indication of the likely reliability or authority of its information. Many systems simply provide the opportunity to restrict search results to a set of top peer-reviewed journals, but the traditional mechanism of peer review is increasingly recognized to be too slow and too arbitrary to suit today's rapidly evolving technical fields. Accordingly, more and more ingenious efforts are being made to develop relatively real-time and preferably automatic methods of indicating the current view of an author's peers about his publication, or better yet, specific assertions made in that publication. Ideally such a system would incorporate the evaluations of commentators with some indication of their relative authoritativeness, based perhaps on previous publication history in that field, and the reception that their own publications have had, or simply professional standing. It would also change over time as the article's conclusions were confirmed, refuted, or modified by subsequent research. Users are, after all, usually interested in the credibility of the data at the time they are searching, rather than at the time a reviewer expressed an opinion in the past.

One suggestion for providing real-time authority ranking emerging from an IBM research group a few years ago illustrates some of the opportunities that current information systems can provide. The system simply kept track of who recommended what, so that over time those whose recommendations most accurately predicted the future consensus on recommendations were given greater weight in subsequent rankings. It is rather as if a Zagat system for restaurant recommendations was constantly adjusting its weighting to give greater influence to those who had proven accurate predictors of value in the past, and reduce the influence of those who had failed to predict accurately.

The sophistication with which the categorization of text can be automated can only increase. With a reliability that no human indexer can match, automatic techniques will eventually rival the best that humans can provide. But this remains well in the future. At present automatic techniques can only provide either a conveniently rapid but very rough categorization of heterogeneous materials, or a precise categorization of thoroughly predictable and homogeneous materials. This will change, but not tomorrow.

Normalizing the query

Many efforts have been made to process the query, often in the hope that this can reduce the importance of pre-processing the text. Much of this work is commonly referred to as 'natural language processing,' or 'NLP', and this rubric again covers a wide range of approaches, of quite variable difficulty.

Fortunately, some of the relatively easy measures are among the most beneficial. Since the early studies on Medline in the late 1960s (Lancaster, 1969) we have known that the chief obstacle to comprehensive recall is simply inability to think of the relevant synonyms for the words that first occur when formulating a query. Synonym thesauri can easily be implemented as part of the search engine, and if hierarchical thesauri are available, it is possible to provide a more sophisticated relevance ranking that returns in sequence documents matching:

- the literal terms of the query
- exact synonyms
- hyponyms
- immediate parent hypernyms
- all the children of the immediate parent hypernyms
- grandparent hypernyms etc.

Such a ranking is sometimes said to incorporate the 'semantic distance' separating the terms of the query and the document. This is especially useful in technical databases

where the occasional user, such as a doctor, may be unaware of the level of generality of the indexing. The result is a much more forgiving information retrieval system; one that reduces the consequences of a poor match between query and documents.

Although this approach works well, using a complex hierarchical thesaurus in this way to expand queries at run-time can involve hundreds of additional searches, imposing a significant burden on the system, and delaying response time appreciably. The obvious solution is to preprocess the target documents, decomposing them into words and phrases that can then be matched against a thesaurus, and then either adding all the synonyms to the record (as well as the hyponyms and hypernyms in separate fields), or simply choosing one canonical term to represent each concept and adding that one term. In the former case, relatively little query processing is needed, while in the latter, the query terms and phrases must be mapped into the same canonical terminology. In either case the process typically reduces machine time by one or two orders of magnitude.

It will be noted that this is in fact indexing, although of an unusually automatic and reliable, if limited, sort, and its adequacy depends primarily on the quality of the thesaurus being used. While the automatic indexing program is looking for matches to its thesaurus, it can easily keep track of the number of occurrences of synonyms for a given concept that it finds in each document, generating a very useful measure of the salience of the concept in the document. This information can be quite useful to the relevance ranking algorithm, enabling it to present the documents where the concepts in the user's query are most prominent before others where the concept may appear only as an incidental mention.

Thus we see that the same simple indexing procedure can contribute at once to improved recall, improved precision, and reduced time to locate the most important documents. The same procedure can even make a crude contribution to what is sometimes called authority ranking, by keeping track of the length of the document being indexed or the number of references it contains, giving what appear to be weightier documents a higher relevance ranking. In general, however, authority is one of the characteristics most conspicuously lacking in web relevance ranking. The attempt to use popularity as a surrogate, measured either by the number of links to a site or the frequency of navigation to a site, works better for some topics than others. Again, medicine is a problematic area because of the large number of commercial sites vying for the attention of the searcher. Pre-inclusion screening is the method most often used in private databases, so the presence of the record in the database is itself a kind of index term representing authority.

