

Paper presented at Electronic Age '96 A conference for the information industry 20-21 April 1996 Ranelagh House, Robertson NSW. To be published in Online Currents, the AusSI Newsletter 20(6):4-9, July 1996 and LASIE 27(3):58-65

Automatic indexing

Glenda Browne

© 1996 Glenda Browne.

Introduction

This paper will examine developments in automatic indexing and abstracting in which the computer creates the index and abstract, with little or no human intervention. The emphasis is on practical applications, rather than theoretical studies. This paper does not cover computer- *aided* indexing, in which computers enhance the work of human indexers, or indexing of the Internet.

Research into automatic indexing and abstracting has been progressing since the late 1950's. Early reports claimed success, but practical applications have been limited. Computer indexing and abstracting are now being used commercially, with prospects for further use in the future. The history of automatic indexing and abstracting is well covered by Lancaster (1991).

Database indexing

Extraction indexing

The simplest method for indexing articles for bibliographic databases is **extraction indexing** , in which terms are extracted from the text of the article for inclusion in the index. The frequency of words in the article is determined, and the words which are found most often are included in the index. Alternatively, the words which occur most often in the article compared to their occurrence in the rest of the database, or in normal language, are included. This method can also take into account word stems (so that *run* and *running* are recognised as referring to the same concept), and can recognise phrases as well as single words.

Computer extraction indexing is more consistent than human extraction indexing. However, most human indexing is not simple extraction indexing, but is **assignment indexing** , in which the terms used in the index are not necessarily those found in the text.

Assignment indexing

For assignment indexing, the computer has a thesaurus , or controlled vocabulary, which lists all the subject headings which may be used in the index. For each of these subject headings it also has a list of **profile words** . These are words which, when found in the text of the article, indicate that the thesaurus term should be allocated.

For example, for the thesaurus term *childbirth* , the profile might include the words: *childbirth*, *birth*, *labor*, *labour*, *delivery*, *forceps*, *baby* , and *born* . As well as the profile, the computer also has **criteria for inclusion** -- instructions as to how often, and in what combination, the profile words must be present for that thesaurus term to be allocated.

The criteria might say, for example, that if the word *childbirth* is found ten times in an article, then the thesaurus term *childbirth* will be allocated. However if the word *delivery* is found ten times in an article, this in itself is not enough to warrant allocation of the term *childbirth* , as *delivery* could be referring to other subjects such as mail delivery. The criteria in this case would specify that the term *delivery* must occur a certain number of times, along with one or more of the other terms in the profile.

Computer database indexing in practice

In practice in database indexing, there is a continuum of use of computers, from no computer at all to fully automatic indexing.

- No computer.
- Computer clerical support, e.g. for data entry.
- Computer quality control, e.g. checking that all index terms are valid thesaurus terms.
- Computer intellectual assistance, e.g. helping with term choice and weighting.
- Automatic indexing (Hodge 1994).

Most database producers use computers at a number of different steps along this continuum. At the moment, however, automatic indexing is only ever used for a **part** of a database, for example, for a specific subject, access point, or document type.

Automatic indexing is used by the Defense Technology Information Center (DTIC) for the management-related literature in its database; it is used by FIZ Karlsruhe for indexing chemical names; it was used until 1992 by the Russian International Centre for Scientific and Technical Information (ICSTI) for its Russian language materials; and it was used by INSPEC for the re-indexing of its backfiles to new standards (Hodge 1994).

BIOSIS (Biological Abstracts) uses computers at all steps on the continuum, and uses automatic indexing in a number of areas. Title keywords are mapped by computer to the Semantic Vocabulary of 15,000 words; the terms from the Semantic Vocabulary are then mapped to one of 600 Concept Headings (that is, subject headings which describe the broad subject area of a document; Lancaster 1991).

The version of BIOSIS Previews available on the database host STN International uses automatic indexing to allocate Chemical Abstracts Service Registry Numbers to articles to describe the chemicals, drugs, enzymes and biosequences discussed in the article. The codes are allocated without human review, but a human operator spends five hours per month maintaining authority files and rules (Hodge 1994).

Retrieval and ranking tools

There are two sides to the information retrieval process: documents must be **indexed** (by humans or computers) to describe their subject content; and documents must be **retrieved** using retrieval software and appropriate search statements. Retrieval and ranking tools include those used with bibliographic databases, the 'indexes' used on the Internet, and personal computer software packages such as Personal Librarian (Koll 1993). Some programs, such as ISYS, are specialised for the fast retrieval of search words.

In theory these are complementary approaches, and both are needed for optimal retrieval. In practice, however, especially with documents in full-text databases, indexing is often omitted, and the retrieval software is relied on instead.

