

Clumping Properties of Content-Bearing Words*

A. Bookstein

Center for Information
and Language Studies
University of Chicago
Chicago, IL 60637
a-bookstein@uchicago.edu

S. T. Klein

Dept. of Math. & CS
Bar Ilan University
Ramat-Gan 52900
Israel
tomi@cs.biu.ac.il

T. Raita

Comp. Sci. Dept.
University of Turku
20520 Turku
Finland
raita@cs.utu.fi

Abstract

Information Retrieval Systems identify content bearing words, and possibly also assign weights, as part of the process of formulating requests. For optimal retrieval efficiency, it is desirable that this be done automatically. This paper defines the notion of serial-clustering of words in text, and explores the value of such clustering as an indicator of a word's bearing content. This approach is flexible in the sense that it is sensitive to context: a term may be assessed as content-bearing within one collection, but not another. Our approach, being numerical, may also be of value in assigning weights to terms in requests. Experimental support is obtained from natural text databases in three different languages.

1. Introduction and Background

Automatic Information Retrieval (IR) has in the past been based on global word-counts — the only indicators previously available for assessing the content-bearing strength of words. But the advent of full text databases has created new possibilities. For example, we now have not only counts of how many times a word occurs in a document, or over a database, but also information about the sequential occurrence pattern of each word. In this paper, we consider how this sequential structure of text can be exploited. In particular, we shall use such information to generate lists of content-bearing words, which can be used as index terms in a formal request, and to develop retrieval weights that can be assigned to each index term.

The distinction between content-bearing and non-content-bearing words is central in IR. Much of traditional information retrieval focuses on the problem of vocabulary control [12, 16]. Although this emphasis is reduced in automated systems, nonetheless the distinction between

*Two of the authors (A.B. and S.T.K.) wish to acknowledge that the material in this paper is based upon research supported by the U.S. National Science Foundation under award number IRI-9307895, and by grant No. 92-00163 from the United States - Israel Binational Science Foundation (BSF), Jerusalem, Israel. T.R. acknowledges support by the Academy of Finland under grant No. 18587.

words that can be used for retrieval purposes and those that cannot remains, for example in the concept of a stop-list [14], words that are presumably lacking in subject content.

But whether or not a word conveys topical information may be collection dependent. For example, the word `computer` may distinguish books on Computer Science from other books in a general collection, but may not be useful for retrieval purposes within a collection restricted to material in Computer Science. For this reason, we would like methods that automatically identify words which are useful in requests, and that can be applied to specific collections — perhaps even subcollections dynamically identified during a preliminary stage of an extended retrieval process. Such methods will be particularly useful if they not only identify, but also quantify a term's index-merit; such a quantification would allow us to assign weights to terms in a vector-based IR system; for such purposes, even an approximate method could be useful, since a slightly incorrect assignment does not lose a good index term.

The problem of identifying content-bearing terms is reasonably well understood theoretically, using a probabilistic framework. As noted by Bookstein and Swanson [7], retrieval decisions have associated expected costs. Thus, if we know the distributional properties of a word in documents relevant and non-relevant to a term, and can assign values to costs, we can compute the expected cost of using and not using that term. The relative expected costs then serve as a conceptual criterion for deciding whether the word should be kept for retrieval purposes.

While such an analysis provides a conceptual criterion for identifying index terms, it is very difficult to apply. A variety of practical filters for content-bearing terms have been devised. Luhn [13], for example, suggested a very simple filter based solely on frequency of occurrence. But the value of a word as a subject indicator is likely to depend on its *pattern* of occurrence as well as its frequency. Bookstein and Swanson [7] suggested that non-content-bearing terms would be Poisson distributed from document to document of approximately the same size, and that content-bearing terms may be identified by departures from this distribution. For this reason, discrepancy from a random (or Poisson) distribution was seen as indicating that a word carried content. And to carry out a test of this discrepancy, we need know only how many occurrences there are of the term in each of a set of documents. This method, while quite effective, was costly computationally. More sensitive maximum likelihood estimation would further increase these costs. But also, it is disturbing to have to depend upon detailed distributional assumptions.

The availability of full text in machine readable form suggests new options. For now, we not only have data on the number of occurrences of a term in each unit, but also topological information of the pattern of occurrences in a *sequence* of text units. This extends the essential idea in [7], that if a word conveys meaning in a semantic sense, then its occurrences will be associated with textual units related to that meaning. For such a word, occurrences will not be random, but rather will tend to occur in clumps, just as content does; this tendency should produce sequential patterns that can be used to distinguish units about that word from units not about the word. It is precisely this nonrandom pattern of occurrence of certain words, in which a word's occurrences distinguish segments of text in a semantically useful way, that makes information retrieval possible.

Thus, while we are not able to detect meaning directly, we are able to exploit statistical

correlates of meaning. In particular, as we scan the text from beginning to end, we can test whether the passages containing a given term tend to be randomly scattered throughout the text. We now suggest that a pronounced tendency of passages containing a term to clump together might be a simple, non-parametric variant of the Poisson criterion. For a collection or sub-collection, the retrieval process might involve first identifying content-bearing terms using a clumping test, and afterwards using standard retrieval procedures based on these terms, as noted above.

Since the key concept of *clumping* is similar to that of *clustering* that is well known in Information Retrieval, it is important to emphasize the difference between these concepts. Traditionally, clustering methods are applied to sets of objects having no internal structure; isolated, but similar, items are gathered together into clusters on the basis of shared features. Typically, in IR the objects brought together are documents, based upon the similarity of their term vectors [17], or conversely, terms, based upon the similarity of their document vectors [1].

In contrast, we are here concerned with passages appearing in sequence, and the *adjacency* property of the passages is critical: we are asking whether passages containing a term appear surprisingly close together. To distinguish the two types of clustering, we shall often refer to the type of clustering being studied here as *serial-clustering* or *clumping*.

