Mathl. Comput. Modelling Vol. 16, No. 4, pp. 69–90, 1992
Printed in Great Britain. All rights reserved

0895-7177/92 $5.00 + 0.00

THEORY OF SEARCH KEYS AND APPLICATIONS IN RETRIEVAL TECHNIQUES USED BY CATALOGERS

L. EGGHE
Limburgs Universitaire Centrum, Library*
Universitaire Campus, B-3590 Diepenbeek, Belgium
and
Universitaire Instelling Antwerpen
Speciale Licentie Informatie en Bibliotheekwetenschap
Universiteitsplein 1, B-2610 Wilrijk, Belgium

(Received March 1991 and in revised form October 1991)

Abstract—This paper constructs a model for studying the performance of search keys of several types (such as, e.g., author/title keys of the form 4/4, 3/3, 3/1/1/1, and so on), and gives a criterion for deciding whether or not to use one (or several) more slashes (/) (i.e., more truncated title words) in a certain system (e.g., an automated library catalog). Some mathematical theorems on search keys are proved, given the probability distribution of letters (more general: symbols) in words. We then study the effect (on search key performance) of enlarging the " alphabet," i.e., when adding new symbols, that can be used in forming the search keys. Changing the alphabet might cause a change of optimal search keys. Criteria for this (not) to happen are given. The last part of the paper deals with the difference in average performance (averaged over all possible systems) of search keys with less or more slashes (/). In general, we can prove that, on the average, introducing slashes does not improve the retrieval performance.

1. INTRODUCTION

Search keys are an important tool in the retrieval of documents. Say we have an automated library catalog. OPAC's (i.e., Online Public Access Catalog's—as the catalogs for use by the non-professionals, e.g., a library user) usually restrict the searching to author, title or subject, since these are the most user-friendly ways, but more experienced searchers, such as catalogers, can save much more time by using a so-called "truncated search key" input. Indeed, catalogers (per definition) are people that add the indexed version of library material (e.g., books, journals) to the library catalog, thus, presenting a much simplified version of the document in question to the catalog. As such, catalogers want to know whether or not a certain book (or other document) is already in the system: if so, they only have to put in the local aspects of the book in question (e.g., shelf-marks); all the other aspects (e.g., subject indexing, author, title, editor, etc.) can be taken from the already catalogued document. For this very frequent retrieval exercise, catalogers better use a quick and effective method to retrieve documents. Truncated search keys are the standard techniques in this matter.

Let us give an example. Let us take the book by L. Egghe and R. Rousseau entitled "Introduction to Informetrics." One can find this book very quickly in a library catalog (if it is there!) by using one of the several possible keys:

- a 3/3-key requires the input EGGINT, i.e., the first 3 letters of the first author's last name concatenated with the first 3 letters of the first (meaningful) word of the title (or author/title may be reversed into title/author);

*Permanent address.

Typeset by $\mathcal{A}_{\mathcal{M}}\mathcal{S}$-TEX

- a 3/1/1/1-key requires the input of EGGITI, i.e., the first 3 letters of the first author's last name concatenated with the first letter of the first meaningful word of the title, and the first letters of the next 2 words of the title.

Other examples are: a 4/4-key (e.g., EGGHINTR), an 8-key (title)(e.g., INTRODUC), a 3/3/1-key (e.g., EGGINTT), and so on. In all these cases, one must define the symbols (i.e., if one is using $*$, $-$, and so on to form the keys, or not), one must say what is meant by a "meaningful" word of the title, and one must define what to do in case the author's name and/or the title is too short, and hence, does not have enough symbols to form the complete key.

In the sequel we assume that these fuzzy aspects have been settled. Hence, we assume that in all cases we know exactly how to form a certain search key.

It is clear that the overall ideal objective is to find zero documents in case the desired one is not in the database, and to find one document in case it is in the database. In practice, however, since by using keys one only uses a very simplified "image" of the document, it is very well possible that more documents satisfy the same key. This causes extra work for the cataloger to distinguish between these documents, and to choose the right one: this must be avoided as much as possible.

Of course, the longer the search key, the more effective it is (at least when using comparable keys, such as 3/3 versus 4/4), but this requires more time for the cataloger to put in (and this, every time that one searches a document).

Several articles deal with this problem, mainly written by staff members of OCLC (Online Computer Library Center), who introduced this technique and studied their practical concern for the OCLC database (see [1–11]). However, to the best of my knowledge, there do not exist models for search keys in which the performance can be deduced by mathematical methods. In this paper, we show that such a mathematical approach is very well possible, and that we can deduce practical conclusions from it. The paper is divided as follows.

In Section 2, we fix the numbers $(p_{ij})_{i,j=1,\dots,N}$ (N = number of symbols to be used), where p_{ij} is the probability that symbol j succeeds symbol i as the first two letters of the first word of, e.g., the title. We also simplify the problem by studying the performance of the 1/1-key versus 2-key. We measure the performance by the average number of documents retrieved when using the 1/1-key, respectively, the 2-key. We show that the intuitive idea of "a 1/1-key is better than a 2-key" is not always true: it depends on the numbers p_{ij}. We also prove three theorems, dealing with three cases: the 1/1-key better or worse than the 2-key and the 1/1-key performs equal when compared to the 2-key. An exact and applicable algorithm is given for a library system, to decide which of the two keys is best in this particular system.

The next section generalizes the above findings to more general search keys; the principles remain the same, however.

Section 4 starts from a fixed alphabet, with fixed probabilities $(p_{ij})_{i,j=1,\dots,N}$, and studies the problem: what happens if we enlarge the alphabet (i.e., when we allow new symbols to be used in the formation of the search keys). In general, a change of alphabet may result in another optimal search key to be used. Criteria for this not to happen are given.

A further generalization, in Section 5, deals with the following problem: study the overall average performance of 1/1-keys versus 2-keys, when the $(p_{ij})_{i,j=1,\dots,N}$ are variable (but N is fixed). This problem can be considered as the study of the "universal" performance of 1/1-keys and 2-keys over all library systems. This problem is far from trivial, but has theoretical interest (using higher dimensional geometry and analysis).

Section 6 makes the link with concentration theory (cf. [12,13]) and Lorenz-curves; the last section generalises the results of Section 5 and Section 6 to more general search keys. Concentration theory (originating from econometrics) has been a major tropic of research by R. Rousseau and also by the author (cf. [12,13]). When applying (and further developing) the concentration theory (i.e., the theory of inequality) to informetrics, one has recognized that many informetric problems can be modelled via the so-called Lorenz order (cf. [12,13] again). Here we also found this relation.

2. PERFORMANCE OF THE 1/1-KEY VERSUS THE 2-KEY FOR A GIVEN LIBRARY RETRIEVAL SYSTEM

Suppose we have a fixed library retrieval system. Hence, it is possible to determine the probability p_{ij}, $i,j = 1,\dots,N$, being the probability to have symbol j succeeding exactly symbol i, in a certain word. More specifically, we can take the first meaningful word from the title as this word, and to look for these two symbols in the beginning of this word, but this is not necessary. Let us define a 1/1-key as a key formed by two symbols, the first symbol coming from another word than where the second symbol comes from. Then a 2-key is defined as a key formed by two consecutive symbols in the same word. Most frequently, a 2-key is formed with the first two symbols of the first meaningful word of the title and a 1/1-key by the first symbol in the first author's last name, concatenated by the first symbol in the first meaningful word in the title (a so-called author/title search key). A 1/1-key can also be formed by the first symbol in the first meaningful word in the title concatenated with the first symbol in the next word in the title. We do not specify this here since we do not need it. What we want to study is the average number of documents retrieved in the library system when using 1/1-keys and 2-keys. The lower this average number is, the better, since these keys are used in retrieval by catalogers. This aspect could be called the "performance" of search keys, or also the "distinguishing power" of search keys.

So, we have a fixed library system, represented by the numbers $(p_{ij})_{i,j=1,\dots,N}$. Note that all p_{ij} satisfy $0 \le p_{ij} \le 1$, and

$$\sum_{i,j=1}^{N} p_{ij} = 1. \tag{1}$$

Furthermore,

$$p_i = \sum_{j=1}^{N} p_{ij} \tag{2}$$

represents the probability for the symbol i to appear as the first symbol of a word. We assume that this is also the probability for symbol i to appear as the first symbol in the second word that we use in the formation of our 1/1-key. Note that also $\sum_{i=1}^{N} p_i = 1$.

Now, the probability of having symbol i as the first symbol of the first word, and symbol j as the first symbol of the second word is $p_i\,p_j$ (due to probability independence, a reasonable assumption in this stage—see also note 2 below). Let there be n documents in the library system. Then, the number of documents in the library system with the 1/1-key "symbol i/symbol j" is

$$n\,p_i\,p_j. \tag{3}$$

This case appears with a probability $p_i\,p_j$, when a cataloger uses the 1/1-key: indeed, the key "symbol i/symbol j" appears as the retrieval request of the cataloger also with a probability $p_i\,p_j$. (We assume here that the same probabilities apply; this is quite natural: we only suppose that the occurrence of symbols (in the respective words) is the same for the documents that are already *in* the database as for the documents that we *are going* to put into the database—see also Note 2 below.)

So, the average number of documents found with the 1/1-key in this library system is ($n\,p_i\,p_j$ documents with probability $p_i\,p_j$, and summing over all symbols)

$$\bar{x}_{1/1} = n \sum_{i,j=1}^{N} p_i^2\,p_j^2 = n\left(\sum_{i=1}^{N} p_i^2\right)^2. \tag{4}$$

Note that Equation (4) is also equal to (by (2)):

$$\bar{x}_{1/1} = n\left(\sum_{i=1}^{N}\left(\sum_{j=1}^{N} p_{ij}\right)^2\right)^2, \tag{5}$$

expressing $\bar{x}_{1/1}$ in function of the p_{ij}'s.