For these documents, which will not be indexed, it is important to ensure the best possible access. To accomplish this, the authors of the documents must be aware of the searching methods which will be used to retrieve them. Authors must use appropriate keywords throughout the text, and ensure that keywords are included in the title and section headings, as these are often given priority by retrieval and ranking tools (Sunter 1995).

The process whereby the creators of documents structure them to enhance retrieval is known as bottom-up indexing. A role for professional indexers in bottom-up indexing is as guides and trainers to document authors (Locke 1993).

One reason that automatic indexing may be unsuited to book indexing is that book indexes are not usually available electronically, and cannot be used in conjunction with powerful search software (Mulvany and Milstead 1994).

Document abstracting

Computers abstract documents (that is, condense their text) by searching for high frequency words in the text, and then selecting sentences in which clusters of these high frequency words occur. These sentences are then used in the order in which they appear in the text to make up the abstract. Flow can be improved by adding extra sentences (for example, if a sentence begins with 'Hence' or 'However' the previous sentence can be included as well) but the abstract remains an awkward collection of grammatically unrelated sentences.

To try and show the subject content, weighting can be given to sentences from certain locations in the document (e.g. the introduction) and to sentences containing cue words (e.g. 'finally', which suggests that a conclusion is starting). In addition, an organisation can give a weighting to words which are important to them: a footwear producer, for example, could require that every sentence containing the words *foot* or *shoe* should be included in the abstract.

Computer abstracting works best for documents which are written formally and consistently. It has been used with some success for generating case summaries from the text of legal decisions (Lancaster 1991).

After recent developments in natural language processing by computers, it is now possible for a computer to generate a grammatically correct abstract, in which sentences are modified without loss of meaning.

For example, from the following sentence:

"The need to generate enormous additional amounts of electric power while at the same time protecting the environment is one of the major social and technological problems that our society must solve in the next (sic!) future"

the computer generated the condensed sentence:

"The society must solve in the future the problem of the need to generate power while protecting the environment" (Lancaster 1991). Text summarisation experiments by British Telecom have resulted in useful, readable, abstracts (Farkas 1995).

Book indexing

There are a number of different types of microcomputer based software packages which are used for indexing.

The simplest are **concordance generators**, in which a list of the words found in the document, with the pages they are on, is generated. It is also possible to specify a list of words such that the concordance program will only include words from that list. This method was used to index drafts of the ISO999 indexing standard to help the committee members keep track of rules while the work was in progress (Shuter 1993).

Computer-aided indexing packages, such as Macrex and Cindex, are used by many professional indexers to enhance their work. They enable the indexer to view the index in alphabetical or page number order, can automatically produce various index styles, and save much typing.

Embedded indexing software is available with computer packages such as word processors, PageMaker, and Framemaker. With embedded indexing the document to be indexed is on disk, and the indexer inserts tags into the document to indicate which index terms should be allocated for that page. It does not matter if the document is then changed, as the index tags will move with the part of the document to which they refer. (So if twenty pages are added at the beginning of the document, all of the other text, including the index tags, will move 20 pages further on).

Disadvantages of embedded indexing are that it is time-consuming to do and awkward to edit (Mulvany 1994). Indexers who use embedded indexing often also use a program such as Macrex or Cindex to overcome these problems.

Embedded indexing is commonly used for documents such as computer software manuals which are published in many versions, and which allow very little time for the index to be created after the text has been finalised. With embedded indexing, indexing can start before the final page proofs are ready.

Embedded indexing will probably be used more in the future: for indexing works which are published in a number of formats; for indexing textbooks which are printed on request using only portions of the original textbook or using a combination of sources; and for indexing electronically published works which are continually adapted. In some of these applications the same person may do the work of the editor and indexer.

The most recent development in microcomputer book indexing software is Indexicon (Version 2), an **automatic indexing package** .

Indexicon

Indexicon -- How it works

Indexicon is published by Iconovex , and is available as an add-on program for MS-Word and WordPerfect on IBM-compatible computers, and for MS-Word on the Macintosh. All versions cost US\$129. Indexicon 2.0 for MS-Word requires MS-Word for Windows 6.0 or above; a 386 or better CPU (486 recommended); Windows 3.1 and 8 MB RAM (Indexicon Spec Sheet 1996).

To use Indexicon, the book to be indexed must be available electronically in a word processing format. The user chooses from six levels of detail, and Indexicon creates an embedded index at that level using the indexing facility available with MS-Word or WordPerfect. The user can then edit the tagged entries in the original document. Indexicon indexes are subject to all the problems of embedded indexes, including the time-consuming editing process.