Another example where the distinction between traditional and serial-clustering occurs is in the area of bitmap compression [1, 2], where each bitmap is associated with a term, and collectively the bitmaps represent the concordance of a document set. In [1], compression took place by clustering terms that had similar occurrence patterns in a manner reminiscent of traditional clustering. On the other hand, in [4], Markov theory is used to model occurrences of individual bits within a map, parallel to the approach taken here.

The idea of serial-clustering has potential application to other problems in IR. IR research is increasingly focusing on full text, in which document structure such as contiguity information as well as term occurrence can be exploited. As a simplified example of what we have in mind, consider a retrieval strategy in which a document is retrieved if more than a threshold fraction of its paragraphs would be independently retrieved on the basis of simple word occurrence. But it is reasonable to increase the retrieval strength of a paragraph if adjacent paragraphs contain key terms. Knowing the tendency of a term to cluster serially could be useful here. Thus the retrieval strength of a unit might be increased by the retrieval strength of nearby units, modified by a measure of its terms' tendency to cluster serially.

2. Measures of Serial-Clustering

The assumption underlying this paper is that occurrences of a term sensitive to content will have a greater tendency to *clump*, or occur in the same textual neighborhood, than those of non-content-bearing terms. But, as is well known, it is not easy to determine by intuitive means whether occurrences of an object genuinely tend to cluster. A person will often see clusters forming even among objects that are randomly distributed. Thus, formal tests of clumping are needed. In this section we shall introduce several candidate measures of serial-

clustering strength, and derive tests of statistical significance for these measures.

The notion of clumping is intuitively simple, but can manifest itself in many ways, depending on how the clumps form. Our measures are intended to capture a clumping effect in a manner that is relatively independent of how the terms were generated. Nonetheless some measures may be more sensitive to specific modes of clumping than others. Below we describe four measures of serial-clustering. These will depend on two consequences of an underlying tendency of terms to cluster serially.

Condensation-Clustering: Consider text as being generated serially, a term at a time. We expect that occurrences of a given content-bearing word will tend to occur in clumps, reflecting the content of a textual neighborhood. If we now break the text up into segments for analysis, an artifact of this tendency of content-bearing words to clump is that some segments will contain more of these terms than expected by chance. This should be true for any detailed underlying serial-clustering mechanism. For convenience, we refer to this consequence of serial-clustering as *Condensation-clustering*.

More generally, we consider a set of entities distributed over a set of receptacles (above, words over a set of *textual units* such as paragraphs or pages). We test whether some receptacles are more effective than others in attracting occurrences of these entities. For example, if the entities are terms distributed over textual units, we expect that non-content-bearing terms will appear to be distributed at random over the textual units, and look for deviations from such a distribution as an indicator that a term bears content. In [7], this approach is taken with whole documents as the textual unit.

Serial-clustering is also reflected by the linear ordering of units containing the term in question, hence the second group of tests.

Linear-Clustering: We now consider objects (e.g. textual units, such as paragraphs in a document) arranged linearly. Some (e.g., units containing a specified term) are “marked” as designated objects, and intermixed with occurrences of unmarked objects. We ask whether the arrangement of the designated objects along the line can be described as random, or whether their occurrences tend to occur in clumps. Thus, whereas condensation-clustering treats the textual units as receptacles and analyzes the *number* of occurrences of a term in the textual units, linear-clustering examines the linear arrangement of textual units marked as containing at least one occurrence, and analyzes the *pattern* of occurrence of marked units. We shall also call this effect clumping, since it is just a higher level expression of the underlying clumping of the individual terms.

The measures discussed below are designed to be evaluated on textual units taken from a single document. However, they can be generalized to treat collections of documents, either by simply merging the documents into a single document for the analysis, or by carrying the analysis out separately on each document and combining the results to get a summary serial-clustering measure for a term.

2.1 Condensation Measure C1 — Term condensation over textual units

If terms are distributed at random over a number of textual units, several may land in a single unit. Thus, we expect the number of units containing at least one occurrence of the term to be less than the total number of occurrences of the term. But if the terms tend to cluster, in the *condensation* sense, then we expect even fewer units will contain a term. A simple measure of condensation-clustering is the ratio of the *actual* number of units that have at least one occurrence of the term to the *expected* number of units having at least a single occurrence, assuming a random distribution. The statistical significance of this measure can be derived as follows:

Suppose that our document is divided into D textual units. (We are currently developing effective methods for accomplishing this division. For the moment, any arbitrary segmentation method — for example, taking successive groups of five sentences, or accepting existing paragraphs — is a possibility.) Consider then the distribution of the occurrences of a term over these units. For that term, we know:

N : the number of units containing the term; and

T : the total number of occurrences of the term.

This information allows a test of the independence assumption. Focusing on the designated term, there are D^T distinguishable ways to place its T occurrences into the D units. Each such distribution is equally likely, given our assumption of independent term placement. To find the probability that exactly N units contain at least one occurrence of the term ($1 \leq N \leq \min(T, D)$), we must compute how many ways this can be done. We first note that there are $\binom{D}{N}$ different ways to select the units which are occupied. For each such set of N units, a placement of terms in these units defines both a partition of the T terms into N classes (of which there are $\left\{ \begin{smallmatrix} T \\ N \end{smallmatrix} \right\}$) and an ordering of the components of the partition (of which there are $N!$); the term in braces denotes a Stirling number of the second kind [11]. Thus there are

$$N! \binom{D}{N} \left\{ \begin{smallmatrix} T \\ N \end{smallmatrix} \right\}$$

distinguishable distributions in which exactly N units contain occurrences of the term, and the probability, $p(N, T)$, of exactly N units containing the term is given by

$$p(N, T) = \frac{N! \binom{D}{N} \left\{ \begin{smallmatrix} T \\ N \end{smallmatrix} \right\}}{D^T}. \quad (1)$$

