NAME-TRACING USING THE ICL CONTENT ADDRESSABLE FILESTORE

A G WARD

ICL

The paper describes the Regional and National Tracing Systems being developed for the UK enabling public service records to be identified and located using unreliable and incomplete name and address information, and explains how CAFS provides an answer to a problem which would otherwise be insoluble. Topics covered include the sizing calculations, the main database organisation, the primary and secondary indexing structures and particular reference is made to the use of CAFS in 'fuzzy' and quorum searches.

1. INTRODUCTION

In common with many administrative organisations which have direct correspondence with and about members of the general public, certain UK Government departments are faced with the problem of relating to members of the public locating records when adequate reference numbers are not quoted. As in a Directory Enquiries service, the available name and address details must be translated into appropriate references before record access is possible.

There will be a reference number which uniquely identifies an individual. There will also be additional information to identify the local office that deals with his affairs.

The clerical methods used hitherto have been unable to provide a tracing service on all documents, and the unofficial rule has been that full tracing would only be attempted on documents passing certain criteria.

Alternative tracing technologies that have been evaluated include the searching of microfiche directories, and computerised searching using standard methods of Soundex coding of surnames. The pros and cons of these techniques are well known. It is only necessary to say here that it was decided that neither could provide a full tracing service at an economically justifiable cost.

The rest of this paper describes the approach being adopted, in which the ICL Content Addressable Filestore 'CAFS' is used to provide an outstanding efficient full tracing facility, first at Regional level, and subsequently at the National level.

The whole country is split into 12 approximately equal-sized regions, each containing the records of some 3.5 million individuals. A prototype regional tracing service with a capacity of about 0.7 enquiries per second will handle all the traces for which information is held within the region. Eventually users will access the national tracing service, which will hold some 48 million records and support a load of more than 7 traces per second.

CAFS is highly relevant to the requirements of this application, because it is a searching engine, able to scan data stored in conventional files on standard discs at a rate approximating (with curr ent discs) to 1 megabyte per second. A search task may contain up to 16 criteria and these may be combined with either boolean or quorum logic. Its effectiveness in filtering stored data far exceeds the capabilities of any conventional

Proceedings of the Tenth International Conference on Very Large Data Bases.

Singapore, August, 1984

software techniques. Since the mainframe driving CAFS is relieved of the searching task, its functions are reduced to those of maintaining the dialogue with terminal users and instructing the CAFS hardware where to search and how much to search in each enquiry. Thus a much more modest mainframe can be used, and the capital cost of the system correspondingly reduced.

The remainder of this paper concentrates on the methods of indexing required to provide adequate search task focussing in the context of the Department's name-tracing application.

2. REGIONAL TRACING STRATEGY

The names, addresses, reference numbers and other relevant information concerning the individuals in the region are held in a large ISAM file. The key of a record consists of the first five characters of the surname, followed by the reference number. (The purpose of which is to make the key unique.) Other fields are mandatory or optional, and fixed or variable length. The average length of a record, weighted according to the frequency of the occurrences of the optional fields, is 133 characters. However, for sizing purposes and to cater for the possibility of expansion we allow for a record size of 200 bytes.

A search enquiry is handled - with elegant simplicity - by using the primary ISAM index to identify the range of blocks in this file which contain all the entries for the given surname or surname stem. These blocks are then searched by the CAFS unit for any entries which contain the specified surname and address words. If any of the blocks has a chained overflow block then another CAFS search is required to scan these overflow blocks.

A hashed-random HRAM file is used in parallel as a reference number index. It is keyed by the reference number and indicates the record(s) in the main ISAM file corresponding to that reference number. It is thus used in tracing individuals for whom the reference number is known but not the name and/or address and/or other relevant data - such as the local office reference.

An outline transaction within the normal online service is used to drive the CAFS unit. It displays a screen format into which the user enters those details of the name and/or address and/or reference number which he has, and which he judges will be most effective in the search. These details are used by the application code in a transaction processing environment to initiate a CAFS search on the ISAM file. The resulting taxpayer details - if any - are displayed on the same screen format. If there are too many to fit onto one screen, a paging mechanism is used to display later ones once the user has seen and dealt with the earlier entries.