Indexicon comes with a primary lexicon containing about 55,000 words, and it allows the user to create specialised lexicons. The lexicons include index values for terms, which indicate how likely it is that a given term should be tagged as an index term. In general this depends on the degree of specialisation of the term, so that a commonly-used term is less likely to be indexed than a more specialised term.

The primary lexicon also includes compound terms, so that pairs of words such as *control tower* , or *remote control* , are always indexed as a pair.

All words in the lexicon also show what part of speech they are. Thus the word *lead* would have two parts indicating that it can mean the metal *lead* or the verb *lead* .

Indexicon will invert proper names if it recognises them, however if the surname carries another meaning (as with the names Brown, Miller, and Young) then the name is not recognised as such. Indexicon can also be set to omit proper names and geographic names if it can recognise them (Iconovex 1996).

Indexicon -- Reviews

Indexicon has been reviewed a number of times. The Indexicon Spec Sheet (1996) on the Internet says: "Indexicon is a tool capable of handling everyone's indexing needs". PC Magazine (13/9/94; quoted on the Indexicon Spec Sheet 1996) says: "With Indexicon, creating an index is as quick and easy as spellchecking".

However, a review of Indexicon Version 1.00b by Mulvany and Milstead (1994) found that it did not live up to the promises on the packaging that it was the "Standard for Indexing" and

could produce "professional quality indexes". In a response to this review, Steven Waldron, President of Iconovex, acknowledged many of the points raised, and stated "The purpose of INDEXICON ... is NOT to replace professional indexers" (Waldron 1994).

The Macintosh version of Indexicon was reviewed by Erfert Fenton (1996). He says: "Before I wrote this review I was skeptical of computer-based indexing programs. Having written it, I'm even more skeptical."

Fenton found that Indexicon missed many terms. When tested on a chapter on Macintosh fonts it missed the terms *pica*, *em dash*, and *leading* (pronounced *ledding*). It started many terms with adjectives (e.g. *slushy winter roads*) and it included many inappropriate entries (e.g. *Uncle Steve Yahoo*). The reviewer found much evidence of the fact that the computer did not understand what it was reading, and was therefore unable to make valid judgments.

In a test using Indexicon Version 2.0 to index a short article on the use of in vitro fertilization to save tigers from extinction, we identified the following problems:

Indexicon did not recognise and invert any of the names in the article. In one case this was because the person's surname had another meaning (*Ann Miller*); in the other two cases the name appeared in a string with other capitalised words and the whole string was indexed (e.g. *Leslie Johnston of National Zoo*).

Indexicon included some inappropriate entries (e.g. *Biologist's hopes*) and some strange constructions (e.g. *Reproductive tract, Nicole's* -- Nicole is a tiger).

Bengal tiger cubs was indexed in direct order, but *Tigress, Siberian* was inverted. Presumably this is because *Bengal tigers* is included in the lexicon as a compound word.

Indexicon does not generate cross-references so these must be identified and added by the indexer at the editing stage.

Finally, Indexicon did not group terms, so that *Tiger* and *Tigers* were given as separate entries.

In this exercise Indexicon set to the highest level of indexing indexed all important terms; in other experiments which we did many important terms were omitted, while non-significant terms were included.

Indexicon -- Potential uses

Iconovex states that Indexicon is suitable for use with documents which would not otherwise be indexed, and as a first step for professional indexers.

It is currently used to index manuals (e.g. corporate policy and procedure manuals), large contracts and large quantities of e-mail. Technical writers who index their own work have been using it as a first step in indexing.

Among indexers, Indexicon is most likely to be useful for specialists, who are more likely to take the time to create specialised lexicons, and to work with the program to enhance its efficacy in their special field. For journal indexing, where the same indexer works with similar material, in a consistent format, year after year, it might be worth taking the trouble to set up a specialised lexicon, and use Indexicon as a first step. But Indexicon is not good enough at picking key concepts and leaving out worthless ones, to be useful, in general, as an aid to indexing books.

If Indexicon improves, and if the embedded indexing software used in word processing programs improves, it may become more cost-effective to start indexing with Indexicon, and then enhance the index by editing.

As the ability of computer software to recognise personal names develops, it may also become useful as a tool for automatically generating name indexes (Feldman, Lawrence e-mail 15/03/96).

Effect of automatic methods on professionals

As computer programs become more sophisticated, and more information appears in electronic form, there will eventually be less 'traditional' indexing work available. This loss may be balanced in the short-term by an increase in the number of databases and an increase in the number of indexing and abstracting projects attempted. The proportion of freelance versus in-house work may also change.

Humans should still be used for important works, which perhaps can be identified by studying usage and citation patterns (Anderson 1993). Indexers and abstracters will have to become more selective, and decide on the quality of the works they might index and abstract, as well as the subject content.