Further discussion of this equation appears as Appendix I. That $\{p(N, T)\}_{N=1}^{\min\{D, T\}}$ is a probability distribution follows from the well known, often defining, property of Stirling numbers of the second kind [11, formula 1.2.6-41]:

$$x^r = \sum_{\ell} \left\{ \begin{smallmatrix} r \\ \ell \end{smallmatrix} \right\} x^{\ell} \quad \text{with} \quad x^{\ell} \equiv x \cdot (x-1) \cdots (x-\ell+1). \quad (2)$$

Using the basic recursion equation of $\left\{ \begin{smallmatrix} T \\ N \end{smallmatrix} \right\}$:

$$\left\{ \begin{smallmatrix} T \\ N \end{smallmatrix} \right\} = N \left\{ \begin{smallmatrix} T-1 \\ N \end{smallmatrix} \right\} + \left\{ \begin{smallmatrix} T-1 \\ N-1 \end{smallmatrix} \right\}, \quad (3)$$

we get for the expected number of textual units containing the term:

$$\begin{aligned} E_{C1} &\equiv \sum_n n p(n, T) \\ &= \frac{1}{D^T} \sum_n n D^n \left\{ \begin{smallmatrix} T \\ n \end{smallmatrix} \right\} \\ &= \frac{1}{D^T} \sum_n D^n \left(\left\{ \begin{smallmatrix} T+1 \\ n \end{smallmatrix} \right\} - \left\{ \begin{smallmatrix} T \\ n-1 \end{smallmatrix} \right\} \right). \end{aligned}$$

By equation (2), $\sum D^n \left\{ \begin{smallmatrix} T+1 \\ n \end{smallmatrix} \right\} = D^{T+1}$, and similarly $\sum D^n \left\{ \begin{smallmatrix} T \\ n-1 \end{smallmatrix} \right\} = D \sum (D-1)^{n-1} \left\{ \begin{smallmatrix} T \\ n-1 \end{smallmatrix} \right\} = D(D-1)^T$. Combining these, we conclude that

$$E_{C1} = D \left[1 - \left(1 - \frac{1}{D} \right)^T \right]. \quad (4)$$

A more intuitive derivation can be based on the indicator random variables, \tilde{x}_i , where:

$$\tilde{x}_i = \begin{cases} 1 & \text{if } i\text{-th document contains the term} \\ 0 & \text{otherwise.} \end{cases}$$

In terms of the $\{\tilde{x}_i\}$, the number of documents containing the term is given by $\sum \tilde{x}_i$, with expected value $E_{C1} = \sum E(\tilde{x}_i) = DE(\tilde{x}_i)$, since each indicator variable has the same expected value. But, since the indicator variables are binary, their expected values are just the probabilities of an indicator variable taking the value one. Since the probability that *no* term hit a document is given by $(1 - 1/D)^T$, the probability that $\tilde{x}_i = 1$ is given by $1 - (1 - 1/D)^T$, which confirms equation (4).

Given the value E_{C1} of the expected number of occupied documents, then if N units actually contain the term, $M_{C1} \equiv N/E_{C1}$ is a measure of condensation strength. If the terms do tend to concentrate within a subset of textual units, then we expect that fewer units will contain the term than predicted by the independence model — that is, that the observed N will be small compared to the value expected on the basis of the independence assumption.

As a test of statistical significance, we can compute the probability $P(N, T)$ that N or fewer units contain the term, given the independence assumption. This value should be unreasonably small for most content-bearing terms. This observation forms the basis of Test C1 of the hypothesis that the terms are randomly distributed. This probability is given by:

$$P(N, T) = \sum_{n=1}^N p(n, T).$$

In this manner we can assign to each term both a measure of condensation strength and the statistical significance of that value: the probability that it will occur in no more units than

it actually does occur in. These quantities measure the tendency of a *single term* to condense into a few textual-units.

The test described above evaluates the statistical significance of the measure for a single word. But, since many words are being studied, we expect some to take statistically significant values just by chance, even if genuine clumping did not occur. Before using such a measure, it is prudent to first assess whether the *set* of terms exhibits a true effect.

Such a global test can be based on the observation that, if there are S different terms, approximately $E_{P_0} = SP_0$ terms will have a given probability-value $P(N, T)$ of P_0 or less.¹ Thus we can divide the probabilities into classes, and compare the values $S(P_k - P_{k-1})$ with the number of terms whose probability-values actually fall between P_k and P_{k-1} . If there is a substantial tendency to clump, then the data should substantially deviate from this; in particular, for intervals with small values of P_k , we expect that the *observed* number of terms falling in the interval will be substantially larger than the number *expected*.

The tests described in this section require us to compute the probabilities, $\{p(N, T)\}$ and $\{P(N, T)\}$. Unfortunately, to compute the required probabilities directly, we need to compute Stirling numbers, for which no closed forms exist. We can, however, use equation (1) to compute a table of the values of $p(N, T)$. The recursion for the Stirling numbers yields the following recursion for $p(N, T)$:

$$p(N, T) = \frac{N}{D} p(N, T - 1) + \left(1 - \frac{N - 1}{D}\right) p(N - 1, T - 1),$$

and for the cumulative probabilities:

$$P(N, T) = \frac{N}{D} P(N, T - 1) + \left(1 - \frac{N}{D}\right) P(N - 1, T - 1).$$

To derive the recursion for $p(N, T)$ recall equation (3): $\{T\}_N: \{T\}_N = N \{T-1\}_N + \{T-1\}_{N-1}$. Proceed by multiplying both sides by $N! \binom{D}{N} / D^T$, expanding the binomial coefficient, and rearranging terms. Alternatively, a combinatoric argument could be used: the probability of N documents holding occurrences of the term is equal to the probability that the first $T - 1$ occurrences go into N documents multiplied by the probability that the last term go into one of those documents, plus the probability that the first $T - 1$ occurrences go into $N - 1$ documents multiplied by the probability that the last term go into a different document.