Another kind of processing that 'natural language' can easily do at run time is disambiguation of polysemous terms using other terms in the query and appropriate thesauri whose terms are classified by subject. If the query contains the term 'bond' other terms in the query may indicate whether synonyms related to financial instruments, chemical affinity, or psychological affinity should be used in the search expansion.

It may be noted, however, that the examples given so far have all dealt with separate concepts, individual or multiple, but not relations among concepts. These, including such basic relations as causality, inclusion or exclusion, and quantitative relationships, create challenges for automatic indexing and remain largely the province of the manual indexer. Parsing the query is not so unmanageable, and when the target documents are highly structured or reliably indexed, is often worth doing. Reasonably effective natural language systems to query instruction manuals, help desks, and SQL databases have existed for years. When the kinds of answers available are known, particularly in a very limited subject domain, it may be worth going to the trouble of trying to anticipate the common locutions a user may employ, although creating a menu query system is often both simpler and more effective.

True natural language parsing, complete grammatical analysis of the user's query, is both surprisingly difficult and of surprisingly limited utility. In the first place, most user queries, even in technical databases, consist of only one or two words, expressing very few concepts; and in the second,

DNI – German Network of Indexers

At long last, a few indexing specialists have gathered to form a group of indexers in Germany. At present there are not enough people to found a German Society of Indexers. Instead we have formed the German Network of Indexers, or in German, Deutsches Netzwerk der Indexer (DNI).

DNI regards itself as a pool for all aspects of indexing, including book and periodical indexing. The network is open to everyone with an interest in indexing in German-speaking countries as well as to all indexers outside German-speaking countries interested in German.

Despite a long book tradition of over 500 years, it is surprising to see that indexing is often neglected in today's German publishing industry. One of DNI's goals is to address this issue.

We are proud to have Dr Robert Fugmann among our members. Undoubtedly he is one of the very few 'big guns' of the German indexing community, as well as a prolific scholarly writer on indexing and knowledge organization, his experience of all kinds of indexing aspects dating back to the early 1960s.

We are also happy to call readers' attention to our web site at `www.d-indexer.org`, launched back in July. There is also a welcome page in English. Apart from listing ourselves, we have included a German bibliography and reviews of German indexes. For those indexers who want to link to the DNI web site, we offer to list them with links to their websites in return.

DNI looks forward to having a lively exchange with our fellow societies of indexers and their members.

Jochen Fassbender
Deutsches Netzwerk der Indexer (DNI)
`info@d-indexer.org`

the nature of the real world precludes many theoretically possible ambiguities. Even if the system in incapable of specifying the direction of the causal arrow in a query such as 'Did a comet cause extinction of the dinosaurs?', which the system interprets simply as 'comets and extinction of dinosaurs,' the results will be saved by the fact that comets are more likely to cause dinosaur extinctions than dinosaur extinctions are to cause comets. There are of course awkward exceptions, and one way to test a supposed 'natural language processing' search engine is to ask for only papers where a drug causes the symptom it is normally prescribed to treat, e.g. 'Can barbiturates cause agitation?' Such 'paradoxical' reactions are not that rare with many drugs such as sedatives, but they are uncommon enough to be swamped by the papers describing therapeutic benefit if the system is merely searching on 'barbiturates and agitation,' and the user has failed to give the system a hint by including a term such as 'paradoxical.'

Most systems that encourage the user to enter 'natural language' queries currently simply break the query down into its nouns and noun phrases, ignoring the relations among these terms, expand these terms into their various synonyms (and in some cases hyponyms) and then rank the documents to be presented to the user by the number of query concepts they contain. If the database is a rich one and the query less demanding, there may be enough documents that contain all the query concepts to satisfy the searcher. If that is not the case, users find themselves looking at documents that contain fewer concepts than were in their original query, and the system must decide which of the original concepts was most expendable. Considerable ingenuity can be expended on trying to determine automatically the sequence in which concepts should be dropped, but generally those that are more common in the database being searched are considered more expendable.

