Prime Number Coding for Information Retrieval

By A. H. Cockayne and E. Hyde

While investigating the possibility of transferring a punched-card file of organic chemical compounds to a magnetic-tape file, a technique was used for coding the file items which is simple, leads to simple programming for the retrieval of items from the file, and, under the circumstances obtaining for the given file, resulted in savings in storage space.

Introduction

During the last six years a unit, called the Fine Chemical Service, has built up a collection of some 25,000 organic chemicals, together with a file of relevant physical, chemical, biological, etc., properties. A typical question made of the file, and the only sort of question which will be considered in what follows, is to request all those compounds recorded which possess certain specified structural features, e.g. all aromatic amino-acids. The majority of the 18,000 samples issued annually by the Fine Chemical Service, in fact, result from demands of this nature. Due to the large number of variables it is not possible to arrange the file in an order which will appreciably simplify the search for information thus specified. At present this search would be carried out by (Hollerith) punched-card methods, but as an experiment it was decided to program the search for a Pegasus computer with magnetic tape. This led to the method of coding the chemical compounds using prime numbers which is described later, and which so far as we know is novel. The method is described, in the main, against the background of the Fine Chemical Service collection as it is thought that this will make it less abstract, but there is no difficulty in making the method more general, as is pointed out in the final paragraph.

Details of the File and Punched-Card System

The main file of typed documents is kept in order of specimen number, a serial number which gives no indication as to the nature of the compound. In addition a file of 80-column punched cards is kept which has been designed to facilitate the extraction from the file of compounds with given structural features. The information contained in the punched cards is essentially of two sorts:

(i) the five-digit specimen number punched according to the normal punched-card conventions. This gives a direct cross-reference to the main file of typed documents;

(ii) information specifying the structural features of the compound.

The difficulties in making a choice of features which will go to make up a workable system should not be underestimated. Our scheme embraces approximately 250 different structural features (e.g. ring systems, amines, chlorides) and 250 positions on a punched card have been allocated, one to each feature. If a feature is present in a molecule the corresponding hole is punched in the card: otherwise not. To find all those compounds possessing a given set of structural features the punched-card file is passed through a Hollerith group-select sorter (suitably plugged to specify the required features) which will extract those cards which have holes punched in all the positions corresponding to the given set. The small number (of the order of a hundred) of cards thus extracted are first scrutinized for the molecular structure of the compound which is written on each card, and those which pass this test are used for entering the main file via the specimen number also written on the punched card. Quite a high proportion of the cards found by the mechanical search might be rejected on the scrutiny, for no claim is made that the system will produce only those cards required. Rather it is looked upon as a method of reducing the area of the manual search by a factor of 100 or 1,000.

Form of Information on Magnetic Tape

When the decision had been taken to experiment with the replacement of the punched-card file by magnetic tape, and the mechanical search by a computer search, the problem arose as to what was the best form for keeping the information on the tape. Since the master file would never be abandoned for policy reasons there was no point in attempting to store information relating to properties of the compounds on magnetic tape; the specimen number was adequate, just as it had been on the punched-card system. However, in the case of the information specifying structural features a change was called for.

By analogy with the punched-card system one could have set aside 250 bits for recording structural features and, if a compound possessed a particular feature, the corresponding bit would have been made unity, and otherwise zero. But for any one compound the number of structural features actually occurring is usually less than 10 and invariably less than 20, so that less than 8% of the bits allocated would have been unity on any one occasion. Thus only a small fraction of the 2^{250} possible configurations of 250 bits would have been realized, and inefficient use would have been made of the storage space. This suggests that more efficient systems of

recording the information might be available, which do not, of course, make it more difficult to establish the presence or absence of a structural feature. The system described below takes advantage of the small proportion of features which occur in any one compound, and is, in the case of the Fine Chemical Service collection, more efficient than the straight binary system.

Coding the Structural Features

The first 208 primes have been allocated to 208 of the structural features. (The drop from 250 to 208 will be explained later.) Care has been taken, based on experience gained with the file on punched cards, to allocate the smaller primes to the more frequently occurring features. For any one chemical compound the primes corresponding to its structural features are selected and multiplied, and the resulting compound (in the mathematical sense) number is used as the code for the chemical compound. No ambiguity arises if the number so found is called the *compound number*.

Advantages of the use of Compound Numbers

The test for the presence of a particular set of features, consists of dividing the compound number by a factor, formed from the primes corresponding to the particular set, and testing for zero remainder. Since an integer can be factorized in only one way, this process will select those compounds which have (at least) the given set of features. The programming of this test is no more complicated than the use of collate and not-equivalent operations, followed by a test for zero result which would otherwise have been required, and in the case of the Fine Chemical Service collection the computer time taken (on Pegasus) is roughly the same (about 6 millisec compared with from 1 millisec to 11 millisec). This method of coding, therefore, fulfils one of the requirements, namely that it should be not too difficult to detect the presence or absence of a feature. On a decimal machine with built-in division there would probably be a considerable simplification as a result of using this method.

It has been found in practice that, with the allocation of primes adopted, the compound numbers did not exceed 2^{76} (which is two Pegasus words), and consequently the storage requirements were cut to less than one-third, for this type of information. When the specimen collection number is taken into account in working out total storage requirements, one finds that the overall saving is 50%, i.e. four Pegasus words per item instead of eight. Since the computer is able to "keep ahead" of the magnetic tape, the running time of a search depends on the length of tape used (and the time to punch out results which for these purposes is a constant) and consequently the time for a search is approximately halved. This is the main advantage of the system in the case of the Fine Chemical Service collection.

A further point to note about the use of compound numbers is that fewer categories of structural feature are needed. For instance, in the case of chlorine atoms in a compound, it would be desirable to distinguish chlorides from dichlorides, and two bits would be allocated, one to -chloride, and the other to -dichloride. In the case of compound numbers, only one prime is used to denote chlorine, and if chlorine occurs twice it is raised to the second power. Thus, in general, one needs fewer primes than bits to cover all cases. This accounts for the drop from 250 bits to 208 primes mentioned above.

General

Although the above description has been entirely in terms of the Fine Chemical Service collection, it is obvious that the method could be used in other cases where the decision whether to select an item from a file or not depends on whether satisfactory replies are received to a set of yes–no questions, e.g. in the case of many edge-punched files. For by a simple extension one could test for the absence of one set of properties and the presence of another set (i.e. break the test into two division sums and two conditional transfers of control).

A considerable difficulty arises, in the general case, in deciding whether the use of primes and compound numbers as described will be advantageous or not. This is because it is difficult to find a general method of estimating the largest compound number which will occur which can be usefully applied in particular cases. In the case of the Fine Chemical Service collection one would have estimated that the largest compound number would be

$$(p_{208})^{17} \doteqdot 6 \times 10^{52}$$

since the largest number of factors allowed for any one compound is 17. This is so much in excess of the largest compound number found, i.e. 10^{22}, as to be useless in deciding whether the system using compound numbers will save storage space. However, in any particular case it should be a simple matter to find the most complex items in a file, and obtain a very good estimate of the highest compound number likely to occur.

Acknowledgement

The authors wish to express their gratitude to the directors of Imperial Chemical Industries Limited for permission to publish this paper.