Similarly, the recursion for the cumulative probability follows from adding both sides of the recursion for $p(N, T)$:

$$\sum_{n=1}^N p(n, T) = \frac{1}{D} \sum_{n=1}^N n p(n, T - 1) + \sum_{n=1}^N p(n - 1, T - 1) - \frac{1}{D} \sum_{n=1}^N (n - 1) p(n - 1, T - 1).$$

The second term is just $P(N - 1, T - 1)$ since $p(0, T - 1) = 0$. Similarly, after changing the summation index in the last sum, we find it cancels all but the final term in the first sum, leaving $N p(N, T - 1)$. Substituting $p(N, T - 1) = P(N, T - 1) - P(N - 1, T - 1)$ and collecting terms produces the recursion we seek.

¹Because of the discreteness of the data, this result is only approximately true. The argument here is somewhat subtle and deferred to Appendix II.

2.2 Condensation Measure C2 — Term distribution over textual units

A second measure of condensation-clustering is based on the detailed distribution of terms over textual units. Continuing with the above example, the probability that m occurrences of a term appear in any *given* unit is given by the binomial distribution: $p_m = \binom{T}{m} (1/D)^m (1 - 1/D)^{T-m}$. Thus we expect Dp_m units to contain m occurrences. This can be measured and tested using the standard chi-square goodness of fit test.

2.3 Linear Measure L1 — Number of clumps

A measure and test of linear-clustering can be based on the runs test. If a document has D textual units, of which N have occurrences of the term being tested, then there are $\binom{D}{N}$ ways of choosing the N units that will contain the term from the total of D units; these are equi-probable, assuming there is no serial-clustering. We next compute the number of ways in which the selection can be made so that K clumps of marked units form; here a clump is defined as an unbroken run of one or more consecutive units with an occurrence of the term.

Each actual placement of tagged and untagged textual units is of the form

$$\square \quad \alpha_1 \quad \square \quad \alpha_2 \cdots \square \quad \alpha_{D-N} \quad \square \quad ,$$

where \square denotes a sequence of zero or more units with the designated term, and each α denotes one unit *not* containing the designated term; there are $D - N$ of these. That is, a box is a *potential* location of a clump of textual units with the term.

To get a placement in which there are exactly K clumps (non-empty boxes) of units with the designated term, we must first choose the K boxes which are to contain the clumps; since there is one more box than there are α 's, we have $D - N + 1$ boxes, and the choice can be made in $\binom{D-N+1}{K}$ ways. We then allocate the N selected units among the K boxes in a manner in which each box has at least one unit; this can be done in $\binom{N-1}{K-1}$ ways [11]. Combining these terms, we conclude that the probability of K clumps of the designated term is given by:

$$p_K = \frac{\binom{N-1}{K-1} \binom{D-N+1}{K}}{\binom{D}{N}},$$

where K can take values between 1 and $\min(N, D - N + 1)$. That $\{p_K\}$ is a probability distribution follows from [11, formula 1.2.6-22].

The expected number of clumps, E_{L1} , can be derived as follows:

$$\begin{aligned} E_{L1} &\equiv \sum_k k p_k = \sum_k k \frac{\binom{N-1}{k-1} \binom{D-N+1}{k}}{\binom{D}{N}} = \frac{D - N + 1}{\binom{D}{N}} \sum_k \binom{N-1}{k-1} \binom{D-N}{k-1} \\ &= (D - N + 1) \frac{\binom{D-1}{N-1}}{\binom{D}{N}} = (D - N + 1) \frac{N}{D}. \end{aligned}$$

Thus we conclude that

$$E_{L1} = N \left(1 - \frac{N-1}{D} \right). \quad (5)$$

As with measure C1, a simple argument, using indicator variables, is possible if only expected values are required. We now define an indicator variable that takes the value one when the i -th document *begins* a clump. The argument proceeds much as before, except we must distinguish the first document from the rest. Thus $E_{L1} = \Pr\{\text{first document begins a clump}\} + (D-1)\Pr\{i\text{-th document begins a clump} \mid i \neq 1\}$. The first quantity is simply the probability that the first document contain a term, and equals N/D . However, if $i \neq 1$, a clump begins at document i if the i -th document contains an occurrence of the term, which happens with probability N/D , and, *given* that the i -th document has an occurrence of the term, that the preceding document (document number $i-1$) does not, which happens with probability $1 - (N-1)/(D-1)$. Thus the expected number of clumps is given by

$$\frac{N}{D} + (D-1) \left[\frac{N}{D} \left(1 - \frac{N-1}{D-1} \right) \right] = N \left(1 - \frac{N-1}{D} \right),$$

as was claimed above.

If K clumps are formed, we can use the formula for E_{L1} to evaluate $M_{L1} \equiv K/E_{L1}$ as a measure of linear-clustering. Since a tendency to cluster linearly would reduce the number of clumps, we are seeking terms for which this ratio is substantially less than one.

We can construct a test of statistical significance for this measure, in parallel to what we did for Test C1. Suppose a term appears in N units. The probability $P_K(N)$ that these N units will form K or fewer clumps, assuming the independence model, is given by

$$P_K(N) = \sum_{k=1}^K p_k.$$

Computing this probability for the actual value of clumps formed yields the significance level for the term. A global test can also be constructed, as was done for Test C1.

2.4 Linear Measure L2 — Gap length between Marked Units

An approximate direct measure of linear-clustering is the *variability* of gaps between marked units. This measure, and its test of significance, can be constructed from the following probabilistic argument.