If we remain better than computers we must show this, and indicate that there are economic returns (to the publisher) and academic returns (to the index or abstract user) from a quality index or abstract.

On the positive side, indexing and abstracting skills will be needed in the development of computer systems, and to check the output from computers. Indexers will be needed to set up and maintain thesauruses, and to train writers as 'bottom-up indexers' so that their work is readily retrievable.

Indexers will have to become entrepreneurial and computer literate. Indexers with skills in the related areas of computing, editing, librarianship and bibliography may be best suited to take advantage of new opportunities. We will have to be able to identify gaps in the organisation of knowledge and to fill those gaps in a commercially effective way. To do this we will have to be computer literate. Not only will we have to know how to use various computer tools for indexing; we will also have to know how information is organised and used electronically, so that we can best understand the needs and make our own contributions.

Acknowledgments

I would like to thank Terry Maguire, language director of Iconovex, the publisher of Indexicon, for a trial copy of the software, and prompt answers to all of my questions. I would also like to thank Jonathan Jerney and Bill Browne for their support and patience while I prepared this talk and paper.

References

Anderson, James D. 1993 ,

Indexing standards: Are they possible? What good are they? Why bother? In: *Indexing, Providing Access to Information: Looking Back, Looking Ahead, Proceedings of the 25th Annual Meeting of the American Society of Indexers*, Alexandria, VA, May 20-22. Port Aransas, Texas: American Society of Indexers.

Farkas, Lynn 1995 ,

Economics and the future of database indexing. In: *Indexers – Partners in Publishing, Proceedings from the First International Conference* , Marysville, Vic, March 31 to April 2. [Melbourne]: Australian Society of Indexers.

Fenton, Erfert 1996 .

Indexicon 1.0: Indexing program for Word 6. Macworld Communications.

Hodge, Gail M. 1994 .

Computer-assisted database indexing: the state-of-the-art. *The Indexer*. Vol. 19, No. 1, pp. 23-27.

Iconovex 1996. *Indexicon 2.0: Automated Indexing for Microsoft Word: User's Guide*. [Bloomington, MN]: Iconovex. Indexicon Spec Sheet 1996.

Koll, Matthew B. 1993.

Automatic relevance ranking: A searcher's complement to indexing. In: *Indexing, Providing Access to Information: Looking Back, Looking Ahead, Proceedings of the 25th Annual Meeting of the American Society of Indexers*, Alexandria, VA, May 20-22. Port Aransas, Texas: American Society of Indexers.

Lancaster F.W. 1991.

Indexing and abstracting in theory and practice. London: Library Association.

Locke, Christopher 1993.

Weaving the Social Fabric: Illuminating Manuscripts. In: *Indexing, Providing Access to Information: Looking Back, Looking Ahead, Proceedings of the 25th Annual Meeting of the American Society of Indexers*, Alexandria, VA, May 20-22. Port Aransas, Texas: American Society of Indexers.

Mulvany, Nancy C. 1994.

Embedded Indexing Software: Users Speak Out. In: *The Changing Landscapes of Indexing, Proceedings of the 26th Annual Meeting of the American Society of Indexers*, San Diego, California, May 13-14. Port Aransas, Texas: American Society of Indexers.

Mulvany, Nancy And Milstead, Jessica 1994.

Indexicon, The Only Fully Automatic Indexer: A Review. *Key Words*, Vol. 2, No. 5, pp. 1, 17-23.

Shuter, Janet 1993.

Standards for indexes: Where do they come from and what use are they? In: *Indexing, Providing Access to Information: Looking Back, Looking Ahead, Proceedings of the 25th Annual Meeting of the American Society of Indexers*, Alexandria, VA, May 20-22. Port Aransas, Texas: American Society of Indexers.

Sunter, Steve 1995.

Humans and computers: partners in indexing. In: *Indexers -- Partners in Publishing, Proceedings from the First International Conference*, Marysville, Vic, March 31 to April 2. [Melbourne]: Australian Society of Indexers.

Waldron, Steven 1994.

Message to INDEX-L@BINGVMB.BITNET on 31/10/94.

Bio

Glenda Browne, PO Box 307 Blaxland NSW Australia 2774, email glendabrowne@optusnet.com.au

Glenda is a freelance indexer with a background in biotechnology and information management. After a stint in sole charge of a small hospital library she moved into library teaching at Mt Druitt TAFE, which she now combines with indexing (in partnership with her husband Jonathan Jerme) and caring for their children. She has indexed articles for the CSIRO Index database, and has created book indexes on subjects ranging from Communication to Chemistry. Glenda's index to *Pharmacology and Drug Information for Nurses* was Highly Recommended in the 1995 AUSSI medal awards.