In practice, it is rarely necessary for the user to enter complete words. The first few characters of the more significant words in the address are usually more effective in providing an accurate match than a complete surname and a complete address word. This saves keying time for the user, besides minimising the effect of inconsistency of spelling between the stored records and the trace documents. After relatively very short experience, users become adept at deciding how much or how little to enter, and at picking out which terms within a name and address will provide the best discrimination.

There is a facility to specify 'fuzzy' characters (eg B?AT, which may be satisfied by BEAT, BOAT, BRAT, and so on.) In addition, if a search fails to find a single address, the user may specify a 'quorum', saying, in effect, : 'Well, can you find any address which satisfies 3 out of the 4 words I specified?' Both of these facilities are mapped directly onto CAFS hardware facilities by the tracing application software.

If the preliminary access to the ISAM low-level index indicates that too many tracks would need to be searched by the CAFS unit, or if no surname is given, the user is informed and invited to refine his search criteria.

The size of the main ISAM file in a single region is derived as follows:

1. size of regional index
   = 5 million names
2. block size
   = 6,447 bytes
3. packing density
   = 90%
4. average record length
   = 200 bytes

5. hence records per block
   = 29

6. hence blocks in file
   = 172,414

7. hence total data size
   = 1060 megabytes

3. NATIONAL TRACING STRATEGY

The significantly greater number of records in the National Tracing facility, together with the increased message throughput which that facility must support, require that the number of tracks searched per enquiry must be considerably reduced below the value which would be established by the relatively simple indexing strategy that can be used in the Regional services.

One approach to this problem would be to split the main ISAM file into a set of 'search units', each of one or more tracks, and hold a separate orthogonal index which identifies all the keywords that occur in the address and the set of search units that contain any occurrences of these keywords. The size of a search unit needs to be estab lished by sizing calculations.

A general approach to accessing a CAFS file may in fact require additional indexes holding, (in the case of this application), surname, parts of post-codes, region and district keys, and so on; or an amalgamated index holding all of these, Indeed, it is worth consider ing whether the primary ISAM index is required at all for enquiries on the file and whether all enquiries can be handled through such secondary indexes.

For speed of access, it is essential that the key to the secondary index are stored as fixed-length records, and it is most appropriate to store these as records in an ISAM file. The key of each record consists of the first few characters of the keyword. In general, the more characters are used in this stem, the finer the resolution in terms of reducing the number of search units identified by each key-word, but the more keyword entries there will be, and the longer it will take to identify any particular key-word. The trade-off between stem-size and resolution must be established by inspection.

Note that the CAFS hardware imposes a minimum length on the records which it is to search, and stems in the keyword index would fall below the limit. However the part of the record beyond the stem field is used to hold pointers, counts, etc, to assist in the processing of the search, and hence the limitation is avoided.

When processing a search for a name and address it is possible to search the keyword index for each significant word in the address, as specified by the user, and from this obtain a set of the search units which contain that word. The CAFS search can then be confined to those search units which are obtained by taking the intersection of all the individual sets.

In practice, in this case, the user may specify a quorum in the normal CAFS sense, and the computation of the set of units to search is more complex than merely taking the intersection of each keyword's search unit set. For example, if A,B,and C are the sets of search units corresponding to three keywords, the simple computation if all three terms must appear in the chosen search units is ;

A AND B AND C

If it is sufficient to find units with any two out of the three, then the computation is ;

(A AND B) OR (B AND C) OR (A AND C).

Names of types of road (CLOSE,AVENUE, STREET,ROAD,LANE,etc) would not normally be required as part of an address-keyword index, although some (CLOSE, LANE for example) could be significant in a surname index. However, for simplicity in creating the indexes it is best to handle all entries in an address as significant, and to set them up in the keyword index. Their entries would identify a great many search units, which would make them useless for delimiting a search. However, a skilled operator would not use such common words in a search, and thus their presence in the keyword index has no significance other than the space they occupy. But if these words were omitted from the keyword index, then the software which accesses that index would not know whether a failure to find a keyword meant that the word was not known or whether it was too common.