Analogously, a 2-key can be described; now the probabilities p_{ij}, instead of $p_i p_j$, are used. We find that the average number of documents with the 2-key in this library system is ($n p_{ij}$ documents with probability p_{ij}, and summing over all documents)

$$\bar{x}_2 = n \sum_{i,j=1}^{N} p_{ij}^2. \tag{6}$$

There is no hope of proving $\bar{x}_{1/1} < \bar{x}_2$, $\bar{x}_{1/1} > \bar{x}_2$ or $\bar{x}_{1/1} = \bar{x}_2$ in general, due to the following examples.

Examples. Take $N = 2$ and $p_{11} = 0.1$, $p_{12} = 0.2$, $p_{21} = 0.3$ and $p_{22} = 0.4$. Then $\bar{x}_{1/1} = 0.3364\,n$ while $\bar{x}_2 = 0.3\,n$. Hence, here $\bar{x}_{1/1} > \bar{x}_2$ (i.e., the 2-key is better here). Take now $p_{11} = 0.1$, $p_{12} = 0.3$, $p_{21} = 0.2$ and $p_{22} = 0.4$. Now, $\bar{x}_{1/1} = 0.2704\,n$ and $\bar{x}_2 = 0.3\,n$. So, now $\bar{x}_{1/1} < \bar{x}_2$. Finally, take $p_{11} = 0.16$, $p_{12} = p_{21} = 0.24$ and $p_{22} = 0.36$. Now $\bar{x}_{1/1} = \bar{x}_2 = 0.2704\,n$. Note that the knowledge of n is not necessary for these comparisons.

Conclusion. We have already an important conclusion: if we want to know what search key to use in our own library system, we must determine the numbers $(p_{ij})_{i,j=1,\dots,N}$: they determine which key is the best to use: if $\bar{x}_{1/1} < \bar{x}_2$, use the 1/1-key; if $\bar{x}_{1/1} > \bar{x}_2$, use the 2-key; in the coincidental case, it is the same. In [14], an extensive experiment was performed in the network comprising my university library and three university libraries in Antwerp. For monographs, one found the following results: for $n = 373,230$ (the number of documents) and $N = 27$ (the alphabet-size), one had $\bar{x}_{1/1} = 1026.01 << \bar{x}_2 = 4267.15$. Right now, they use there an author/title key of size 4/4, and one wonders if allowing more words from the title would result in smaller numbers of retrieved documents. Based on the above result, we could advise that a further introduction of slashes (e.g., 4/2/2, ...) would indeed be beneficial. Note, also, that the calculation of $\bar{x}_{4/4}$ versus, e.g., $\bar{x}_{4/2/2}$ is very time-consuming. In the case of $\bar{x}_{4/4}$, there are $2\,(27)^4 = 1,062,882$ probabilities to be calculated (see Section 3, for an explanation of this number).

Note 1. Remark that, of course, we always have that a 2-key is better than a 1-key: $\bar{x}_2 < \bar{x}_1$ since, obviously,

$$\sum_{j=1}^{N} p_i^2 > \sum_{i,j=1}^{N} p_{ij}^2,$$

(if there is at least one i for which there are at least two $p_{ij} \neq 0$—an evident requirement). Indeed, the above relation follows by (2).

Note 2. In the building up of formulae (3) and (4) we assumed two things:

(a) To have symbol i as the first letter in the first word is independent from having symbol j as the first letter in the second word.

(b) The probabilities of the symbols *in* the database are the same as the probabilities of the symbols of the books that *we are going* to add to the database.

In practice, none of the above assumptions are perfectly correct. For the sake of simplicity, however, they are needed. Furthermore, we think we can assume the above statements. Indeed, with (a) we really mean that in 1/1-keys there is "much more" independence between the symbols than in 2-keys, which is acceptable. More refined studies could be in order, however. Assumption (b) is—we think—even more evident: in short terms, the type of the documents that a library buys is the same as the ones that are already in the system. Of course, here also a more refined model might be considered. We think, however, that in this stage, the simpler model must be developed and studied before we go to more sophisticated theories.

Important remark. The main objects, as defined in this section, are $\bar{x}_{1/1}$ and $\bar{x}_2$, and they both involve squares of probabilities ($p_i p_j$ or p_{ij}). The reason for this is that these averages are the average number of documents that a *cataloger* encounters when using one of these search keys. This is the *real* important problem. In such kind of studies, search keys have a certain "fatalistic" effect: what we want is, of course, to keep the $\bar{x}$'s as low as possible, but one always has that the search key (be it 1/1-, or 2-, or whatever) with probability p in the database (hence, for which there are $n\,p$ documents in the database), will be used by the cataloger with a probability p. Otherwise stated, the search keys corresponding to a lot of documents in the database are the most frequently encountered by catalogers—a negative finding, indeed! (See also the remark after Lemma 1.) Still, this type of study is the most important one. Studying only the average number of documents in the database, for a certain search key *does not* reflect the real performance of this key, as explained above. This is, however, what is done so far in articles on search keys. We agree, of course, that the larger key one wants to study (e.g., 4/4-key as above), the more difficult it gets to investigate the $\bar{x}$'s (e.g., $\bar{x}_{4/4}$ versus other key-averages). We claim, however, that a knowledge of $\bar{x}_{1/1}$ versus $\bar{x}_2$ explains a lot concerning the addition (or deletion) of a slash (/), also in more general search-keys.

We leave as an open problem to study possible relationships (for a given search key) between the average number of documents in a database (easy to calculate in practice, but less important in itself) and the average number of documents found by a cataloger's retrieval (as explained: difficult to calculate in practice, but really important).

We now proceed with three theorems, each representing on of the three cases discussed above.

THEOREM 1. *If for all $i,j = 1,\ldots,N$,*

$$p_{ij} = p_i\,p_j, \tag{7}$$

then $\bar{x}_{1/1} = \bar{x}_2$. The converse is not true: i.e., there exists a system of symbol probabilities (p_{ij}) such that $\bar{x}_{1/1} = \bar{x}_2$, but for which there exists at least one couple (i,j) for which $p_{ij} \neq p_i\,p_j$.

PROOF. If $p_{ij} = p_i\,p_j$ for every $i,j = 1,\ldots,N$, then $\bar{x}_{1/1} = \bar{x}_2$, obviously by (4) and (6). The converse is not true: take p_{ij} satisfying (7) and interchange (e.g., in $N = 2$) p_{11} and p_{12}, and also p_{21} and p_{22}. Denote the new system by $(p'_{ij})_{i,j=1,2}$. Then, since $(p_{ij})_{i,j=1,2}$ satisfies $p_1 = p_{11} + p_{12}$, $p_2 = p_{21} + p_{22}$, $p_1 + p_2 = 1$, $\sum_{i,j=1}^{2} p_i^2\,p_j^2 = \sum_{i,j=1}^{2} p_{ij}^2$; the same is true for $(p'_{ij})_{i,j=1,2}$, but $p'_{ij} \neq p'_i\,p'_j = p_i\,p_j$, for every $i,j = 1,2$. A concrete example is offered by: $p'_{11} = 3/16$, $p'_{12} = 1/16$, $p'_{21} = 9/16$, $p'_{22} = 3/16$. Then ($p'_1 = 1/4$, $p'_2 = 3/4$),

$$\sum_{i,j=1}^{2} p'^2_i\,p'^2_j = \sum_{i,j=1}^{2} p'^2_{ij} = 0.390625,$$

$p'_1 + p'_2 = 1$, $p'_1 = p'_{11} + p'_{12}$, $p'_2 = p'_{21} + p'_{22}$ but $p'_{11} \neq p'^2_1$, $p'_{12} \neq p'_1 p'_2$, $p'_{21} \neq p'_2 p'_1$, $p'_{22} \neq p'^2_2$. ∎

The result of Theorem 1 says that, if the occurrence of symbols as the second in a word is independent of what the first symbol is, then the two keys perform alike. We will, further on, consider more general cases.

LEMMA 1.

(a) *For variable $(x_i)_{i=1}^{M}$, such that $0 \leq x_i \leq 1$, for all $i = 1,\ldots,M$, and $\sum_{i=1}^{M} x_i = 1$, we have that $\sum_{i=1}^{M} x_i^2$ is minimal for all $x_i = \frac{1}{M}$.*

(b) *Let $(x_{ij})_{i,j=1}^{M}$ be such that*

$$\sum_{i=1}^{M}\left(\sum_{j=1}^{M} x_{ij}\right)^2 = k \tag{8}$$

is fixed. Then $\sum_{i,j=1}^{M} x_{ij}^2$ is minimal, if $x_{ij} = \frac{1}{M}\sum_{j'=1}^{M} x_{ij'}$, for all $i,j = 1,\ldots,M$.

Furthermore, the indicated values for which these minima are attained, are unique.

PROOF. Both proofs are an application of the method of the multiplicators of Lagrange (constraint extrema) (cf., e.g., [15]).

