Rank Order Distributions and Secondary Key Indexing

W. B. SAMSON AND A. BENDELL
Dundee College of Technology, Bell Street, DUNDEE, DD1 1HG

The performance of a secondary index depends greatly upon the distribution of secondary key values, especially when these are not unique. The nature of these distributions is discussed and a model for the minimum indexing time is proposed. Normally, at the time the database is designed, little is known about the nature of the data to be stored. A technique is described for modelling the underlying distribution of a secondary key population, based on a small sample from that population. Alternative indexing strategies may be compared on the basis of this model distribution at an early stage of design. Possible strategies for improving indexing performance are discussed.

1. INTRODUCTION

Database performance modelling presents the designer with a host of problems. For example, when attributes of records are to be indexed the assumptions made about these records and attributes can have a profound effect on the performance estimates. It is well known that in a file, certain records are much 'busier' than others. This effect is carried through to the indexes by which these records are located. Performance estimates can therefore be made very inaccurate if an assumption of equal frequency of access is made. For example, a 'busy' record could also be an awkward record to access – it could involve overflows which could increase the access time. This assumption could therefore lead to underestimated access times which can be embarrassing for the design team.

A rather less well known problem is that of indexing on non-unique attributes – for example surnames, keywords, physical characteristics. There are varying degrees of naivety in approaching this problem. Probably the most disastrous approach is that of assuming that the attributes are in fact unique. Figure 1 illustrates the relative access times, for various loading factors, of a distribution of unique attributes, compared with a distribution of non-unique attributes (Samson and Davis[1]). The index model used is a hash table with ten index entries per page and quadratic hash overflow. A somewhat less naive approach is based on the assumption that all attributes are non-unique to the same degree – ie an average frequency per attribute is assumed to be the actual frequency of all attribute values in the index. Although the average access times are now not dissimilar to those encountered in practice, the range of access times is much greater. This can lead to serious problems when a busy record also happens to be indexed on a high frequency attribute. For example, in a file of criminal records a criminal named Smith is more likely to lead to very long access times based on surname alone than a criminal named Maitland-Titterton.

Non-unique attributes can be classified according to their distribution when arranged in descending order of observed frequency. Such a distribution is called a 'rank order distribution'. These distributions merit careful study by all those involved in the evaluation of secondary index performance. In this paper, a technique is described for the comparison of alternative indexing strategies, at an early stage of database design. A model for optimal index performance is proposed and ways of improving the performance of a badly sub-optimal index are indicated.

The authors will concern themselves exclusively, in this paper, with rank order distributions of key frequency. The matter of some records being busier than others is a separate problem and will not be considered further here.

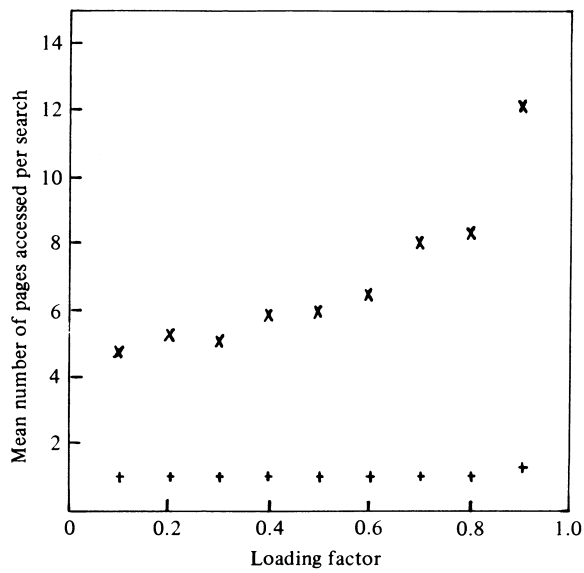

Figure 1. Comparison of hash table access times for unique (+) and non-unique (X) keys.

2. RANK ORDER DISTRIBUTIONS

It has been known for many years that attributes of entities which are not, by their nature, unique have a frequency distribution which is far from uniform. Examples of this include the frequencies with which human surnames occur. In the United Kingdom the surname Smith is about twice as common as the next most common surname. Similarly, the letter 'e' is the most frequent letter in a piece of text. A set of fingerprints consisting of ten ulnar loops is very common while there is probably no person on earth with ten whorls.

Probably the most famous early account of the rank order distributions associated with such phenomena comes from Zipf[2] who studied, among other things, the frequencies with which words appeared in text.

Non-unique attributes form the bulk of material stored in most databases and records are often accessed through indexes based on these attributes – ie secondary key

indexes. Some indexing techniques will store the secondary key value along with a list of corresponding primary key values. The length of such a list will depend on the underlying distribution of the secondary keys and so the index must be tailored to suit the distribution.