4. SIZING THE INDEXES

To establish the optimum size o f the keyword stem it is necessary to obtain some feel for the number of keywords involved in the national index. Information to contribute to such calculations has been available from

two files (A and B)of different sub-populations of Great Britain.

File A contains about 2.16 million addresses. Each address contains on average 5.85 words. There are 197,523 separate words in these addresses. About 10% of these are spurious, resulting from misspellings, mistypings, and so on. In addition, about 40% of the words contain numeric characters, and are thus house or flat numbers, or parts of postcodes.

Examination of the numbers of occurrences of each stem in this file, for various stem sizes, indicates that the probabilities of a given stem being contained in fewer than n addresses, for various values of n, are as shown in Table 1. This table shows, for example, that for a 4-character stem :

50% of all discret stems occur in 3 addresses or less.

70% of all discrete stems occur in 11 addresses or less

100% of all discrete stems occur in 401,186 addresses or less.

Some of the interim counts were obtained for the stem sizes of 20,8,5,4, and 3 characters at regular intervals. The total numbers of entries for each stem size are shown in Table 2.

The stem which occurs most commonly is 'GLAS', which occurs in 18% of all File A addresses.

The corresponding figures for File B, which contains just over 500,000 address-es, are as shown in Tables 3 and 4.

The commonest stem in File B is 'RD' which occurs in 42% of all addresses.

The differences between the tables for File A and File B are significant File B addresses contain far fewer words, have fewer words per address, and many more instances of the words that do exist. We may assume that this is typical for all large con-urbations.

File A, on the other hand, has far more rare words, (54% of all words appear in only one address!), and more words per address. We may assume that this is more typical of the country as a whole, as it contains both large conurbations - Edinburgh and Glasgow - and many rural areas. It is thus an appropriate base for extrapolation to a full national index.

Of the various stem sizes considered, a stem of 4 characters appears to give the best secondary index size. It is estimated that there will be about 70,000 entries for the whole country, though Tables 2 and 4 show that it is difficult to predict the level at which the number of entries will tail off, and allowance must be made for a great deal of variance in this figure.

Examination of the numbers of occurrences of selected words from File B, after analysing various numbers of addresses, shows that the number of occurrences is directly proportional to the number of addresses analysed, although the ratio :

$$\frac{\text{number of occurrences}}{\text{number of addresses}}$$

is very different for different words.

The tabulated figures give the probabilities in terms of the numbers of addresses containing any given words. For sizing purposes we need to know the corresponding probabilities for the numbers of search units containing these words.

Let : N = number of distinct stems resulting in 1 or more addresses

t = number of addresses hit by most common stem

ni = number of stems which hit just i addresses (for i = 1, 2,3, ...t)

$pi = \frac{ni}{N}$ i = 1,2,3, ...t

TF = total number of search units in ISAM file

We wish to know the number of search units hit by a stem that hits x addresses. The probability that a particular search unit is hit by a stem

= 1 - Prob (not hit)

= 1 - Prob (not hit by 1st address and
not hit by 2nd and
not hit by 3rd and
not hit by xth)

$$= 1 - (1 - \frac{1}{TF})^x \qquad (1)$$

Thus the number of search units hit by a stem which hits x addresses is :

$$TF\ (1 - (1 - \frac{1}{TF})^x) \qquad (2)$$

If we have a given number of search units in a file, we can use Formula (2) to estimate the number of these search units holding occurrences of a given word as the number of addresses

containing that word varies. The numbers for a population size of 48 million addresses, and a file size of 552,000 tracks, each of 85 addresses, (ie with a single-track search unit size) are shown in Table 5. The significant data in this table are those showing the percentage of search units hit. It is apparent that a keyword need only occur in 2% of all addresses before it becomes unusable for refining a search.

Corresponding tables could be produced for other search unit sizes.

From this point on, the most important design consideration is the efficiency of the representation chosen for search uni sets in the keyword index. Next it is necessary to ensure that the various main and index files are so laid out that the optimum combination of indexed accesses and CAFS searches can be supported. Thereafter, conventional sizing calculations can be used to establish the CAFS disc usage for the average task, and matching this against the time available yields a figure for the number of CAFS engines required to be concurrently active.