(a) Let

$$g(x_1,\ldots,x_M,\lambda) = \sum_{i=1}^{M} x_i^2 + \lambda\left(\sum_{i=1}^{M} x_i - 1\right). \tag{9}$$

Then, the sufficient condition for an extremum is

$$\frac{\partial g}{\partial x_{i_0}} = 2x_{i_0} + \lambda = 0,$$

and implies

$$x_{i_0} = -\frac{\lambda}{2}, \qquad \text{for all } i_0 = 1, \ldots, M.$$

Hence, all x_i must be equal. Since $\sum_{i=1}^{M} x_i = 1$, we then have that $x_i = \frac{1}{M}$, for all $i = 1, \ldots, M$. In this case, the minimal value is $\sum_{i=1}^{M} x_i^2 = \frac{1}{M}$.

(b) Let

$$h(x_{ij}\ (i,j = 1, \ldots, M), n) = \sum_{i,j=1}^{M} x_{ij}^2 + \eta \left(\sum_{i=1}^{M} \left(\sum_{j=1}^{M} x_{ij} \right)^2 - k \right). \tag{10}$$

For every $i_0, j_0 = 1, \ldots, M$, we now have

$$\frac{\partial h}{\partial x_{i_0 j_0}} = 2\, x_{i_0 j_0} + 2\eta \left(\sum_{j=1}^{M} x_{i_0 j} \right) = 0,$$

if

$$x_{i_0 j_0} = -\eta \sum_{j=1}^{M} x_{i_0 j}. \tag{11}$$

Now (11) implies that all $x_{i_0 j}$ $(j = 1, \ldots, M)$ are equal; hence, for all $i, j = 1, \ldots, M$,

$$x_{ij} = \frac{1}{M} \sum_{j'=1}^{M} x_{ij'}\,. \tag{12}$$

The indicated minima are unique since, in (a), the values $x_i = \frac{1}{M}$, $i = 1, \ldots, M$, correspond to the tangent point of the ball $\sum_{j=1}^{M} x_i^2 = r$ (minimal radius) and the hyperplane $\sum_{i=1}^{M} x_i = 1$; in (b), the values (12) correspond to the tangent point of the ball $\sum_{i,j=1}^{M} x_{ij}^2 = r$ (minimal radius) and the "hyperquadric" (8) (different form this ball). Hence, these minimal are unique. ∎

REMARK. From this lemma, it follows that $\bar{x}_{1/1} \geq \frac{n}{N^2}$, as well as $\bar{x}_2 \geq \frac{n}{N^2}$ (and, of course, $\bar{x}_{1/1}, \bar{x}_2 \leq n$). Note the value $\frac{n}{N^2}$ is the average number of documents *in the database* per 1/1-key as well as per 2-key, while $\bar{x}_{1/1}$, resp. $\bar{x}_2$, are the average numbers of documents *as retrieved by the catalographer* per 1/1-key, resp. per 2-key, (cf. also the important remark above).

This lemma has the following two theorems as consequences.

THEOREM 2. *If the p_i, $i = 1, \ldots, N$, are all equal (hence, equal to $\frac{1}{N}$), and if there is $i, j = 1, \ldots, N$, such that $p_{ij} \neq p_i\, p_j$, then*

$$\bar{x}_{1/1} < \bar{x}_2.$$

PROOF. The $(p_i)_{i=1,\ldots,N}$ represent the minimal situation discused in Lemma 1(a) while, since $p_{ij} \neq p_i\, p_j$ for a certain $i, j = 1, \ldots, N$, the $(p_{ij})_{i,j=1,\ldots,N}$ does not represent the minimal situation, we have that

$$\left(\sum_{i=1}^{N} p_i^2 \right)^2 < \sum_{i,j=1}^{N} p_{ij}^2. \tag{13}$$

Indeed the minimal value of $\left(\sum_{i=1}^{N} p_i^2 \right)^2$ equals $\frac{1}{N^2}$ and so does the minimal value of $\sum_{i,j} p_{ij}^2$. Hence, Equation (13) is true. ∎

THEOREM 3. *Let $(p_i)_{i=1,\ldots,N}$ be fixed and not all equal, and suppose also that there is a $i,j = 1,\ldots,N$ such that $p_{ij} \neq p_i p_j$. Let*

$$p_{ij} = \frac{p_i}{N}, \tag{14}$$

for every $i,j = 1,\ldots,N$. Then

$$\bar{x}_{1/1} > \bar{x}_2.$$

PROOF. We are given a situation of variable $(p_{ij})_{i,j=1,\ldots,N}$, such that $p_i = \sum_{j=1}^N p_{ij}$ are fixed for every $i = 1,\ldots,N$. Hence, Lemma 1(b) can be applied, since the fact that all p_i are fixed implies that

$$\sum_{i=1}^N \left(\sum_{j=1}^N p_{ij}\right)^2 = \sum_{i=1}^N p_i^2$$

is fixed. Equation (14) and Lemma 2(b) now imply that

$$\sum_{i,j=1}^N p_{ij}^2 < \sum_{i,j=1}^N p_i^2 p_j^2,$$

(since $(p_{ij})_{i,j=1,\ldots,N} \neq (p_i p_j)_{i,j=1,\ldots,N}$). Hence

$$\bar{x}_{1/1} > \bar{x}_2.$$

In fact, (although it is interesting in itself) we do not need Lemma 1(b) here; we can also apply Lemma 1(a): for p_{ij} as in (14), we have:

$$\sum_{i,j=1}^N p_{ij}^2 = \frac{1}{N}\sum_{i=1}^N p_i^2 < \sum_{j=1}^N p_j^2 \sum_{i=1}^N p_i^2,$$

by Lemma 1(a), using that $(p_1,\ldots,p_N) \neq (\frac{1}{N},\ldots,\frac{1}{N})$. ∎

In short, we proved so far:

CONCLUSION.

(i) $p_{ij} = p_i p_j$, for all $i,j \Rightarrow \bar{x}_{1/1} = \bar{x}_2$,

(ii) $p_{ij} \neq p_i p_j$, for an i,j and $p_i = \frac{1}{N}$ for all $i \Rightarrow \bar{x}_{1/1} < \bar{x}_2$, and

(iii) $p_{ij} \neq p_i p_j$, for an i,j, $p_i \neq \frac{1}{N}$ for an i, and $p_{ij} = \frac{p_i}{N}$ for all $i,j \Rightarrow \bar{x}_{1/1} > \bar{x}_2$.

We now will extend these results to more general search keys.

3. AVERAGE NUMBER OF DOCUMENTS RETRIEVED IN THE LIBRARY SYSTEM WITH MORE GENERAL SEARCH KEYS

Keys of the same "size" (i.e., using the same number of symbols) can now be studied, analogously with Section 2. Let we give the example of a 3-key versus a 1/1/1-key. In this case, we need the probabilities:

$$(p_{ijk})_{i,j,k=1,\ldots,N},$$

where p_{ijk} denotes the probability that a word (e.g., the first meaningful word of the title) starts with symbol i, immediately followed by symbol j and then immediately followed by symbol k. We now have the relations:

$$p_i = \sum_{j,k=1}^N p_{ijk}, \tag{15}$$

for every $i = 1,\ldots,N$, and, of course,

$$\sum_{i,j,k=1}^N p_{ijk} = 1. \tag{16}$$

On the average, in this system, we find

$$\bar{x}_{1/1/1} = n \sum_{i,j,k=1}^{N} p_i^2 p_j^2 p_k^2 = n \left(\sum_{i=1}^{N} p_i^2 \right)^3 \tag{17}$$

documents with an 1/1/1-key and

$$\bar{x}_3 = n \sum_{i,j,k=1}^{N} p_{ijk}^2 \tag{18}$$

documents with a 3-key. So, as in Section 2, we must determine the probabilities $(p_{ijk})_{i,j,k=1,\dots,N}$ and compare (17) with (18), using (15), to know whether we better use an 1/1/1-key than a 3-key, or vice-versa. Of course, the larger the length of the keys is, the more intricate the test, since the probabilities $(p_{ij\dots\ell})$ number a total of N^d, where d is the length of the key under study. As a general rule, we can say that the knowledge of $\bar{x}_{1/1}$ versus $\bar{x}_2$ gives sufficient information of how a library system behaves when more slashes (/) are added. By no means, in an automated system, it is difficult to obtain the matrix $(p_{ij})_{i,j=1,\dots,N}$, since N is (usually) between 26 and 30, or 35. We close this part by checking the analogues of Theorems 1, 2 and 3 of the previous section.

Theorem 1′. *If for all* $i, j, k = 1, \dots, N$,

$$p_{ijk} = p_i \, p_j \, p_k, \tag{19}$$

then $\bar{x}_{1/1/1} = \bar{x}_3$. *The converse is not true.*

Proof. $\bar{x}_{1/1/1} = \bar{x}_3$ by (17)–(19). The converse is not true: take $p_{ijk} \neq p_{ij} \, p_k$,where p_{ij} form the counterexample in Theorem 1. ∎

Theorem 2′. *If the* p_i *are all equal, and if there is* $i, j, k = 1, \dots, N$, *such that* $p_{ijk} \neq p_i \, p_j \, p_k$, *then*

$$\bar{x}_{1/1/1} < \bar{x}_3.$$

Proof. The proof is again based on Lemma 1(a). ∎

Theorem 3′. *Let* $(p_i)_{i=1,\dots,N}$ *be fixed and not all equal and suppose also that there is a* $i, j, k = 1, \dots, N$, *such that* $p_{ijk} \neq p_i \, p_j \, p_k$. *Let*

$$p_{ijk} = \frac{p_i}{N^2}, \tag{20}$$

for every $i, j, k = 1, \dots, N$. *Then*

$$\bar{x}_{1/1/1} > \bar{x}_3.$$

Proof. Again this can be proved using Lemma 1(b) or 1(a). ∎

In the same way other keys (of equal length) can be compared. Note also the following trivial relations between keys of unequal length:

$$\bar{x}_{1/1/1/1} < \bar{x}_{1/1/1} < \bar{x}_{1/1}, \tag{21}$$

$$\bar{x}_4 < \bar{x}_3 < \bar{x}_2, \tag{22}$$

and also relations such as:

$$\bar{x}_{3/1/1/1} < \bar{x}_{3/1/1},$$

$$\bar{x}_{4/4} < \bar{x}_{3/3} < \bar{x}_3,$$

and so on.