Few authors have taken the distribution of non-unique attributes into account when designing secondary key indexes. The most important reason for this is that these distributions are not well understood and that such theoretical studies as have been made have not appeared in the computing literature. Nevertheless, as figure 1 shows, the effect on index lookup times can be dramatic. Here the distribution is the special case of a Zipf distribution in which the frequency of the nth most frequent secondary key value is given by

$$freq(n) = freq(1)/n$$

where $freq(1)$ is the frequency with which the most frequent secondary key value occurs.

In this paper a model is proposed which provides designers with a value for minimum indexing time for comparison with any proposed design. An empirical technique will also be demonstrated to show how the form of the rank order distribution for a population may be estimated from a sample of attribute values.

Whilst much of the early work on rank order distributions centred around the rank-size rules or Zipf[4] distributions, there is no reason why any multi-parameter rank order distribution cannot be constructed and fitted to the observed data using some estimation technique. The rank-size rule does not, in general, provide a completely satisfactory fit and can sometimes be objected to on distributional grounds (eg Yule[3]).

It is not practical within this paper to examine the full range of distributional forms that have been proposed. Instead, we limit our attention to the Zipf distributions, which are in many ways the simplest rank order distributions, and at the same time provide an effective, flexible and widely accepted model for the majority of observed distributions for secondary keys.

3. ZIPF DISTRIBUTIONS

In general, the probability of occurrence of the rth most frequent attribute value in a randomly chosen record is

$$P(r) = k/(r+c)^a \quad r = 1,2,...,N; \; -1 < c < \text{infinity}; \; a > 0 \tag{1}$$

where a, c, and N are parameters to be estimated and, to ensure that the probabilities sum to unity,

$$k = 1/\sum_{i=1}^{N} (r+c)^{-a} \tag{2}$$

This distribution is a step function which approximates closely to a straight line in the log-log plane when c is 0. The effect of increasing or decreasing N is to increase or decrease the length of the tail of the distribution. A large value of N means that a large proportion of attribute values appear once in the observed distribution and so are effectively unique. This, in turn, implies more convenient secondary indexing because only a small proportion of the attributes fall into the early non- unique categories.

The effect of a low value of a is to cause non-uniqueness to extend further along the rank axis. In the limit, when $a = 0$, the distribution is a rectangular one with equal probabilities for all attribute values. At the other extreme, in the limit as a tends to infinity, the probability of the most frequent attribute value is 1, and all other probabilities are 0. When c is zero, the distribution approximates to a straight line in the log-log plane with slope $-a$. However, when c is non-zero, the distribution in the log-log plane no longer approximates to a straight line but is curved. The continuous analogue of this line is asymptotic to a line with slope $-a$ in the log-log plane. The curvature is convex upwards for $c > 0$ and convex downwards for $-1 < c < 0$. Another point of interest is that the effect of increasing the parameter c from an arbitrary value is almost indistinguishable from the effect of reducing the parameter a. The problem is compounded when we are dealing with uncertainties due to sampling. This gives rise to problems when an attempt is made to estimate values of c and a for an observed distribution.

4. MINIMUM INDEX LOOKUP TIME

It is assumed that all entries are equally likely to be consulted. So, for an arbitrary key, the mean number of pages to be accessed for an index lookup to find all entries for that key is the weighted mean of the number of memory pages in which entries for each key are to be found; ie

$$\frac{\sum_{i=1}^{n} f_i p_i}{\sum_{i=1}^{n} f_i} \tag{3}$$

where

f_i is the frequency of entries for the ith key;

p_i is the number of memory pages in which entries for the ith key are to be found;

and n is the number of different key values.

This caters for both the case where entries for one secondary key value spread over several pages; and the

Table 1

a	Least possible mean number of pages accessed
0	1.00
0.3	1.00
0.4	1.01
0.5	1.04
0.6	1.19
0.7	1.66
0.8	2.96
0.9	6.16
1.0	13.01

Showing least possible mean number of pages to be accessed for a secondary key distribution where
c = 0
number of categories = 10000
loading factor = 0.7
page size = 512 bytes
index entry size = 16 bytes
total number of entries = 26214

case where entries for several secondary key values appear on the same page giving one access per key value even though it occupies only part of a page.

If it is assumed that the size of an index entry is fixed and that a maximum of m entries may be stored on a page, then the smallest possible value for p_i is given by

$$p_i = ||f_i/m|| \quad (4)$$

Where $||a||$ indicates the smallest integer which is greater than or equal to a.

