This is your introduction to the MIR Tutorial series. It attempts to answer the questions: What? For whom? Why? How? The "How" takes the form of interactive publishing in which you are invited to contribute. Part of our aim is to make the MIR computer indexing and retrieval techniques widely available, so we include in full the Free Software Foundation's GNU General Public License. This license provides the legal means to ensure there is the maximum freedom (and minimum restriction) for all who wish to understand, use, and further develop techniques of computerized indexing.

++

1. COMPUTER INDEXING AND RETRIEVAL TECHNIQUES

++

These tutorials are about people and information. People need information. The MIR (Mass Indexing and Retrieval) project has one objective: to make available leading edge technology which may be used to enable people to find information quickly and easily within large quantities of computerized data. The technology is being shared through five sets of tutorials, each accompanied by software with source code.

The tutorial series subtitle is "Finding Information in a Gigabyte World". A gigabyte is 1,073,741,824 characters of data. Visualize a stack of computer paper 140 feet high, or a library of 500 books, or 10,000 hours of reading. More and more, it is becoming commonplace for people to search through quantities of data of that magnitude. The one certainty is that no-one ever wants to read through a pile like that, even at computer speeds, in order to find an item of information. So our focus in this project is on computerized indexing and retrieval techniques. Well designed index structures and logic can reduce time for a complex search down to seconds or a fraction of a second.

The Mass Indexing and Retrieval project got under way in March, 1991. Five interactive tutorials based on the research are being released in the period September 1992 through September 1993. A reference text based on the series is scheduled for December 1993. Each tutorial has eight or more sections, and invites inputs from readers. The tutorials are on paper and available at a modest price. Accompanying each shipment are diskettes

containing all executable programs and source code described in that tutorial.

With certain clearly marked exceptions, the tutorials themselves are copyright, and not to be copied. In contrast, the software may be freely copied, used, revised, and further distributed within the terms of the Free Software Foundation's GNU General Public License.

What is meant by "interactive" tutorials? I believe that many minds are better than one, and that everybody gains through "open architecture" sharing. The quality of the final software and the final published version of the tutorials depends on your questions and suggestions. While we have worked on hundreds of different databases, I know that there are always other interesting challenges. That is why I am encouraging you to share technical insights, ideas, clearer wording, source code amendments and even whole new programs. I look to you in particular to expand the range of worked examples; send in real world data that may be included. Tutorials are meant to be a dialogue. This to me is the exciting part of a learning situation... the more people pitch in with their ideas, and the more enthusiasm they show, the more everybody learns (including the teacher!)

There is an incentive to sharing within two to three months... free updates. If a purchaser of a tutorial sends us content that is significantly helpful, we will modify the master version and send an updated copy of the tutorial and the related software. This way you can (at no extra cost) receive more than one release of a tutorial and get the benefit of other people's improvements (and see your name in the acknowledgements) well in advance of final publication of the cumulative series.

Watch for sections like this in the interactive tutorials:

>>>>>>> QUESTION:

Are you with me so far? I may be too close to this stuff, and assume that you should know what is in my mind. What parts need clarification? Send in your comments. Make a copy of the RESPONSE file which comes with the software; it is already personalized. Fill in the relevant sections, and identify any other files that you are sending. The RESPONSE file contains the FAX and e-mail numbers and the mail address. If sending anything lengthy by normal post, please put it on a PC-compatible diskette.

<<<<<<<

If you are a dedicated technical person, you may wish to look ahead at the TUTORIAL ONE topic on "First Steps in Data Analysis". It is included on the sampler so that you may get a feeling of the level and style of the tutorials. But please come back to these introductory topics. They give the context that is necessary to take full advantage of the technical detail. When you get a complete copy of TUTORIAL ONE, give special attention also to the second topic; it shows how you can be involved in this project to make the tools and technology even better.

We continue with an overview of each of the five tutorials, the final cumulative publication, and the schedule of releases.

++++++++++++++++++++++++++++++++
1.1 Tutorial ONE...
Database Analysis
++++++++++++++++++++++++++++++++

>>>>>>> QUESTION:
Contest!! "Database Analysis" is a humdrum title. We could use snappy headings for everything... for the tutorials, the topics within each, and even individual sections. Maybe our Table of Contents could be as neat as Jerry Weinberg's The Secrets of Consulting... "The Law of the Jiggle", "The Edsel Edict", "The Bigness is Not the Horse", and so on like that. Make notes as you read, and send in a batch of headings.
<<<<<<<

[This section is copied from topic 1.2 in the first tutorial.]

The purpose of MIR Tutorial ONE is to enable you to analyze computerized data from an indexing perspective.