If the N marked units are distributed at random over the D units of the document, then the probability that a randomly chosen unit not be marked is $\gamma = 1 - N/D$. Thus the probability of a run of r ‘blanks’ between two successive marked units is given approximately by the geometric distribution: $p_r = \gamma^r(1 - \gamma)$.

If there is pronounced clumping, the *variability* of lengths of runs of blanks will be greater than expected on the assumption of independence. But the variance of the geometric distribution is $\sigma^2 \equiv \gamma/(1 - \gamma)^2$. Thus the ratio (empirical variance of run-lengths)/ σ^2 could be used as a measure for the degree of linear-clustering.

2.5 Linear Measure L3 — Markov Model

The tests analyzed above were non-parametric, in that they did not assume any detailed distributional model. In [5] we developed Markov models of bitmap generation as part of an attempt to compress database concordances. We here note that estimation of the parameters of the Markov model implicitly constitutes a test of serial-clustering. In developing the model, we think of ourselves as sequentially moving from one textual unit to the next, and changing state according to whether or not we meet with the designated term in the current unit. We assume the state changes can be described as a Markov chain. If the model is valid, studying the parameters of the model gives us a very sensitive test of the clumping proclivity of terms. Thus the parameters of the Markov model constitute measures of linear-clustering. Furthermore, as indicated in [5], we can also develop tests of the model itself; thus we can assess whether the model is valid, and if so, which terms it selects as candidates for index terms on the basis of their tendency to clump.

3. Experiments

We next test whether our basic hypothesis, that words that tend to cohere spatially also tend to bear content, is valid. We first evaluate our methods on the Hebrew Bible, which consists of 309631 words, partitioned into 929 chapters; there are 41197 different words. To select candidates for “good” index terms, we used measures M_{C1} and M_{L1} . In these tests, we treated each chapter as a textual unit. Subsequent tests on the English Bible and on a French language database will be described below.

3.1 Preliminary Global Test

Before presenting our experimental results, we carried out a global test to ascertain that the overall distribution of words is not consistent with randomness. Table 1 summarizes the results of global Test C1, restricting ourselves to the 3083 terms that occur often — that is, appear in at least 10 textual units. For each word, we can compute the probability that its C1 value would be taken by a term that didn’t clump. In this way, each term is given a probability value. The first column defines a set of probability intervals $(P_{i-1}, P_i]$. The second column is the expected number of terms whose probability values fall into a given probability range, if the independence model is assumed. In order to emphasize the degree to which our data deviates from the expected values beyond what can be expected from random fluctuations, we also ran the following simulation. For each term, we took its number of occurrences T , and randomly assigned each occurrence to one of the D units; this determined the number N_s of units containing the term in the simulation. Then the actual number N_a and the simulated number N_s of recipient units were converted to probability-values. The third and fourth columns show the number of terms taking probability values in the given intervals for the simulated and actual results respectively. For technical reasons, 12 out of the 3083 terms were not included in Table 1: the table size needed to calculate the corresponding $P(N, T)$ values exceeded the capacity of our machine. Note that while the simulated results

nicely agree with the expected ones, the actual data show a very strong bias towards the class of very low probabilities, with 73% of the terms having probability-values smaller than 0.01. Thus we can say with confidence that a substantial number of terms occurring “often” do exhibit a genuine tendency to cluster linearly.

Probability Interval	Expected number	Simulation result	Actual number
0.00 – 0.01	12.86	14	2239
0.01 – 0.05	47.92	45	115
0.05 – 0.10	92.85	114	237
0.10 – 0.25	252.89	276	145
0.25 – 0.50	352.91	368	28
0.50 – 0.75	281.65	246	2
0.75 – 1.00	2026.88	2008	305

Table 1: *Comparative table for global Test C1*

3.2 Experimental procedure

We next evaluated the effectiveness of our criteria for isolating words that are meaningful and thus candidates for use in retrieval requests. First, the list of the 3083 terms appearing in at least 10 chapters was presented to a Bible-expert, who was asked to classify the terms into those he judged as content-bearing and those not. This was done preceding, and independently of, the subsequent analysis, and without knowledge of its results.

Several peculiarities of the Hebrew language made this task difficult. For example, Hebrew grammar may generate thousands of variants of most words, since prepositions and articles may be prefixed, and possessive pronouns may be suffixed. Therefore, single words may often represent entire phrases. On the other hand, many terms may be parsed in more than one way and most have several homonyms. Thus for deciding about whether a term carried content, all its interpretations must be taken into account. Nevertheless, a global judgement is still possible, based on the expert’s knowledge (or rather assessment) of the likelihood of the various possible interpretations of a term. In our case, only 512 out of the 3083 terms were marked as content-bearing.

We then produced two lists of these 3083 terms: the first list was ordered by increasing condensation values M_{C1} , and the second by increasing linear-clustering values M_{L1} . The values of M_{C1} varied between 0.060 and 1.030, and for M_{L1} between 0.276 and 1.039, where a value of 1.000 would be expected of a non-clumping term. To quantify the effectiveness of the clumping-based measures, classical IR measures were adapted as follows: the clumping-based selection process is analogous to a document retrieval process, except that good words, rather than good documents, are being retrieved. To identify the relevant words, we relied on the

judgement of an expert, defining the set of *relevant* items as the words that he marked as content-bearing. At any given threshold for the clumping measure, we consider the set of *retrieved* items to be the words in the vocabulary list with values of the clumping measure above that threshold.

Continuing with this analogy, for each threshold for the clumping measure, we assessed the effectiveness of our word retrieval process by the following values:

Precision — number of relevant words in the retrieved set divided by the size of the set; and

Recall — number of relevant words in the retrieved set divided by total number of words selected as relevant by our judge.

Agreeing with the standard use of these measures in IR, simultaneous Recall and Precision values of 1 would indicate a perfect match between the expert’s partition and that generated by one of the clumping measures.