Using equations (3) and (4) we can calculate the minimum possible index lookup time for any frequency distribution of key values. The values in Table 1 were calculated using a simulation program (see below) for a number of rank order distributions where the c parameter is equal to zero, the number of categories in the underlying population is equal to 10000 and the a parameter varies in value between zero and one. Key values are effectively unique when $a=0$.

Index designers should compare their indexing performance based on simulation, as suggested in this paper, with the optimal values computed as shown above. If the performance is found to be badly sub-optimal, the reason will almost certainly be that entries for different high frequency keys are mixed on a high proportion of pages, with the result that some values of p_i will be substantially larger than the optimal values calculated as in equation (4) above. Any difference between the actual value of p_i and the optimal value will be caused by overflows which are due to collisions between entries for the ith key and entries for other keys.

The solution to a problem of excessive index lookup time will, of course, depend on the architecture of the indexing mechanism being used, but it is important to ensure that entries for different high frequency keys are directed to different pages, and that this separation is maintained throughout the overflow process. This will almost certainly mean that a logical 'slot' in the system will coincide with a physical page. A comparison of some standard indexing methods with optimal performance figures will be the subject of a future paper. Preliminary results suggest that mixing entries for a high-frequency key with those for much lower frequency keys have only a small influence on the mean number of pages accessed.

5. ESTIMATION OF A ZIPF DISTRIBUTION FROM A SAMPLE

The literature on estimation in Zipf rank order distributions has largely concentrated on large sample (total number of observed attributes) techniques, eg Carroll[5] and reference [6]. In the simple case where an a-priori estimate of c can be obtained, then since from (1) $\ln(P(r))$ is linear in $\ln(r+c)$, it follows that estimates of N and a, and hence of the probabilities of unobserved attribute values, can be obtained by fitting a straight line (say by least squares) to the plot of $\ln(\textit{frequency})$ against $\ln(r+c)$. However, estimation for small samples is of practical interest since this is all the data that may be readily available at the time of systems design. With small samples it is possible that the rank order of category frequencies in the sample differs substantially from that in the population, especially in the tail of the distribution. In addition, the observed shape in the sample must differ somewhat from the underlying shape of the population due to the discreteness inherent in the sample as well as to sampling fluctuation. The implication is that the standard large sample methods of estimation which treat the sample as if it approximates closely to the population could cause serious error.

Bendell and Samson[4] discussed the problems of estimation in the 1-parameter Zipf model and their extension into the 3-parameter form.

The three parameter form of the rank order distribution involves complex estimation and is expensive of computer time. First estimates for values of c and a are determined from an iterative fit to the categories with ranks 1 to 3. The observed number of categories may be taken as an initial estimate of N. We then perform a search employing minimum chi-squared estimation to numerically identify the optimum estimates of the parameter values. That is, the estimates of c, a and N are chosen so that

$$\chi^2 = \sum_{i=1}^{N} (O_i - E_i)^2/E_i \quad (5)$$

is a minimum, where O_i denotes the observed frequency in the sample category of rank i and E_i denotes the corresponding expected frequency for the category of rank i, in the sample. These expected frequencies, E_i, are determined by taking the average over 100 samples from the theoretical distribution corresponding to the c and a values under consideration.

Unfortunately, in approaching the optimum point the parameter values are very interdependent, and the optimum is, typically, located in a narrow valley. A typical contour diagram for fixed N is shown in figure 2. The fit, for a population with a large number of categories, is not sensitive to N over a large range of N values. The figure suggests that some form of pattern search algorithm should be used to determine the minimum. A number of 'textbook' algorithms were tried, but they all suffered from the problem that values of

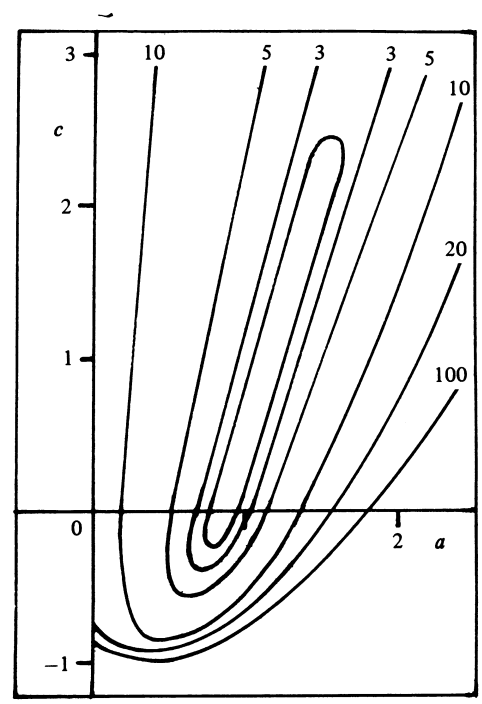

Figure 2. Contour diagram showing chi-squared values.

chi-squared are determined using monte-carlo simulation and so successive values of chi-squared for a given point were not identical. As the 'valley floor' is reached, the search tends to wander wastefully and to retrace its steps on occasion. This is very expensive of computer time and so the simple heuristic method described below was used. This heuristic uses the fact that the valley floor is relatively straight.