More ambitious systems devote considerable attention to phrase decomposition, permitting a portion of a phrase to continue to represent a concept before it is dropped. This can be particularly important in technical subjects such as medicine, where extensive qualifiers of variable semantic value often create very long noun phrases. It can be useful for a system to recognize that in the phrase 'percutaneous endoscopic gastrostomy' 'percutaneous' is probably more expendable than 'gastrostomy,' which has some claim to be a partial representative of the complete concept. Quite elaborate relevance ranking schemes can be implemented that take into account the semantic distance between query terms and document terms, as well as the number of query concepts present in the target document and their degree of completeness. Such considerations will be superfluous in many simple queries, but they will contribute to the overall goal of making the search system more forgiving, thereby increasing the number and type of searches that can be dealt with successfully (yet another measure of search system quality).

Many other approaches to retrieving relevant documents and extracting desired information from unindexed full-text have been, and are being, evaluated. For reviews see the standard works by Salton (1988), Baeza-Yates *et al.* (1999), and Witten (Witten *et al.*, 1999; Witten and Frank, 2000).

For current research see:

- Text Retrieval Conference sponsored by the National Institute of Standards and Technology: `http://trec.nist.gov/`
- Association for Computing Machinery Special Interest Group in Information Retrieval: `http://www.acm.org/sigir/`
- Journal of the American Society for Information Science and Technology: `http://www.asis.org/Publications/JASIS/jasis.html`

Such systems (whether simple Google, NLP, or automatic data mining) all work, as long as the criteria for success are not too stringent. Can they ensure that all documents relevant to a user's query are found? Can they avoid returning anything that is not relevant? Can they ensure that searchers will spend as little time as possible finding what they want? Can they address the greatest number of query types? The answer to all these queries is clearly 'no.' All of these objectives can be addressed more adequately by pre-processing the target text in one way or another: enforcing stringent inclusion criteria that guarantee quality, applying specific index terms to the target documents that are matched in a user interface tailored to searching on those terms, and many other devices for normalizing both target and query. But all these methods are more labor-intensive and costly. The question today is not whether to index or not, but when and how to index.

Indexing in the biomedical century

The traditional selection pressures on the evolution of the practice of database indexing, increasing publication volume and changing vocabulary structure, have today been joined by the ever-increasing importance of real-time currency. The rate of change in the biomedical sciences is outstripping the capabilities of the traditional publication cycle for print media. In rapidly evolving fields such as molecular biology or oncology, the need for near-instantaneous dissemination of research results is acute. The evolution of new tools and solutions is certain to provide many surprises as this evolution continues, but some directions can be noted.

The challenge is not just to provide web publishing platforms but to provide systems and tools that facilitate the incorporation of new results into existing knowledge as quickly as possible. When doctors query a literature database today, whether in person or through an automatic trigger from the prescription they are writing or the chart they are completing, it verges on criminal negligence for the database host not to provide them automatically with any very recent information that changes the drug's safety profile. Since regulatory authorities now make decisions such as withdrawal of a drug from the market or the issuing of a new warning available on the web in standardized formats, it is a trivial task for the database host/vendor (rather than the original producer) to automatically monitor the regulatory sites, detect and parse any changes reflecting the altered status of a drug, and convert them into a record in an ancillary database. This database can then be searched automatically along with the primary database, and provide a pop-up or sidebar that warns the physician of significant

changes. The major database does not have to be reloaded and its original records need not be altered. It is quite remarkable how many opportunities of this sort now exist.

A major strategy for rendering information more readily accessible involves developing the thesaurus of terms into browsable knowledge structures including very specific information for which the literature target is merely a supporting reference, which usually does not need to be consulted directly.

Medical databases as usual have led the way, with the introduction of modular textbooks that can be updated in segments. *Scientific American Medicine* was an early example, and the *Merck Manual* is today most easily used online, but one of the best current illustrations of a series of hierarchical statements designed for browsing is Michael Kauffman's *Outlines of clinical medicine* (`http://www.outline.med.com/about.htm`), where clinical observations and recommendations are carefully structured to permit individual references and updates.In such cases, the 'index' becomes a series of assertions that the literature references support. To the extent that the assertions represent the content of the article cited, they can replace it for many purposes, although anyone interested in the methodology and details of the research will have to turn to the article itself.