The first topic, source code guidelines, explains the perspectives that have been built into the software that is provided with the tutorials. People who wish to improve on the technology are shown how to share their insights and C language source code.

Methods of data gathering affect the cost, the quality and the complexity of the task of indexing. An index adds value to data, so we pay attention to some marketing considerations.

Data analysis has to do with recognizing various forms in which data is accumulated, and detecting the inconsistencies (common in large sets of data) that make indexing more challenging. Data format offers possibilities and imposes limitations that will face searchers who wish to extract information. How might the data be structured in a way that better suits the needs of searchers? The reader is provided with a variety of software tools for this critical data analysis function.

The ability to identify patterns in byte sequences quickly is critical to keeping indexing costs low. We examine a series of software tools for this purpose.

Worked examples are provided of the analysis stage. These topics are at a "nuts and bolts" level... use such and such a program, here is the input, here is the output, and here is what the results mean. The sequence is from simplest to most complex... simple ASCII text, ASCII with markup, fielded text, fixed length records, the addition of packed numbers, then various forms of binary data

Data deblocking is explained at this stage since it may be required in order to finish analysis of the data.

At the end of TUTORIAL ONE, the participant has detailed exposure to the techniques of data analysis, and is able to use a selection of analysis tools (source code provided) to recognize and interpret a wide range of data types.

++
1.2 Tutorial TWO...
Secrets of Data Preparation
++

The first topic sets out a simple ASCII text format which makes data suitable for automated indexing. Careful planning of data sequence and layout can speed up response to search requests. What the searcher sees later depends on a series of decisions made during data preparation.

Example: What is to be the unit of search (article, paragraph, computer record, a fixed length record, etc.)? A second topic delves into other issues in data organization: the use of invisible fields, pointers, parameter controls, data that must remain accessible to other software and the handling of multimedia data.

Standard Generalized Markup Language enhances the end user's ability to control layouts of records found during search. It may be embedded in data without hindering automated indexing. We look at how to distinguish flexible versus fixed display, how to handle oversize tables, etc.

Data pre-processing describes the task of converting data to a standardized production format. In some cases, it's easy. If the analysis has been thorough, there should be few surprises. Yet experience shows that setting up the pre-processing sequence can still be the most expensive aspect of all. We look at a series of standardized tools to make the job easier and more efficient.

Worked examples show how to use combinations of standardized tools and custom software. We look at how to extract data from several kinds of typesetting codes. This section is intended to be as practical as possible, so readers are invited to submit sample real world data.

One of the surprises for you in this tutorial is a detailed analysis of why compression before indexing makes more sense than would at first appear. The standardized ASCII layout can be used as an intermediate step toward a compressed version which greatly increases the indexing capacity of a personal computer. We examine some integerizing techniques and software.

At the end of TUTORIAL TWO, the user can make decisions about the layout of data, and implement those decisions using a variety of data conversion tools. Source code for these tools is provided with TUTORIAL TWO. You will have been exposed to issues in writing custom data conversion tools. The user is able to compress large databases into integerized format, in order to make it practical to index them on a personal computer.

++

1.3 Tutorial THREE...
Keys to Automated Indexing

++

Indexing basics start with an explanation of index formats, and how they may be combined through Boolean logic. We look at grouping indexes within separate field lists, and also at how to tag index items within a global index list.

The topics on search term selection show how to go beyond simple word indexing to enable search on word fragments, phrases, topics and numeric or date ranges. Files are created for each "field" in a database. We look at means to upgrade these field files and to ensure strict quality control over the indexes.

Specialized index preparation leads us into "fuzzy search" of alternate verb forms (search on "is", calls up "were", "shall be", "was", "isn't", "to be", etc.) and nouns (possessives, plurals, etc.) Search on synonyms and correlates is related; the power depends on how much context is taken into account to distinguish homonyms... words of one spelling with radically different meanings. Pattern indexing provides extra speed where the searcher may specify extended word sequences. The issue of "relevance" of found records carries the discussion further into automated subject recognition.

Automated indexing is critical to limiting costs; one efficient set of software programs (called an "inversion engine") can be used to build the indexes for virtually any data originally expressed in alphabetic letters, digits, and other keyboard characters. The structure of the index is critical to how quickly the retrieval software can perform Boolean combinations... ((this-word OR that-phrase) AND something else AND NOT another term). The automated indexing software creates indexes in a format geared to high speed Boolean operations when used for search.

We look at software (source code provided) for two "inversion engines", one using strings, the other working from integerized data.

At the end of TUTORIAL THREE, the user is familiar with the tools necessary to set up and create computer indexes, tailoring the index types according to the needs of searchers in the target database.