The algorithm for determining the minimum chi-squared value operates by finding the minima in a for two fixed values of c. These two minima are joined by a straight line (up the centre of the valley) and the minimum is then determined on this line, giving the best values for c and a. This algorithm is much less time consuming than any of the more general pattern search methods and leads to similar results and so it is preferable to the other methods.

6. EXAMPLE

As an example we consider the frequencies of occurrence of surnames in the Isle of Man telephone directory for 1974, where the total number of records is 10741. The total number of different surnames is 3345. The number of unique surnames is 1968. The most common surname occurs 189 times.

This population has been completely observed and its distribution is shown in figure 3. A small sample of 400 was drawn at random from this population and the minimum chi-squared algorithm applied to estimate the c and a parameters for the underlying population. The resulting estimated population distribution is shown as a broken line in figure 3. This can be compared with the true underlying distribution from which the sample was drawn.

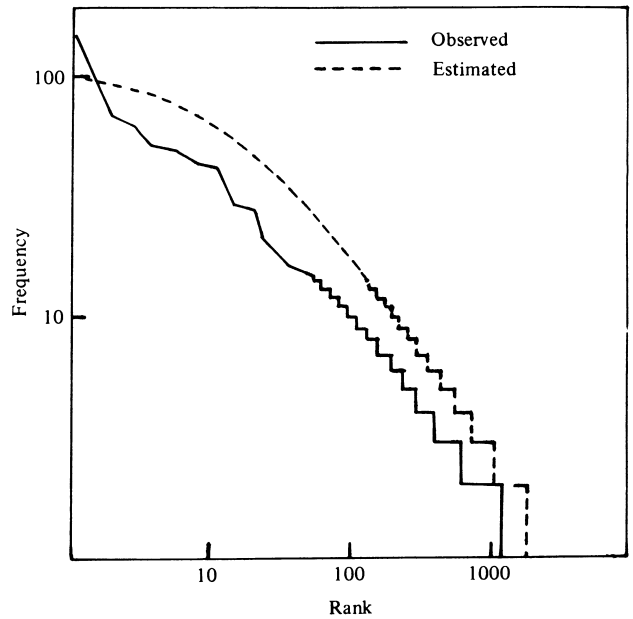

Figure 3. Isle of Man telephone directory. Rank-order distribution of surnames.

6.1 Conclusions

This work described in this paper illustrates the fact that a relatively small sample from a secondary key distribution may be used to select a good indexing strategy from those available, at modest cost. On the basis of our experience we would recommend that index designers should sample the values of secondary keys which are to be indexed and obtain estimates for the parameters of the underlying distribution using the method described in this paper. Optimal indexing times may be calculated and compared with the simulated performance of the index. If the performance is found to be significantly sub-optimal then the index should be re-configured according to the suggestions made in section 4. If, on the other hand, the performance is similar to the optimal one then the designer may be confident that he has a well configured index.

Acknowledgements

The authors are grateful to members of the PRECI Collaboration for helpful suggestions, and particularly so to Dr S M Deen, the Coordinator of the PRECI project. One of the authors (WBS) acknowledges financial support from the Science and Engineering Research Council under grant GR/B/82288.

REFERENCES

1. W.B. Samson and R.H. Davis, Search times Using Hash Tables for Records with Non-unique keys, *The Computer Journal*, Vol 21, p210 – 214, (1978).
2. G.K. Zipf, Human Behaviour and the Principle of Least Effort, Addison Wesley, (1949).
3. G.U. Yule, A Statistical Study of Vocabulary, Cambridge University Press, (1944).
4. A. Bendell and W.B. Samson, The use of rank order distributions for the estimation of the probabilities of rare events., *Proceedings of the Third National Reliability Conference*, 29 April to 1 May, 1981, National Centre for Systems Reliability, Warrington, p 2B/1/1 – 2B/1/11, (1981).
5. J.B. Carroll, The Psychological Record, Vol 2, p379, (1938).
6. D.R. McNeil, Estimating an Author's Vocabulary, *Journal of the American Statistical Association*, Vol 68, p92 – 96, (1973).