Today many examples are emerging in which the interface that a database searcher actually uses is really just a portal to many quite heterogeneous databases, including conventional text databases as well as relational databases with numeric data, and highly structured knowledge management systems. This allows each kind of data to be managed with the optimal software, and such hybrid systems can be designed to provide very complex reports combining numeric with text data.

Needless to say, every effort is being made to automate the access points to this heterogeneous and constantly changing information, but manual 'curation' remains essential for many of the most useful databases. The National Center for Biotechnology Information (`http://www.ncbi.nlm.nih.gov`) provides one of the best gateways into the current international web of biological information resources, many of which illustrate the shifting balance from indexed literature to knowledge structures with supporting references. An example of the purely pragmatic balance struck by many of these resources is AceView (`http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/`). When users search on a gene they are presented with a brief summary of what is known about the gene's location, variants, protein products, etc. followed by a chronologically arranged series of specific research findings involving the gene, each with a link to the paper (or abstract of the paper) in which it was reported. Despite the chronological arrangement of each gene's literature, the functionality for the user is not in principle different from the browsable index of topics in the traditional literature database – its specificity, or 'granularity,' is simply much greater. And it represents a momentarily effective balance among competing goals such as currency, economy, and convenience. It illustrates again just how very closely adapted to a specific context the successful database must be today.

A variety of methods have been developed to help beleaguered biomedical indexers cope with their ever-changing concept taxonomy. One frequently useful approach is to permit indexers to enter concepts in the literal terminology encountered in the document they are indexing, focussing their attention on the relatively few conventions needed to capture and normalize the relations among the concepts (e.g. associating specific drugs with specific effects). The

Clinton in German

Mein Leben, the German edition of *My life* published by ECON Verlag, contains 1,472 pages, almost 500 pages more than the American edition. According to a report about Clinton's book promotion tour in Germany, given in the major daily *Hamburger Abendblatt* of 10/11 July 2004, the reason for this is that 'Clinton added chapters which specifically deal with German-related issues.'

However, a closer look at both editions failed to confirm the claim by *Hamburger Abendblatt*. There are 55 chapters in both editions. Apart from a few additional pictures, there is no clue that the book was substantially adapted for the German audience. In fact, the main reason for the 500-page difference seems to be the book's format.

Although the font size is the same in both editions, the German one has five lines less per page than the American version; also, the line width is shorter. Adding this calculation to the fact that German text tends to be longer than its English equivalent, it becomes clear that the German edition is by no means 500 pages larger. Maybe this is just a marketing claim.

Turning to the index is a mere disappointment. ECON did not bother to create a subject index. Instead, there is just an index of names, as is unfortunately often the case with biographies published in Germany. The length of this index is a meagre 12 pages – not even 1 percent of the book's length. Many name entries consist of 10, 15, 20, or more undifferentiated page references, the record being Al Gore with 90! There are no page ranges; instead, 'f.' and 'ff.' are used – still a common practice in German indexes.

Moreover, many name entries with just a few page references are missing in the German version; as for names given in both editions, there is not much difference in the number of page references for each entry. Surprisingly, even Helmut Kohl (nine each) and Gerhard Schroeder (four and three in the English and German editions respectively) did not get more page references, another indication that there is no supposedly German-related coverage.

It is regrettable that ECON failed to provide a subject index, thus denying its readers proper access to the book's content. Apart from the incomplete index of names, the absence of a subject index also means that ECON has not published a true German counterpart of Clinton's book.

Jochen Fassbender, *Germany*

computer then looks up the literal term in a thesaurus, puts the preferred term in the field that will be searched, and stores the literal term in an invisible field for future reference. It is invisible but still associated with the document in which it was found. Any literal expressions not found in the thesaurus are routed to an expert human curator of the thesaurus. The preferred term for any term encountered thus has to be looked up only once, by the curator, rather than by each indexer in each document, and if subsequently it becomes necessary to revise the relationships among terms previously considered synonymous, it can be done without having to review the original document.

Today the venerable periodical index, transformed in the past 50 years into the database index, is in the process of still further transformations. Driven by the constantly accelerating pace of change in both its technology and its content, the pressure for automation and integration of all aspects of knowledge management can only increase.