3.2.1 Results for Hebrew

For each cutoff point on the list, denote by S the set of retrieved items. Recall and Precision vary with the selection of the cutoff point. Raising it reduces the size of the set S , so that the Recall will probably get smaller; on the other hand, if the ordering criterion is effective, the reduced set will tend to have a higher density of relevant items, and thus a higher Precision. By letting the threshold value vary, we get a series of sets S , and thereby can create a Recall-Precision curve.

The curves displayed in Figure 1 were obtained as follows. After ordering the terms by increasing values of M_{C1} , each position is in turn considered as a possible threshold, and the Recall and Precision of the set of terms above that position are evaluated. Finally, Precision is plotted as a function of Recall. This was then repeated for M_{L1} . If the terms had been ordered randomly, we would have expected to find the same proportion of relevant items in any subset, so that the Recall-Precision curve would be the solid horizontal line shown in the figure, at Precision level = (number of relevant terms) / (size of the full list of terms), or $512/3083 = 0.166$ in our example. Both curves generated by our measures clearly differ from this line, showing high Precision values at low Recall, and decreasing slowly as Recall increases. To allow a further comparison, a third curve is included, based upon terms ordered by their *inverse document frequency* (*idf*). We include this curve as an example of a measure based upon global statistics often used in IR applications [14]. Of course, we are using much more information to generate our curves; the graphs suggest that the additional information contained in the internal structure of documents does indeed have retrieval value.

An interesting example, that illustrates the limitations of our methodology, is the word **dena**, which ranked high for both measures, yet was assessed as non-content-bearing by the expert. The word **dena** is a Hebrew word meaning **she judged**; our evaluator apparently didn’t consider this word as content bearing because of its frequent occurrence in the Bible. Though we may disagree with this evaluation, we were surprised to find it is ranked tenth

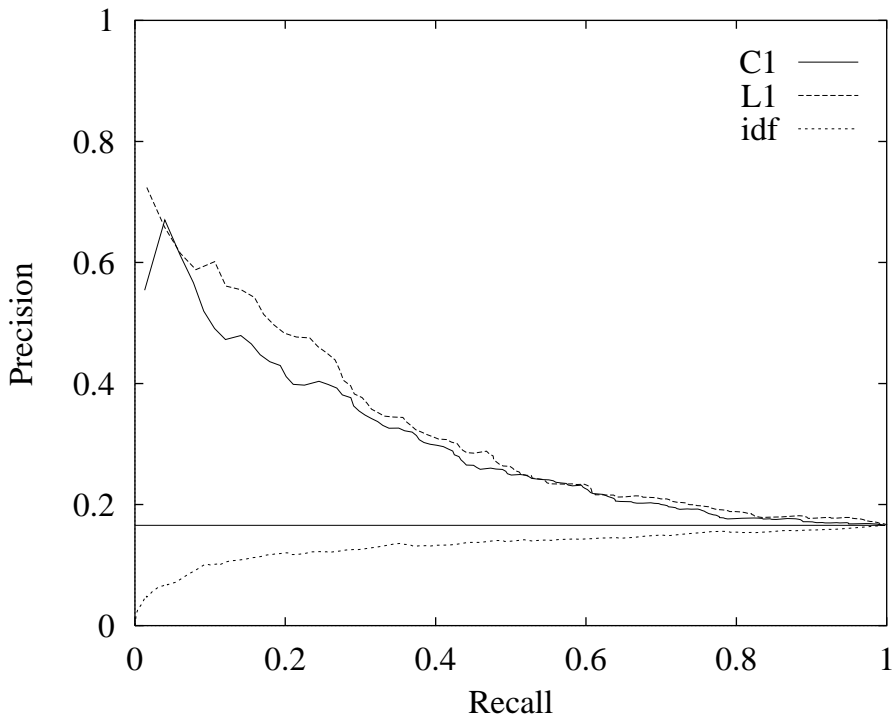


Figure 1: *Recall-Precision curves for Hebrew Bible*

by M_{C1} ($M_{C1} = 0.210$) and first by M_{L1} ($M_{L1} = 0.276$). The reason is that **dena** is also an Aramaic word (meaning **this**), which explains its strong clumping tendency: **dena** appears overall 49 times: 48 of these appearances have the Aramaic meaning and are concentrated in the few chapters in the books of Daniel and Ezra that are not in Hebrew; only one occurrence carries the Hebrew meaning. Thus, although **this** is not content-bearing, it is the *language* characteristic that accounts for the clustering.

We wondered why other such Aramaic words weren't identified in this way. It turned out that other Aramaic words were either not frequent enough to be included in the test set, or had the same spelling (since vowels are omitted) as a non-content bearing Hebrew word, which dominated the statistical measure.

The fact that our first test database was in Hebrew might have had a negative impact on our results, due to the peculiarities of this language. To compensate for this, we repeated the experiment with the King James Version of the English Bible.

3.2.2 Results for English

The King James Bible contains 10644 different terms. Requiring again $N \geq 10$, we got a set of 2472 terms, 254 of which were marked as content-bearing by the expert. Figure 2 again depicts the Recall-Precision curves, which show an improvement over the curves in Figure 1. For the English Bible, even *idf* performed slightly better than random choice, though still evidently worse than the two clumping measures.