We continue our study on search key performance of a fixed library system, but now we investigate the effect of a change of the alphabet.

4. EFFECTS IN SEARCH KEY PERFORMANCE OF A CHANGE OF THE ALPHABET

After a certain time that a library system has been used, one might feel the need of enlarging the possibilities of forming search keys, i.e., of allowing more symbols for the formation of the search keys. This need is realistic in many senses and occurs often in practice. Indeed, consider a small (young) library in which one wants to use search keys as retrieval tool for catalogers. It seems logical to use—beyond the 26 letters of the alphabet— only a few other symbols. After several years, any search key, based on the same set of symbols, deteriorates in performance, since the size of the library increases. Sooner or later one must increase the size of the search key or (keeping the same search key) the size of the alphabet. Typical in this connection is the use of the symbols "e, è, é, ê, ..." For a small library, one might agree on only using the symbol "e." When the library grows, there might be a need to use all symbols "e, è, é, ê, ... " in the search keys. This is especially the case for a library that had almost no books written in French in the beginning but, when time passes, buys more and more of these books. A concrete example is offered by a library of a school or university, where one has added courses in French at a certain time. Another example is offered by the library system to which my library belongs (cf. Supra). At regular time intervals, the board or directors of the system re-evaluates the list of used symbols, based on occurrence tables of these symbols. If a certain symbol has been used increasingly during the past time one might consider splitting it up in a logical way, or if there is a request from a staff member of one of the participating universities to an all-new symbol (e.g., ñ, Å, ł, ...), one might add it to the list of usable symbols in the formation of the search key. This is in particular important when new languages are taught in the university (e.g., ñ in Spanish, ł in Polish, ø in Danish, and so on). It is, of course, clear that in this case, the average number of documents retrieved (with a certain key) will be lower, and hence, that we have a better performance (see also further on).

One can wonder, however, if a certain key that was optimal for the original alphabet, will still be optimal for the new alphabet. We will show that this is not always true: changing the alphabet might cause a change of the search key and hence of the cataloger's habits (in order to continue the optimal retrieval). We furthermore give criteria for which such a change is not necessary. We finally give some hints on which kind of alphabet-enlargements might give the best improvement of search key performance.

Let us fix the old symbol-probabilities as above, by $(p_i)_{i,\dots,N}$. Denote by $(q_{i'})_{i'=1,\dots,M}$, the new symbol-probabilities, where $M > N$. To express, that the new symbols are a refinement of the old ones, we say, that there exists a partition $\mathcal{P}$ of $\{1,\dots,M\}$ such that, for every $i \in \{1,\dots,N\}$, there exists exactly one set $C_i \in \mathcal{P}$ such that

$$p_i = \sum_{i' \in C_i} q_{i'}, \tag{23}$$

(i.e., the symbol with number i has been refined into the symbols with numbers $i' \in C_i$). This clearly models the practical habit of adding symbols.

EXAMPLES.

(1) In the old system, all symbols, è, ê, é, e, are read (and used) as e; in the new system we allow for all of them.

(2) In the old system, several symbols that do not occur very frequently (e.g., ~, $*$, $^{\circ}$, and so on) are 'attached' to one symbol (a so-called "wild card"), say $*$. In the new system, they all can be used.

(3) Symbols that are not refined are still included in the above formalism. In this case, the set C_i is a singleton!

Let us study the performance of 1/1-keys and 2-keys in both systems. In the new system, let us denote by $\bar{x}^*_{1/1}$, resp. $\bar{x}^*_2$, the average number of documents retrieved with a 1/1-key, respectively a 2-key. We clearly have (when at least one C_i is not a singleton) the following theorem.

THEOREM 4.

$$\bar{x}^*_{1/1} < \bar{x}_{1/1}, \tag{24a}$$

$$\bar{x}^*_2 < \bar{x}_2. \tag{24b}$$

PROOF. This is evident since (4) and (6) also imply

$$\bar{x}^*_{1/1} = n \left(\sum_{i'=1}^{M} q_{i'}^2 \right)^2, \tag{25}$$

and

$$\bar{x}^*_2 = n \sum_{i',j'=1}^{M} q_{i'j'}^2. \tag{26}$$

Hence,

$$\begin{aligned} \bar{x}^*_{1/1} &= n \left(\sum_{i=1}^{N} \sum_{i' \in C_i} q_{i'}^2 \right)^2 < n \left(\sum_{i=1}^{N} \left(\sum_{i' \in C_i} q_{i'} \right)^2 \right)^2 \\ &= n \left(\sum_{i=1}^{N} p_i^2 \right)^2 = \bar{x}_{1/1}, \end{aligned}$$

by Equation (23). An analogous argument applies for the proof of $\bar{x}^*_2 < \bar{x}_2$, now also using the next lemma. ∎

LEMMA 2. *For all* $i, j = 1, \dots, N$,

$$p_{ij} = \sum_{(i',j') \in C_i \times C_j} q_{i'j'}. \tag{27}$$

PROOF. Since the event "occurrence of the symbols with numbers i, j" in the old system is the disjoint union of the events "occurrence of the symbols with numbers i', j'" in the new system, the result follows immediately. ∎

There is no hope of proving that $\bar{x}_{1/1} < \bar{x}_2$ implies $\bar{x}^*_{1/1} < \bar{x}^*_2$, or analogous things.

EXAMPLES.

(1) We will construct an example for which $\bar{x}_{1/1} < \bar{x}_2$ (for the old alphabet) and such that $\bar{x}^*_{1/1} > \bar{x}^*_2$. In this, we will use Theorems 2 and 3 of Section 2.

Let $N = 2$ and $(p_{ij})_{i,j=1,2}$ satisfy the conditions of Theorem 2, e.g., $p_1 = p_2 = 1/2$, $p_{11} = 1/8$, $p_{12} = 3/8$, $p_{21} = 1/6$ and $p_{22} = 1/3$. Hence, $\bar{x}_{1/1} = 0.25\,n < 0.295\,n = \bar{x}_2$. Now the numbers $(q_{i'})_{i'=1,2,3,4}$, defined as $q_{1'} = p_{11}$, $q_{2'} = p_{12}$, $q_{3'} = p_{21}$ and $q_{4'} = p_{22}$ are not all equal. Furthermore, if we define

$$q_{i'j'} = \frac{q_{i'}}{4},$$

for all $i', j' = 1, 2, 3, 4$, we are perfectly within the conditions of Theorem 3, with $M = 4$ and $(q_{i'j'})_{i',j'=1,2,3,4}$. Hence, we are assured of the fact that

$$\bar{x}^*_{1/1} > \bar{x}^*_2.$$

(In fact, $\bar{x}^*_{1/1} \approx 0.08703\,n > 0.07378\,n \approx \bar{x}^*_2$.)

(2) We now construct an example for which $\bar{x}_{1/1} > \bar{x}_2$, but $\bar{x}^*_{1/1} < \bar{x}^*_2$. We let ourselves guide again by Theorems 2 and 3.

Let $N = 2$, $p_1 = 1/3$, $p_2 = 2/3$, $p_{11} = p_{12} = 1/6$ and $p_{21} = p_{22} = 1/3$. Hence, $\bar{x}_{1/1} > \bar{x}_2$ (according to Theorem 3). Indeed: $\bar{x}_{1/1} \approx 0.30864\,n > 0.27778\,n \approx \bar{x}_2$. Now $(q_{i'})_{i'=1,\dots,6}$ with all $q_{i'} = \frac{1}{6}$ is clearly a refinement of $(p_i)_{i=1,2}$ (symbols $1'$, $2'$ refine symbol 1 and symbols $3'$, $4'$, $5'$, $6'$ refine symbol 2). According to Theorem 2, if we make at least one $q_{i'j'} \neq q_{i'} q_{j'}$ then $\bar{x}^*_{1/1}$ must be smaller than $\bar{x}^*_2$. Take, e.g., $q_{1'1'} = 1/72$, $q_{1'2'} = 3/72$, and all other $q_{i'j'} = 1/36$ (i.e., 34 times). We then have $\bar{x}^*_{1/1} \approx 0.027778\,n < 0.02816\,n \approx \bar{x}^*_2$.

The above examples are a negative property of search keys: changing the symbols used in catalography (and this is always necessary, now and then) might change the search key that one can use in an optimal way. Besides the fact that one has to look for the best one each time, one has also the uncomfortable fact, that cataloger's habits change, which is not so advisable.

Starting with a system for which $\bar{x}_{1/1} < \bar{x}_2$ must give a certain advantage for $\bar{x}^*_{1/1} < \bar{x}^*_2$ to happen (and the same for both >). In fact, the next two theorems show that, usually, the orders are not reversed (we will comment on these results after the proofs).