Two trends appear robust at the moment. On the one hand the index term is being expanded into an informative statement embedded in a structure (either hierarchical or non-hierarchical) of similar statements, which can be searched or browsed in order to locate documents that contain more detailed support for the statements. This trend is being limited primarily by issues of time and cost in very rapidly changing research areas, because at the moment it still requires a good deal of manual intervention to be effective. But it has obvious utility as an educational tool as well as a purely reference resource, and the need for manual curation will gradually be reduced. On the other hand, real-time annotations, provided either by automatic systems that monitor key websites for important changes, or by human commentators, are becoming ever more crucial to help searchers retain some orientation in the expanding sea of documentation and information, where their own field of competence has shrunk again each morning. While the traditional database index term which provided a convenient, standardized access point into the periodical literature is thus being gradually transformed into knowledge statements, indications of information quality or currency, more discursive annotations, and even less familiar progeny, its basic function – providing a standard vocabulary for searching and browsing text – and its basic tools – thesauri, taxonomies, and ontologies – are today more widely valued and utilized than ever. Indeed it appears that the index term may long survive the printed periodical articles it was meant to serve.

Notes

1 The numeric database is intended to keep track of rapidly changing numeric values, and to permit their use in numeric calculations. It is typically implemented in a relational architecture with fixed-length fields. The text database is intended to keep track of character strings such as words and sentences located throughout documents of indeterminate length, and its search engine has traditionally been implemented as an inverted index, which is optimized for records whose contents do not need to be altered frequently. The expression 'database indexing' is often used to refer to the process by which the computer builds the inverted index or other system of keys out of text that is being 'loaded' into the database.

2 Poole's *Index to periodical literature* began publication in 1802.

3 For example:
- Bibliographie der deutschen Zeitschriftenliteratur 1896–
- Internationale Bibliographie der Zeitschriftenliteratur aus allen Gebieten der Forschung. 1897–
- Repertorium der technischen Journal-Literatur 1823–1908, Berlin
- John Shaw Billings Index Medicus 1879–2004. The print version will finally end at the end of 2004.
- Engineering Index 1895–

4 For example:
- Chem Abstracts 1907–
- Biological Abstracts 1927–
- Education Index 1929–

5 For example:
- SDC/Orbit
- Lockheed/Dialog
- BRS
- Mead DataCentral/Lexis
- Derwent Patent database

6 For example:
- Online Review 1977–1992, Learned Information, Oxford
- Online 1977, published by Information Today after 2002 (`http://www.infotoday.com/online/default.shtml`)

In the United States the first National Online Information Meeting was held in New York in March 1980.

References

Baeza-Yates, R. and Ribeiro-Neto, B. (1999) *Modern information retrieval*. New York: Addison Wesley.

Briggs, R. (1985) Knowledge representation in Sanskrit and artificial intelligence. *AI Magazine*, **6**(1): 32–39.

Cuadra/Elsevier (1989) *Directory of online databases.*

Greenhouse, S. (2000) The future of database indexing. *Key Words*, **8**(4): 125–126, 132. Updated version at http://mysite.verizon.net/vze2bpts/ASIkwarticle.htm

Jurafsky, D. and Martin, J. H. (2000) *Speech and language processing*. New York: Prentice Hall.

Klement, S. (2002) Open-system versus closed-system indexing: a vital distinction. *The Indexer* **23**(1): 23–31.

Lancaster, F. W. (1969) MEDLARS: report on the evaluation of its operating efficiency. *American Documentation* 20: 119–142, reprinted (1997) in *Readings in information retrieval*, ed. K. Spark-Jones and P. Willett. San Francisco: Morgan Kauffman.

Salton, G. (1988) *Automatic text processing: the transformation, analysis, and retrieval of information by computer.* New York: Addison-Wesley.

Sowa, J. F. (1999) *Knowledge representation: logical, philosophical and computational foundations*. London: Brooks Cole.

Witten, I. H , Moffat, A. and Bell, T. C. (1999) *Managing gigabytes: compressing and indexing documents and images.* San Francisco: Morgan Kaufmann.

Witten, I. H. and Frank, E. (2000) *Data mining.* San Francisco: Morgan Kauffman.

Harry Diakoff has been a consultant on database creation and indexing protocols for over 30 years. His clients have included many pharmaceutical companies and a few law firms representing the health industry. During most of this time he has been working as part of Kaim Associates of Wilton CT, but from 1996 until 2001 he consulted directly to Ovid Technologies, where he helped develop deduping and natural language capabilities. While Ovid was a public company, between 1994 and its sale to Wolters Kluwer in 1998, he also served on its board of directors. Email: harry.diakoff@verizon.net