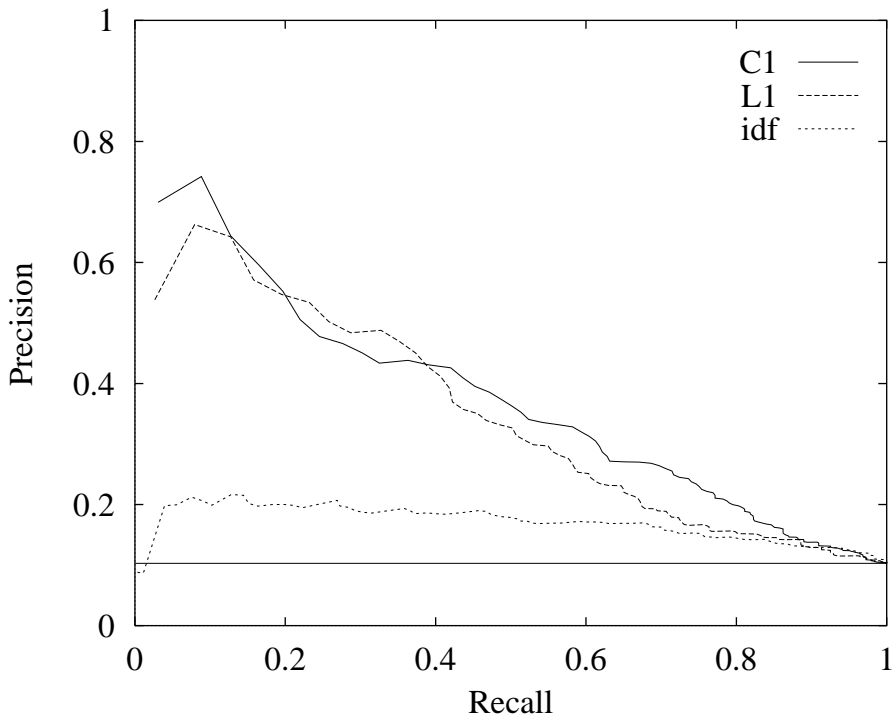


Figure 2: *Recall-Precision curves for English Bible*

To gain some insight into the nature of our measures, we next give a few examples of the rankings of certain terms. When we order terms by increasing values of M_{C1} and M_{L1} , the following four terms were among the first 30 elements for both measures: **Elisha**, **Ahab**, **sockets**, **Saul**. Clearly, since the Bible follows chronological order, there are many consecutive segments mentioning Elisha or Saul, but these occur almost only in the chapters reporting about their eras; this is what produces the strong clumping tendency.

It is interesting to note that we found both **her** and **she** among the first hundred elements for M_{C1} , whereas **his** and **he** are much further down the list. The reason is that the overwhelming majority of personalities reported on by the Bible are male. The feminine pronouns are thus used only in those rare contexts dealing with famous women. This is a good example of terms which would have been considered as stop-words in a general database, but which, for the specific text at hand, clearly bear content and might actually be good keywords for queries about women in the Bible. But this example does remind us that while clumping is *associated* with content, the association is statistical, and we must be careful not to confuse this with semantic identity. In IR, when we speak of meaning, we often in fact intend to indicate content that distinguishes some passages from others. Both **her** and **she** serve this function in the Bible.

An informative example, which shows the difference in character between condensation and linear-clustering, is the term **Psalms**: this term is the second element ($M_{L1} = 0.251$) on the linear-clustering list, but is the next to last ($M_{C1} = 1.021$) for condensation-clustering. Indeed, **Psalms** appears 85 times, scattered over 83 chapters, so it is not clustered at all from the condensation point of view. However, these 85 chapters are all in the Book of Psalms, which

consists of 150 consecutive chapters. Thus these occurrences form a strong linear-cluster. Similarly in the Hebrew Bible, the word used for Psalm appears 57 times, in 57 different chapters of the Book of Psalms; it appears as the sixth element in the M_{L1} -list, and as the last element in the M_{C1} -list.

3.2.3 Results for French

The previous experiments were based on segments of a single document, wherein we expect the terms to cluster serially. As a consequence of our earlier work on concordances [4, 5], we were led to test as well the tendency of terms to cluster serially over a set of *independent* documents, ordered as they would be in a realistic database. We expected the strength of linear-clustering would now be diminished. However, because of the tendency of documents to be ordered temporally, and by author, it is plausible that some linear-clustering would occur.

We thus extended the experiment to a large, multi-document, full-text database: the *Trésor de la Langue Française* (TLF). This database consists of about 680 megabytes of French language material, made up of 39757 complete documents including novels, short stories, poetry and essays, by a variety of authors. The bulk of the texts are from the 17th through 20th centuries. The number of different terms is 439201. We treated each document in the TLF as a textual unit. We chose $N \geq 1400$ as a requirement on the number of documents a term must occur in to be considered. There were 3595 terms in this range, of which only 171 were marked as content bearing by an expert in French literature. Figure 3 displays the Recall-Precision results for TLF.