THEOREM 5. *Let the system* $(p_{ij})_{i,j=1,\dots,N}$ *be such that* $\bar{x}_{1/1} < \bar{x}_2$. *Let* $(q_{i'j'})_{i',j'=1,\dots,M}$ *be a refinement of the* (p_{ij}), *such that for all* $i,j = 1,\dots,N$, *there exists one* $(i'(i), j'(j)) \in C_i \times C_j$ *for which*

$$p_{ij} = q_{i'(i),j'(j)} + \varepsilon_{ij}, \tag{28}$$

such that

$$\varepsilon_{ij} \le \frac{1}{2n}(\bar{x}_2 - \bar{x}_{1/1}), \tag{29}$$

for all i,j. *Then* $\bar{x}^*_{1/1} < \bar{x}^*_2$.

PROOF. By Theorem 4 and the fact that $\bar{x}_{1/1} < \bar{x}_2$, we have:

$$\left(\sum_{i'=1}^{M} q_{i'}^2\right)^2 < \left(\sum_{i=1}^{N} p_i^2\right)^2 < \sum_{i,j=1}^{N} p_{ij}^2. \tag{30}$$

Now, by Lemma 2,

$$\sum_{i,j=1}^{N} p_{ij}^2 = \sum_{i,j=1}^{N} \left(\sum_{(i',j')\in C_i\times C_j} q_{i'j'}\right)^2 = \sum_{i',j'=1}^{M} q_{i'j'}^2 + \sum_{i,j=1}^{N} \sum\nolimits^{*} q_{i'j'}\, q_{i''j''}\,, \tag{31}$$

where $\sum^*$ is over all $(i',j'), (i'',j'') \in C_i \times C_j$ such that $(i',j') \neq (i'',j'')$. The last term is calculated as follows:

$$\alpha = \sum_{i,j=1}^{N} \sum\nolimits^{*} q_{i'j'}\, q_{i''j''} = 2\sum_{i,j=1}^{N} \sum\nolimits^{**} (p_{ij} - \varepsilon_{ij})\, q_{i''j''} + \sum_{i,j=1}^{N} \sum\nolimits^{***} q_{i'j'}\, q_{i''j''}\,, \tag{32}$$

where $\sum^{**}$ is over all $(i'',j'') \in C_i \times C_j$ for which $(i'',j'') \neq (i'(i), j'(j))$, and $\sum^{***}$ is over all $(i',j'), (i'',j'') \in C_i \times C_j$ for which $(i',j') \neq (i'',j'')$, and none of the (i',j') or (i'',j'') are equal to $(i'(i), j'(j))$. But

$$\sum_{i,j=1}^{N} \sum\nolimits^{**} (p_{ij} - \varepsilon_{ij})\, q_{i''j''} = \sum_{i,j=1}^{N} (p_{ij} - \varepsilon_{ij}) \sum\nolimits^{**} q_{i''j''}\,.$$

By Lemma 2, this is equal to:

$$\sum_{i,j=1}^{N} (p_{ij} - \varepsilon_{ij})\, \varepsilon_{ij}\,.$$

Also,

$$\sum_{i,j=1}^{N} \sum\nolimits^{***} q_{i'j'}\, q_{i''j''} \le \sum_{i,j=1}^{N} \left(\sum\nolimits^{**} q_{i'j'}\right)\left(\sum\nolimits^{**} q_{i''j''}\right) = \sum_{i,j=1}^{N} \varepsilon_{ij}^2\,.$$

Hence, Equation (32) becomes

$$\alpha \le 2\sum_{i,j} p_{ij}\,\varepsilon_{ij} - 2\sum_{i,j=1}^{N} \varepsilon_{ij}^2 + \sum_{i,j=1}^{N} \varepsilon_{ij}^2 = 2\sum_{i,j=1}^{N} p_{ij}\,\varepsilon_{ij} - \sum_{i,j=1}^{N} \varepsilon_{ij}^2 < 2 \max_{i,j=1,\dots,N} \varepsilon_{ij}\,.$$

If we apply (29),we then have that

$$\alpha < \sum_{i,j=1}^{N} p_{ij}^2 - \left(\sum_{i=1}^{N} p_i^2\right)^2, \tag{33}$$

and so, by (30), (31) and (33),

$$\begin{aligned}\sum_{i',j'=1}^{M} q_{i'j'}^2 - \left(\sum_{i'}^{2} q_{i'}^2\right)^2 &= \sum_{i',j'=1}^{M} q_{i'j'}^2 + \alpha - \left(\sum_{i'=1}^{M} q_{i'}^2\right)^2 - \alpha \\ &> \sum_{i,j=1}^{N} p_{ij}^2 - \left(\sum_{i=1}^{N} p_i^2\right)^2 - \alpha > 0.\end{aligned}$$

Consequently,

$$\bar{x}^*_{1/1} < \bar{x}^*_2. \qquad ■$$

Note. It follows from the above proof that, instead of (29), it suffices to suppose the weaker condition

$$\sum_{i,j=1}^{N} p_{ij}\,\varepsilon_{ij} < \frac{1}{2n}\left(\bar{x}_2 - \bar{x}_{1/1}\right). \tag{34}$$

Analogous as Theorem 5, we have the following theorem.

Theorem 6. *Let the system $(p_{ij})_{i,j=1,\ldots,N}$ be such that $\bar{x}_{1/1} > \bar{x}_2$. Let $(q_{i'j'})_{i',j'=1,\ldots,M}$ be a refinement of the (p_{ij}), such that, for all $i = 1,\ldots,N$, there exists one $i'(i) \in C_i$ for which*

$$p_i = q_{i'(i)} + \varepsilon_i, \tag{35}$$

such that

$$\varepsilon_i \le \frac{1}{2n}\left(\bar{x}_{1/1} - \bar{x}_2\right), \tag{36}$$

for all i. Then $\bar{x}^*_{1/1} > \bar{x}^*_2$.

Proof. The proof is completely analogous to the one above. ■

Note. As in the above theorem, we can replace (36) by the weaker condition

$$\sum_{i=1}^{N} p_i\,\varepsilon_i \le \frac{1}{2n}\left(\bar{x}_{1/1} - \bar{x}_2\right). \tag{37}$$

Interpretation. Both theorems above have practical value. Indeed, they both deal with the case that (old) symbols are refined into several others, for which all but one have small probability. Examples 1 and 3, in the beginning of this section, are indeed also examples for which these conditions can be true. If an old symbol is a "mixture" of new symbols, for which the new probabilities are of the same magnitude (as is probably the case in Example 2), the order of performance of 1/1-keys versus 2-keys might be changed. This last case, however, is interesting in another way as explained now.

Another problem that one can study in relation to the change of the alphabet is the following: given a system $(p_{ij})_{i,j=1,\ldots,N}$, $M > N$, a natural number, and a partition $\mathcal{P} = \{C_i \mid\mid i = 1,\ldots,N\}$ of $\{1,\ldots,M\}$, what is the best way to refine (p_{ij}) into $(q_{i'j'})_{i',j'=1,\ldots,M}$, i.e., to minimise $\bar{x}^*_{1/1}$ or $\bar{x}^*_2$? The answer to this problem still brings us back to Example 2, as explained above. First, we need a lemma which is only a slight generalization of Lemma 1(b).

LEMMA 3. *Let $M > N$ be a fixed natural number and $\mathcal{P} = \{C_i \mid\mid i = 1, \dots, N\}$ a fixed partition of $\{1, \dots, M\}$.*

(a) *Let $(q_{i'j'})_{i',j'=1,\dots,M}$ be such that*

$$\sum_{i,j=1}^{N} \left(\sum_{(i',j') \in C_i \times C_j} q_{i'j'} \right)^2 = k$$

is fixed. Then, $\sum_{i',j'=1}^{M} q_{i'j'}^2$ is minimal for

$$q_{i'j'} = \frac{1}{\#(C_i \times C_j)} \sum_{(i'',j'') \in C_i \times C_j} q_{i''j''} = \frac{p_{ij}}{\#(C_i \times C_j)} ,$$

for every $(i', j') \in C_i \times C_j$ and every $i, j = 1, \dots, N$.

Here $\#$ denotes "the cardinality of the set" (i.e., the number of elements in the set).

(b) *Let $(q_{i'})_{i'=1,\dots,M}$ be such that*

$$\sum_{i,j=1}^{N} \left(\sum_{i' \in C_i} q_{i'} \right)^2 = k$$

is fixed. Then $\sum_{i'=1}^{M} q_{i'}^2$ is minimal for

$$q_{i'} = \frac{1}{\#C_i} \sum_{i'' \in C_i} q_{i''} = \frac{p_i}{\#C_i} ,$$

for every $i' \in C_i$ and every $i = 1, \dots, N$.

PROOF. The proof goes along the lines of the one of Lemma 1(b). ∎

COROLLARY 1. *Let the system $(p_{ij})_{i,j=1,\dots,N}$ be fixed, as well as M and $\mathcal{P}$, as in Lemma 3.*

(a) *Then, the refinement of this alphabet that minimises $\bar{x}_2^*$ must have probabilities equal to*

$$q_{i'j'} = \frac{p_{ij}}{\#(C_i \times C_j)}, \tag{38}$$

for every $(i', j') \in C_i \times C_j$ and every $i, j = 1, \dots, N$.

(b) *The refinement of this alphabet that minimises $\bar{x}_{1/1}^*$ must have probabilities equal to*

$$q_{i'} = \frac{p_i}{\#C_i}, \tag{39}$$

for every $i' \in c_i$ and every $i = 1, \dots, N$.

PROOF. This follows immediately from formulae (23), (25)–(27). ∎

INTERPRETATION. The above result says that a refinement of an alphabet is the most effective if the new probabilities are a more or less equal division of the corresponding old probabilities (cf. Example 2, in the beginning of this section). But in this case one might have to change the search key that one uses (cf. this section).