5. CONCLUSION

This paper has given a short outline of a system design which exploits the searching capabilities of CAFS to provide an economic and efficient solution to a well-known information retrieval problem. It will be supplemented in presentation by confirmatory figures derived during the first year of live implementation.

TABLE 1

Prob	addresses per separate word			
%	stem=20	stem=5	stem=4	stem=3
100	398,698	399,056	401,186	412,844
95	102	234	317	737
90	26	76	180	277
85	11	29	72	219
80	6	14	33	142
75	4	8	18	70
70	3	5	11	36
65	2	4	7	19
60	2	3	5	12
55	1	2	4	8
50		2	3	5
45		2	2	4
40			2	3
35			1	2
30				2
25				2
20				1
Total stems	197,523	97,814	54,343	22,862

Probabilities of random words occurring in n or less addresses (File A)

TABLE 2

Addresses	number of stems for stem sizes of:				
	20	8	5	4	3
100,000	39,793	35,459	24,874	17,502	9,811
200,000	53,309	50,914	33,853	22,631	11,907
300,000	65,089	61,830	39,918	26,066	13,318
400,000	76,050	71,930	45,352	28,963	14,449
500,000	85,706	80,876	50,130	31,513	15,409
600,000	94,896	89,345	54,454	33,689	16,200
700,000	103,058	95,820	58,154	35,608	16,947
800,000	111,461	104,524	61,944	37,504	17,612
900,000	119,601	111,980	65,569	39,252	--
1,100,000	134,354	125,475	72,078	42,448	19,236
1,250,000	143,275	133,547	75,869	44,279	19,857
1,500,000	158,157	147,034	82,029	47,126	20,754
1,750,000	173,699	161,136	88,354	50,059	21,653
2,000,000	188,042	174,108	94,130	52,707	22,422
2,164,968	197,523	182,665	97,814	54,343	22,862

Number of stems for different address counts and stem sizes (for sections of File A)

TABLE 3

Prob	addresses per separate word			
%	stem=20	stem=5	stem=4	stem=3
100	211,457	211,457	211,457	211,457
95	202	369	740	2,890
90	93	155	293	1,070
85	64	94	167	531
80	48	69	108	319
75	37	52	77	201
70	30	40	59	128
65	24	32	45	87
60	20	26	35	63
55	16	21	28	42
50	13	17	21	29
45	10	13	16	19
40	8	10	12	13
35	6	8	9	7
30	4	6	6	5
25	3	4	4	3
20	2	3	3	2
15	1	2	2	1
10		1	1	
Total stems	19,314	13,273	8,695	3,849

Probabilities of random words occuring in n or less addresses (File B)

TABLE 4

Addresses	number of stems for stem sizes of:				
	20	8	5	4	3
100,000	14,452	14,131	10,348	6,802	2,928
200,000	16,528	16,106	11,607	7,583	3,283
300,000	17,720	17,233	12,331	8,069	3,530
400,000	18,576	18,041	12,836	8,391	3,695
500,000	19,314	18,718	13,273	8,695	3,849

Number of stems for different address counts and stem sizes (File B)

TABLE 5

Addresses	%	Units	%
48	0.0001	47	0.0085
100	0.0002	98	0.0177
480	0.0010	469	0.0850
1,000	0.0020	976	0.177
4,800	0.0100	4,669	0.846
10,000	0.0208	9,683	1.75
48,000	0.100	44,954	8.14
100,000	0.208	89,531	16.2
200,000	0.417	164,541	29.8
300,000	0.625	227,385	41.2
400,000	0.833	280,036	50.7
480,000	1.00	315,938	57.2
600,000	1.25	301,104	65.4
700,000	1.43	392,066	71.0
800,000	1.67	418,007	75.7
900,000	1.86	439,740	79.7
1,000,000	2.08	457,948	83.0
2,000,000	4.17	535,975	97.1
3,000,000	6.25	549,270	99.5
10,000,000	20.8	552,000	100.0

Proportions of units corresponding to proportions of addresses for single-track search units.