++++++++++++++++++++++++++++++++++

1.4 Tutorial FOUR... Search Engines and Information Retrieval

++++++++++++++++++++++++++++++++++

This is the most technical of the five tutorials. Everything up to this point has been the concern of the indexer. Now we turn to the "run time" or retrieval software. Retrieval describes the search process... specifying a search, performing Boolean logic on combinations of terms, identifying data that meets the search criteria, and making the selected data available to the searcher.

Under the topic dealing with Search Engine Servers, we review an SFQL (Structured Full Text Query Language) server which is provided with TUTORIAL FOUR. Alternate server options (CD-RDx et al) will be reviewed.

Search Engine Client (interface) software is deliberately left outside the "copyleft" software set; no single interface can encompass the range of features desirable for all data types and search situations. We comment on current issues in standardization.

Search extensions include:

> optimization of index structures;

> search across multiple databases at a time; and

> dynamic definition of search objects.

By the end of TUTORIAL FOUR, the user has available the know-how and software to analyze, prepare, index, and provide search capability for a diverse range of data types and search requirements. Any engine-independent interface built to SFQL specifications may be used to implement search at high speed across large quantities of data.

++

1.5 Tutorial FIVE...

Related Topics and Applications

++

The list of related topics and applications will continue to grow, based on reader comments on earlier tutorials. Our experience in CD-ROM preparation has already led us to include the following areas of interest:

> Encryption: We believe that encryption merely dissuades the idle browser and raises costs to the determined criminal. We discuss straight-forward methods that serve these purposes admirably. Even where the technique is known, it takes an inordinate amount of computer time for the thief to identify the seed values.

> Data cleaning combines the benefits of indexing with spell checking to enable low cost cleanup of massive databases.

> Records and Information Management (RIM) is a full discipline in its own right. The technology and plummeting costs of full text archiving is bringing about a revolution in RIM philosophy and methods of records retention. There are some simple tricks that can be applied to archiving with spectacular results.

> Correlation studies using indexed retrieval and high speed Booleans can change the nature of research. A cell in a correlation table turns out to be a search count. Mainframe, move aside. The PC is here.

++++++++++++++++++++++++++++++++++

1.6 The MIR Tutorials: The Book and CD-ROM

++++++++++++++++++++++++++++++++++

As the five interactive tutorials are released there will be an ongoing revision and updating process. This will reflect your responses and improvements on the content, and encompass many of the samples and suggestions that you have made. The first four reworked tutorials will be put together with Tutorial FIVE and be published as an ongoing reference work. We will decide closer to the final publication date (December 1993) whether the final version will be

> loose-leaf, or

> bound as a reference or text book, and/or

> electronic (ASCII, WordPerfect, and PageMaker files) on a CD-ROM.

Whatever the form of the tutorial text, all programs, source code and worked examples will be supplied on a CD-ROM.

++
1.7 Timing of successive releases
++

Here is the distribution and response schedule:

TUTORIAL NUMBER	Tutorial plus software distributed	Responses requested
ONE	September 1992	November 1992
TWO	December 1992 *	February 1993
THREE	March 1993	May 1993
FOUR	June 1993	August 1993
FIVE	September 1993	November 1993
CD-ROM/Book	December 1993 **	n.a.

* For those who purchase two or more tutorials, the second tutorial is accompanied by a three ring binder, adequate for the whole series.

** The CUMULATIVE edition is provided without charge to everyone who purchased all five interactive tutorials.

The above shipment dates are subject to a thirty day grace period (acts of God, etc.)

The major unknown in the Mass Indexing and Retrieval project is the readiness of the marketplace to deal with copyleft and the notion that, through sharing, the benefits of $800,000 in development can be picked up for less than $500. This is the old marketing problem of perception of value. We are taking the risk of volume pricing; we are betting that there are enough people in the field who can recognize value based on the sampler and the first tutorial. Marpex Inc. reserves the right to discontinue the project if there is insufficient demand.

What about organizations with a burning need to proceed in advance of the above schedule? Some tutorial sets can be made available prior to the release dates under special circumstances. We are open to your proposals. Your technical, financial and hardware resources are welcome, provided they contribute to advance the project within our open architecture ground rules. And, within limits, we do offer consulting services.

1.8 Summary

This completes our introduction to the series of five tutorials on how to enable people to retrieve information from large accumulations of data. Related high speed indexing and retrieval software is being distributed under the "copyleft" rules of the Free Software Foundation. Interactive publishing enables you to:

> study the techniques in the tutorials and examples;

> put the source code to use, personally or commercially, without payment of license fees;

> further develop the computer source code; and

> contribute your insights.