This concludes the study of the average number of documents found (using a certain search key), when the system of symbol-probabilities are given (i.e., when the library system is fixed). We will now try to prove further results about the average (of the above mentioned average number of documents) over all systems (i.e., with varying probabilities of symbol-occurrence).

5. PERFORMANCE OF THE 1/1-KEY VERSUS THE 2-KEY AVERAGED OVER ALL SYSTEMS

5.1. Definition of the Problem

Given a fixed system of probabilities $(p_{ij})_{i,j=1,\dots,N}$ as in Section 2, we were able to calculate the average numbers of documents retrieved by using a 1/1-key.

$$\bar{x}_{1/1} = n \left(\sum_{i=1}^{N} p_i^2\right)^2, \tag{40}$$

and by using a 2-key

$$\bar{x}_2 = n \sum_{i,j=1}^{N} p_{ij}^2. \tag{41}$$

We showed several results but, dependent on the vector $(p_{ij})_{i,j=1,\dots,N}$ we had the different relations: $\bar{x}_{1/1} = \bar{x}_2$, $\bar{x}_{1/1} < \bar{x}_2$ or $\bar{x}_{1/1} > \bar{x}_2$. We now wonder what more we can prove, when averaging these $\bar{x}$-values over all (p_{ij})'s that are possible.

Besides the theoretical interest in such a result, we obtain in this way "universal" knowledge about 1/1-keys versus 2-keys. This knowledge can always be used, when there is no knowledge available about the (p_{ij})-values, or for general purposes.

We will fix the notation and start with some lemmas that are needed in the sequel.

5.2. Notation and Preliminary Facts

Based on the formulae (40) and (41), we note that $\bar{x}_{1/1}$ depends on the vector $(p_1, \dots, p_N)$ while $\bar{x}_2$ depends on the vector $(p_{ij})_{i,j=1,\dots,N}$. We therefore define the following objects for the 1/1-key, respectively, the 2-key.

For the 1/1-key

Let

$$\Omega_1 = \left\{ (x_i)_{i=1,\dots,N} \in [0,1]^N \parallel \sum_{i=1}^{N} x_i = 1 \right\}, \tag{42}$$

and $X_{1/1}$ be the distribution function of the $\bar{x}_{1/1}$ (divided by n):

$$P_{\Omega_1}(x \le X_{1/1} \le x') = P\left(\left\{ (p_i)_{i=1,\dots,N} \parallel \left(\sum_{i=1}^{N} p_i^2\right)^2 \in [x, x'] \right\} \mid \Omega_1 \right), \tag{43}$$

the conditional probability w.r.t. Ω_1. Here P denotes the usual Lebesgue measure in $\mathbb{R}^N$. Let $f_{1/1}$ be the corresponding density function. Hence, the average that we want to study (the average of $\bar{x}_{1/1}$ over all probabilities $(p_1, \dots, p_N)$ in Ω_1): $\mu_{1/1}$ is (cf. (4)):

$$\mu_{1/1} = n \int_{\frac{1}{N^2}}^{1} x\, f_{1/1}(x)\, dx. \tag{44}$$

The integration interval comes from the fact that, for all $(p_1, \dots, p_N) \in \Omega_1$

$$\frac{1}{N^2} < \left(\sum_{i=1}^{N} p_i^2\right)^2 \le 1, \tag{45}$$

and these extrema are attained ($\frac{1}{N^2}$ for all p_i equal to $\frac{1}{N}$, and 1 if of all p_i are zero except one $p_i = 1$).

Of course, the big problem is the determination of $f_{1/1}$. This turns out to be non-trivial and will be solved in the sequel. In the same way, we define the same quantities for the 2-key.

For the 2-key

Let

$$\Omega_2 = \left\{ (x_{ij})_{i,j=1,\dots,N} \in [0,1]^{N^2} \parallel \sum_{i,j=1}^{N} x_{ij} = 1 \right\}, \tag{46}$$

and X_2 be the distribution function of the $\bar{x}_2$ (divided by n):

$$P_{\Omega_2}(x \le X_2 \le x') = P\left(\left\{ (p_{ij})_{i,j=1,\dots,N} \parallel \sum_{i,j=1}^{N} p_{ij}^2 \in [x,x'] \right\} \middle| \Omega_2 \right), \tag{47}$$

the conditional probability w.r.t. Ω_2. Here P denotes the usual Lebesgue measure in $\mathbb{R}^{N^2}$ (there cannot be any confusion between this P, and P above). Let f_2 be the corresponding density function. Hence, the average of $\bar{x}_2$ over all probabilities $(p_{ij})_{i,j=1,\dots,N}$ in Ω_2 is (cf. (41)):

$$\mu_2 = n \int_{\frac{1}{N^2}}^{1} x f_2(x)\, dx. \tag{48}$$

Note that we have here the same integration interval since

$$\frac{1}{N^2} \le \sum_{i,j=1}^{N} p_{ij}^2 \le 1 \tag{49}$$

and these extrema are attained as in the previous case. Again, we must be able to find f_2 and compare it with $f_{1/1}$. This will be done now.

5.3. Calculation of $f_{1/1}$ and f_2

We can now prove the following result:

THEOREM 7. *Let Ω_1 be as in (42). Let f be the density function of the distribution function Z, defined as:*

$$P_{\Omega_1}(x \le Z \le x') = P\left(\left\{ (x_i)_{i=1,\dots,N} \parallel \sum_{i=1}^{N} x_i^2 \in [x,x'] \right\} \mid \Omega_1 \right). \tag{50}$$

Then f is defined on the interval $[\frac{1}{N},1]$ and

$$f(x) = \frac{N-1}{2\left(1-\frac{1}{N}\right)^{(N-1)/2}} \left(x - \frac{1}{N}\right)^{(N-3)/2}, \tag{51}$$

for all $x \in [\frac{1}{N},1]$.

PROOF. That f is defined on the interval $[\frac{1}{N},1]$ follows directly from the arguments about formula (45). Now the set

$$\left\{ (x_i)_{i,=1,\dots,N} \parallel \sum_{i=1}^{N} x_i^2 \in [x,x'] \right\} \cap \Omega_1,$$

is the sector between the balls in Ω_1 (a $N-1$ dimensional space), being the intersection of the sector between the balls in $\mathbb{R}^N$:

$$\left\{ (x_i)_{i=1,\dots,N} \parallel \sum_{i=1}^{N} x_i^2 \in [x,x'] \right\}$$

and Ω_1. The volume of the former sector must be calculated in Ω_1, hence in $N-1$ dimensions, since we work conditionally w.r.t. Ω_1.

By definition of f, it hence suffices to calculate the derivative (w.r.t. x) of the volume of a ball in $\mathbb{R}^{N-1}$, being the intersection with Ω_1 of the ball in $\mathbb{R}^N$ with radius x and center 0.

The center of this ball in $\mathbb{R}^{N-1}$ is the point at shortest distance from 0 to the hyperplane Ω_1, hence the point $(\frac{1}{N}, \dots, \frac{1}{N}) \in \mathbb{R}^N$. The radius of the intersected ball in Ω_1 is equal to r, where

$$r^2 = \| \left(\frac{1}{N}, \dots, \frac{1}{N}\right) - X\|_N^2 , \tag{52}$$

where X is any vector $(x_i)_{i=1,\dots,N}$ of the intersected ball. Therefore, we have:

$$\sum_{i=1}^{N} x_i^2 = x, \tag{53}$$

and

$$\sum_{i=1}^{N} x_i = 1 \tag{54}$$

(cf. Figure 1, for $N = 3$). Hence,

$$r^2 = \sum_{i=1}^{N} \left(x_i - \frac{1}{N}\right)^2 = \sum_{i=1}^{N} x_i^2 - \frac{2}{N} \sum_{i=1}^{N} x_i + \sum_{i=1}^{N} \frac{1}{N^2} = x - \frac{1}{N}, \tag{55}$$

by (53) and (54).

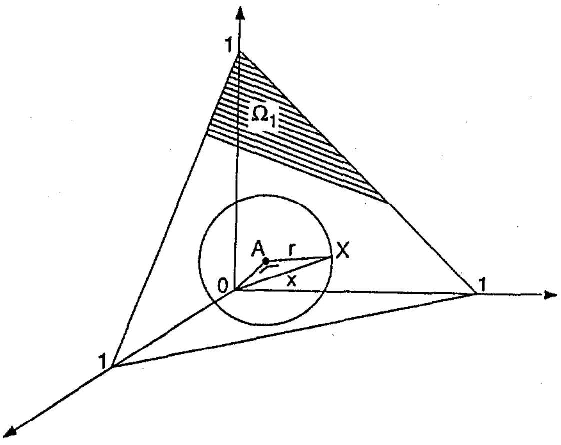

Figure 1. Illustration of the proof of Theorem 7, $A = \left(\frac{1}{N}, \dots, \frac{1}{N}\right) \in \Omega_1$, $X = (x_i)_{i=1,\dots,N} \in \Omega_1$, $N = 3$.

Now the volume $V(r)$ of a ball with radius r in $\mathbb{R}^{N-1}$ is given by (cf. [15]):

$$V(r) = \frac{\sqrt{\pi^{N-1}}}{\Gamma\left(\frac{N-1}{2} + 1\right)} r^{N-1} , \tag{56}$$

where Γ is the classical gamma function. Hence, by (55):

$$V(x) = \frac{\sqrt{\pi^{N-1}}}{\Gamma\left(\frac{N-1}{2} + 1\right)} \left(x - \frac{1}{N}\right)^{(N-1)/2} . \tag{57}$$

Now

$$C\, V(x) = \int_{\frac{1}{N}}^{x} f(x')\, dx',$$

for $C > 0$, such that

$$\int_{\frac{1}{N}}^{1} f(x')\, dx' = 1.$$

Hence, we have the requirement

$$C \frac{\sqrt{\pi^{N-1}}}{\Gamma\left(\frac{N-1}{2}+1\right)} \left(1 - \frac{1}{N}\right)^{(N-1)/2} = 1,$$

so

$$C = \frac{\Gamma\left(\frac{N-1}{2}+1\right)}{\sqrt{\pi^{N-1}}} \frac{1}{\left(1-\frac{1}{N}\right)^{(N-1)/2}}. \tag{58}$$

Hence,

$$C\, V(x) = \frac{1}{\left(1-\frac{1}{N}\right)^{(N-1)/2}} \left(x - \frac{1}{N}\right)^{(N-1)/2}. \tag{59}$$

The derivative (w.r.t. x) is the function f:

$$f(x) = \frac{N-1}{2\left(1-\frac{1}{N}\right)^{(N-1)/2}} \left(x - \frac{1}{N}\right)^{(N-3)/2},$$

for all $x \in [\frac{1}{N}, 1]$. ∎

COROLLARY 2.

$$f_2(x) = \frac{N^2-1}{2\left(1-\frac{1}{N^2}\right)^{(N^2-1)/2}} \left(x - \frac{1}{N^2}\right)^{(N^2-3)/2}, \tag{60}$$

for all $x \in [\frac{1}{N^2}, 1]$.

PROOF. Interpreting Theorem 7 above for N^2 instead of N, we hence have Ω_2 in (46), instead of Ω_1. The definition of f_2 then implies (60) (based on (51)). ∎

For $f_{1/1}$, we cannot immediately apply Theorem 7, due to the special form of (43). In exactly the same way as in Theorem 7, we can however prove the following theorem.

THEOREM 8. *Let* Ω_1 *be as in (42). Let* g *be the density function of the distribution function* U*, defined as:*

$$P_{\Omega_1}(x \leq U \leq x') = P\left(\left\{ (x_i)_{i=1,\ldots,N} \parallel \left(\sum_{i-1}^{N} x_i^2\right)^2 \in [x, x'] \right\} \mid \Omega_1\right). \tag{61}$$

Then, g *is defined on the interval* $[\frac{1}{N^2}, 1]$*, and*

$$g(x) = \frac{N-1}{4\left(1-\frac{1}{N}\right)^{(N-1)/2}} \left(\sqrt{x} - \frac{1}{N}\right)^{(N-3)/2} \frac{1}{\sqrt{x}}, \tag{62}$$

for all $x \in [\frac{1}{N^2}, 1]$.

PROOF. The proof follows the lines of the proof of Theorem 4, now using that $\sum_{i=1}^{N} x_i^2 \in [\sqrt{x}, \sqrt{x'}]$, and taking appropriate derivations. ∎

COROLLARY 3.

$$f_{1/1}(x) = \frac{N-1}{4\left(1-\frac{1}{N}\right)^{(N-1)/2}} \left(\sqrt{x} - \frac{1}{N}\right)^{(N-3)/2} \frac{1}{\sqrt{x}}, \tag{63}$$

for all $x \in [\frac{1}{N^2}, 1]$.

PROOF. This follows immediately from Theorem 5 and (43). ∎

5.4. Calculation of $\mu_{1/1}$ and μ_2

We are now in a position to calculate $\mu_{1/1}$ and μ_2, using formulae (44) and (48).

Theorem 9.

$$\mu_{1/1} = n\left[1 - \frac{4\,(N^2+N+2)\,(N-1)}{(N+1)(N+3)N^2}\right], \tag{64}$$

and

$$\mu_2 = n\left[1 - \frac{2\,(N^2-1)}{N^2\,(N^2+1)}\right]. \tag{65}$$

Furthermore:

$$\mu_{1/1} < \mu_2. \tag{66}$$

Proof. It only requires to calculate

$$\mu_{1/1} = \int_{\frac{1}{N^2}}^{1} \frac{N-1}{4\,(1-\frac{1}{N})^{(N-1)/2}} \left(\sqrt{x} - \frac{1}{N}\right)^{(N-3)/2} \sqrt{x}\,dx, \tag{67}$$

and

$$\mu_2 = \int_{\frac{1}{N^2}}^{1} \frac{N^2-1}{2\,(1-\frac{1}{N^2})^{(N^2-1)/2}} \left(x - \frac{1}{N^2}\right)^{(N^2-3)/2} x\,dx. \tag{68}$$

These integrals are evaluated via partial integration and yield, after simplification, the formulae (64) and (65).

It is now easy to prove that, for all $N \geq 2$, $\mu_{1/1} < \mu_2$. ∎

Important remark. The arguments given here are correct in the supposition that all vectors $(p_{ij})_{i,j=1,\dots,N}$ are equally possible (this was expressed by using the classical Lebesque measure of $\mathbb{R}^{N^2}$ and of $\mathbb{R}^N$ in the arguments—cf. Subsections 5.2 and 5.3). This is certainly not true in practice, but an acceptable supposition for a "first try" (certainly within these rather complex formulae). So, in our arguments, the density functions are increasing with x, being the volumes of balls with radii as in formula (55). This also explains, that $\lim_{N\to\infty} \mu_{1/1} = \lim_{N\to\infty} \mu_2 = n$; within our supposition, this is true: for very high N, most of the density is concentrated in the values of x close to 1. So, in such a system, most of the books have the same keys (the ones for $x < 1$ are not occurring any more) and, hence, the averages are very high. This is clearly a mathematically correct result, but without any practical consequences.

In practice, the density functions must increase from $r = 0$ up to the "average" values of $\sum_{i=1}^N p_i^2$ (resp., $\sum_{i,j=1}^N p_{ij}^2$), (these vectors have, in fact, the largest probability to occur in an automated library system!), and then, must decrease for the higher values of $\sum_{i=1}^N p_i^2$ (resp., $\sum_{i,j=1}^N p_{ij}^2$), up to 1. Indeed, these higher values correspond to the cases where only one or a few letters (symbols) are in use, which is never occurring.

So, a refinement of the above argument can be given by changing the integration interval $[\frac{1}{N^2}, 1]$ into $[\frac{1}{N^2}, \alpha_N]$, where $\alpha_N < 1$. This yields more realistic values of $\mu_{1/1}$ and μ_2, denoted as $\mu_{1/1}^*$ and μ_2^*. We have the following (easily calculated) result.

Theorem 10. *$\mu_{1/1}^*$ and μ_2^*, as described above, have the values:*

$$\mu_{1/1}^* = n\,\frac{\left(\sqrt{\alpha_N} - \frac{1}{N}\right)^{(N-1)/2}}{\left(1-\frac{1}{N}\right)^{(N-1)/2}} \left[\alpha_N - \frac{4}{N+1}\left(\sqrt{\alpha_N}\left(\sqrt{\alpha_N} - \frac{1}{N}\right) - \frac{2}{N+3}\left(\sqrt{\alpha_N} - \frac{1}{N}\right)^2\right)\right], \tag{69}$$

and

$$\mu_2^* = n \frac{\left(\sqrt{\alpha_{N^2}} - \frac{1}{N^2}\right)^{(N^2-1)/2}}{\left(1 - \frac{1}{N^2}\right)^{(N^2-1)/2}} \left[\alpha_{N^2} - \frac{2}{N^2+1}\left(\alpha_{N^2} - \frac{1}{N^2}\right)\right]. \tag{70}$$

Approximately, for high N (as in ordinary alphabets), this yields:

$$\mu_{1/1}^* \approx n \exp\left[\frac{1}{2}\left(1 - \frac{1}{\sqrt{\alpha_N}}\right)\right] \alpha_N{}^{(N+3)/4}, \tag{71}$$

$$\mu_2^* \approx n \exp\left[\frac{1}{2}\left(1 - \frac{1}{\sqrt{\alpha_{N^2}}}\right)\right] \alpha_{N^2}{}^{(N^2+2)/2}. \tag{72}$$

Furthermore, if $\alpha_N < 1$ and α_N is independent of N, $\mu_{1/1}^*$ and μ_2^* are decreasing with N and, for values of α_N not too close to 1, one has:

$$\mu_2^* < \mu_{1/1}^*. \tag{73}$$

We leave open the problem of finding an exact form of the distribution of the $(p_{ij})_{i,j=1,\dots,N}$ and of the $(p_i)_{i=1,\dots,N}$ (and, consequently, of the $\sum_{i,j=1}^N p_{ij}^2$ and $\left(\sum_{i=1}^N p_i^2\right)^2$).

6. RELATION WITH CONCENTRATION THEORY AND APPLICATIONS

Let f be an increasing density function (as $f_{1/1}$ and f_2 in the previous section), on the interval $[a, b]$. Then, we can consider the Lorenz-curve of this density: it is formed by the graph of the function (for $y \in [0, 1]$):

$$\mathcal{L}(f)(y) = (b-a) \int_0^y f(y'(b-a) + a)\, dy'. \tag{74}$$

For functions f as described, $\mathcal{L}(f)$ looks like in Figure 2: an increasing curve below the first bissectrice, and for which $\mathcal{L}(f)(0) = 0$, $\mathcal{L}(f)(1) = 1$. We refer to [12,13] for some basic notions on concentration theory.

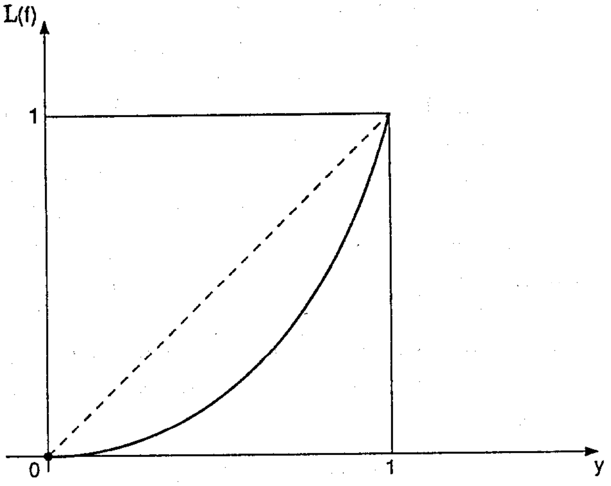

Figure 2. Graph of $\mathcal{L}(f)$.

We can show the following relation between the mean μ and curve $\mathcal{L}(f)$.

LEMMA 4. *Let f be an increasing density function on the interval $[a, b]$. Then, the mean μ can be written as*

$$\mu = b - (b-a)L, \tag{75}$$

where L is the area under the Lorenz-curve $\mathcal{L}(f)$ of f.

PROOF. Partial integration yields

$$\mu = \int_a^b x f(x)\,dx = \int_a^b x\,d(\varphi(x)),$$

where

$$\varphi(x) = \int_a^x f(x')\,dx'. \tag{76}$$

Hence

$$\mu = [x\,\varphi(x)]_a^b - \int_a^b \varphi(x)\,dx = b - \int_a^b \varphi(x)\,dx; \tag{77}$$

but

$$\begin{aligned}\int_a^b \varphi(x)\,dx &= \int_a^b dx \int_a^x f(x')\,dx'\\ &= (b-a)\int_a^b dx \int_a^y f(y'(b-a)+a)\,dy',\end{aligned}$$

where $x' = y'(b-a)+a$ and $x = y(b-a)+a$. Hence,

$$\begin{aligned}\int_a^b \varphi(x)\,dx &= (b-a)^2 \int_0^1 dy \int_0^y f(y'(b-a)+a)\,dy'\\ &= (b-a)\int_0^1 \mathcal{L}(f)(y)\,dy = (b-a)\,L.\end{aligned}$$

So,

$$\mu = b - (b-a)\,L. \qquad \blacksquare$$

COROLLARY 4. *For $N \geq 4$, $\mu_{1/1}$ and μ_2 can be expressed as:*

$$\mu_{1/1} = 1 - \left(1 - \frac{1}{N^2}\right) L_{1/1}, \tag{78}$$

and

$$\mu_2 = 1 - \left(1 - \frac{1}{N^2}\right) L_2, \tag{79}$$

where $L_{1/1}$ and L_2 represent the area under the Lorenz-curves $\mathcal{L}(f_{1/1})$, $\mathcal{L}(f_2)$ of $f_{1/1}$ and f_2, respectively.

PROOF. This follows readily from Lemma 4 and Corollaries 2 and 3 (see formulae (60) and (63)): $f_{1/1}$ and f_2 are increasing in the interval $[\frac{1}{N^2}, 1]$. ∎

NOTE. The result of Theorem 9 can now be explained intuitively, using (78) and (79). Since f_2 (formulae (60)) represents a much more concentrated situation than $f_{1/1}$ (formulae (63)), we have, that

$$\mathcal{L}(f_{1/1}) > \mathcal{L}(f_2),$$

and, hence,

$$L_{1/1} > L_2.$$

By (76) and (79), we now see that

$$\mu_{1/1} < \mu_2,$$

the result of Theorem 9. It is not clear how to interpret the result of Theorem 10 in this framework.

7. EXTENSION TO LARGER SEARCH KEYS

Let us consider, for example, a 1/1/1-key versus a 3-key. An analogous argument as the one leading to Corollary 2 yields (based on the preliminary results in Section 2)

$$f_3(x) = \frac{N^3 - 1}{2\left(1 - \frac{1}{N^3}\right)^{(N^3-1)/2}} \left(x - \frac{1}{N^3}\right)^{(N^3-3)/2}, \tag{80}$$

for $x \in [\frac{1}{N^3}, 1]$. For $f_{1/1/1}$, one finds, based on Section 2, Theorem 8 and Corollary 3, (analogous argument):

$$f_{1/1/1}(x) = \frac{N - 1}{6\left(1 - \frac{1}{N}\right)^{(N-1)/2}} \left(\sqrt[3]{x} - \frac{1}{N}\right)^{(N-3)/2} \frac{1}{\sqrt[3]{x^2}}. \tag{81}$$

The analogous results as in Theorem 9 and Theorem 10 can also be proved here:

$$\mu_{1/1/1} < \mu_3, \tag{82}$$

and

$$\mu_3^* < \mu_{1/1/1}^*. \tag{83}$$

We also have, that ($\sim$ is in the O-sense of Landau, cf. [16]):

$$f_{1/1}(x) \sim x^{(N-4)/4}, \quad \text{on } \left[\frac{1}{N^2}, 1\right], \tag{84a}$$

$$f_{1/1/1}(x) \sim x^{(N-7)/6}, \quad \text{on } \left[\frac{1}{N^3}, 1\right], \tag{84b}$$

$$f_2(x) \sim x^{(N^2-3)/2}, \quad \text{on } \left[\frac{1}{N^2}, 1\right], \tag{84c}$$

$$f_3(x) \sim x^{(N^3-3)/2}, \quad \text{on } \left[\frac{1}{N^3}, 1\right]. \tag{84d}$$

So, by (75),

$$\begin{aligned} \mu_3 - \mu_{1/1/1} &= \left(1 - \frac{1}{N^3}\right) \int_0^1 (\mathcal{L}(f_{1/1/1})(y) - \mathcal{L}(f_3)(y))\, dy \\ &> \left(1 - \frac{1}{N^2}\right) \int_0^1 (\mathcal{L}(f_{1/1})(y) - \mathcal{L}(f_2)(y))\, dy, \end{aligned}$$

by the above approximate formulae (84). We followed here the same intuitive argument as the one in the note following Corollary 3. Hence, using (75) again:

$$\mu_3 - \mu_{1/1/1} > \mu_2 - \mu_{1/1}. \tag{85}$$

Note that this result is only intuitive and, furthermore, based on equal probability of the vectors $(p_{ijk})_{i,j,k=1,\ldots,N}$ and $(p_i)_{i=1,\ldots,N}$.

Obviously, we can note the following trivial relations: $\mu_3 < \mu_2$ and $\mu_{1/1/1} < \mu_{1/1}$.

References

1. M.J. Coe, Uniqueness of compression codes for bibliographic retrieval, *Bull. Med. Libr. Assoc.* **58**, 587–597 (1970).
2. G.P. Guthrie and S.D. Slifko, Analysis of search key retrieval on a large bibliographic file, *J. Libr. Automation* **5**, 196–200 (1972).

3. F.G. Kilgour, P.L. Long and E.B. Leiderman, Retrieval of bibliographic entries from a name-title catalog by use of truncated search keys, *Proc. Amer. Soc. Inf. Sci.* **7**, 79–82 (1970).
4. F.G. Kilgour, P.L. Long, E.B. Leiderman and A.L. Landgraf, Title-only entries retrieved by use of truncated search keys, *J. Libr. Automation* **4**, 207–310 (1971).
5. B. Kjell, Performance of Kilgour's truncation algorithm in files of different subjects, *J. Amer. Soc. Inf. Sci.* **25**, 70–71 (1974).
6. A.L. Landgraf and F.G. Kilgour, Catalog records retrieved by personal author using derived search keys, *J. Libr. Automation* **6**, 103–108 (1973).
7. A.L. Landgraf, K.B. Rastogi and P.L. Long, Corporate author entry records retrieved by use of derived truncated search keys, *J. Libr. Automation* **6**, 156–161 (1973).
8. P.L. Long and F.G. Kilgour, A truncated search key title index, *J. Libr. Automation* **5**, 17–20 (1972).
9. T.C. Lowe, Direct access memory retrieval using truncated record names, *Software Age* **1**, 28–33 (1967).
10. T.C. Lowe, Effectiveness of retrieval key abbreviation schemes, *J. Amer. Soc. Inf. Sci.* **22**, 374–381 (1971).
11. J.D. Smith and J.E. Rush, The relationship between author names and author entries in a large on-line union catalog as retrieved using truncated keys, *J. Amer. Soc. Inf. Sci.* **28**, 115–120 (1977).
12. L. Egghe and R. Rousseau, Elements of concentration theory, In *Informetrics 89/90. Proc. 2nd Internat. Conf. on Bibliometrics, Scientometrics and Informetrics,* (Edited by L. Egghe and R. Rousseau), pp. 97–137, Elsevier, London, Canada, (1990).
13. L. Egghe and R. Rousseau, Transfer principles and a classification of concentration measures, *J. Amer. Soc. Inf. Sci.* **42** (7), 479–489 (1991).
14. E. Decoutere, Toepassing van search-keys bij het catalogiseren, Licentiaatsthesis, Universitaire Instelling Antwerpen, (1991).
15. R. Courant and F. John, *Introduction to Calculus and Analysis, Vol. 2*, Wiley, (1974).
16. C. Pisot and M. Zamansky, *Mathématiques Générales*, Dunod, (1966).