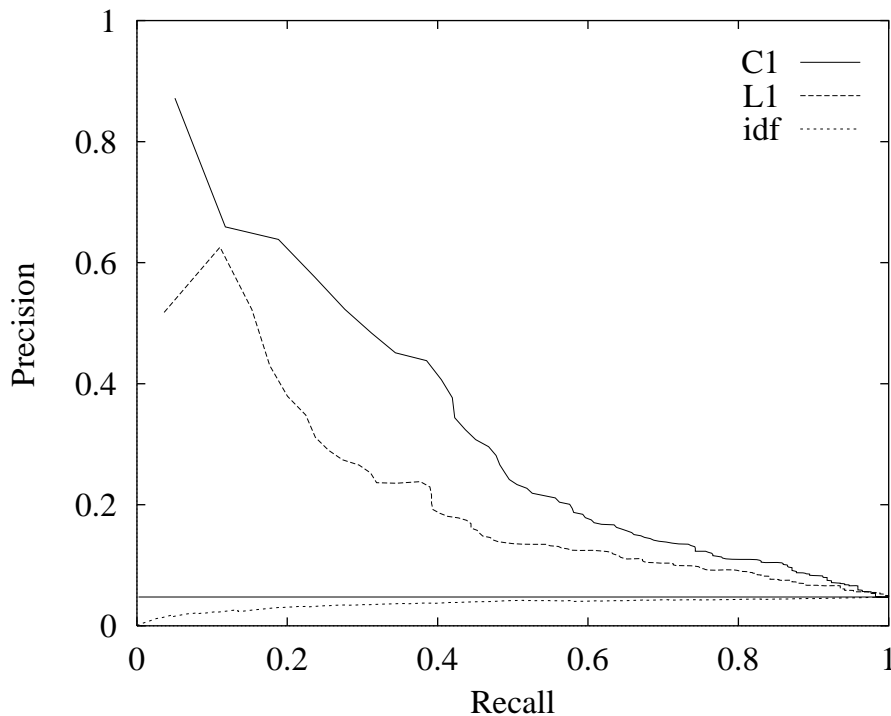


Figure 3: Recall-Precision curves for TLF

We note again the same pattern as for our tests on the Hebrew and English Bible. Since the number of relevant terms is very small, the expected precision on the assumption of a random distribution is only about 0.05 and *idf* again gives an almost horizontal line at this level. But the curves for both measures show similar Precision values for low Recall than did the Bible files, so the improvement over *idf* is even more evident. As typical examples, some of the strongest terms for both measures were archaic forms of words that appear moderately often, like *luy*, *estoit*, *avois*, *tousjours*, etc. These terms would possibly be considered as stop-words, devoid of content; their modern counterparts, at least of the above examples, belong to the 100 most frequent terms. But in their old form, they are excellent key words for locating texts written in the 17th century or earlier. Again, they carry content in the sense of distinguishing some passages from others. An example similar to the Psalm case for the Bibles is the term *chapitre*: this is the first element ($M_{L1} = 0.187$) on the linear-clustering list, but ranked 2817-th ($M_{C1} = 0.653$) for condensation-clustering.

It is interesting to compare the results just described with those of an earlier experiment. In [3] we performed a parallel analysis, but used the probabilities of measures as a criterion for goodness, rather than the measure itself. The intention there was to give us some protection against statistical fluctuations of the measure. However, we noted a number of words that did not appear to bear content, yet ranked high by our probability measure.

The reason was that most words do show a small tendency to clump, but not one large enough to gain prominence on a probability scale. However, the words we referred to above had a very high frequency of occurrence. For such words even small tendencies can be distinguished from chance. Thus, if a word occurs very frequently, even a slight tendency to clump might be associated with a very low probability, and therefore appear to be desirable by the probability measure.

It is now possible to see whether using the measure directly treats these terms differently. In TLF, there are 691 terms that appear in more than 5000 documents; of these, all but 21 ranked high (i.e., above the threshold $p = 0.005$ chosen in [3]) in at least one of the probability based measures, and 460 ranked high in both. Only 14 of these 691 terms have been marked as content bearing by the expert. On the other hand, using the measures M_{C1} and M_{L1} , these stopwords are not concentrated on the top of the lists, and many of them (1e, 1a, un, et, etc.) appear at the very bottom. This comparison emphasizes how dangerous it is to use very tempting probability measures as an indicator of quality.

3.3 Relation between linear and traditional clustering

We emphasized in the introduction the distinction between the type of clustering we study in this paper and traditional clustering. These approaches to clustering are complementary: in [1] it is the *terms* themselves that form the clusters, while in this paper we examine the linear-clustering, or clumping, of the *occurrences* of a single term. However, since the association between items is evaluated in traditional clustering by means of item co-occurrence, we expect that the occurrence pattern of items strongly tied together in a single cluster should in fact have a strong clumping tendency. For if two terms occurred purely at random, it is highly unlikely that they would have an association strong enough to tie them together. But terms

that do appear in a single cluster would be expected not only to clump, but to clump in similar ways. We now have an opportunity to test this hypothesis.

In [1], a hierarchical variant of traditional clustering was applied to terms in the Hebrew and English Bibles and in TLF. Some of the clusters obtained that way are presented in [2]. In the following experiment, we took the terms belonging to a given cluster and evaluated their ranks in the lists ordered by the values produced by M_{C1} and M_{L1} . The Hebrew cluster included terms related to the court of the tabernacle, among others (translated, with the ranks of the term in the lists of M_{C1} and M_{L1} , respectively, following the term in parentheses): **the court** (59, 31), **cubit** (60, 645), **length** (243, 646), **two** (403, 644), **five** (69, 1193), and several more numerals. These words are related to each other, as the Bible tends to give exact dimensions (note the words **length** and **cubit**) in certain detailed descriptions. The English cluster included terms like **kings** (426, 297), **reigned** (124, 57), **chronicles** (357, 48), etc, and the French cluster **lettre** (157, 29), **écrire** (383, 387), **dîner** (134, 95), etc. The interesting thing to note is that these terms also rank relatively high for both measures, thus also have a linear and condensation clustering tendency. This strongly suggests that the clustering considered in [1], though working in another dimension, is positively correlated with the serial-clustering of the present work, so that those terms which tend to form clusters with other, similar, terms, also exhibit internal, linear-clustering.

Language	Size of cluster	Number of terms	Measure C1		Measure L1	
			average	Z-value	average	Z-value
Hebrew	28	3083	649	-5.33 ($< 10^{-7}$)	975	-3.39 (0.0003)
English	19	2472	667	-3.49 (0.0002)	349	-5.44 ($< 10^{-7}$)
French	26	3595	456	-6.61 ($< 10^{-10}$)	241	-7.68 ($< 10^{-14}$)

Table 2: *Description of clusters. Average refers to the ranks of the clustered terms as imbedded in the total set of terms. Z-value is the standard normal deviate of the Mann-Whitney statistic, followed by one-sided significance level in parentheses.*

Table 2 summarizes the statistics describing the terms making up the three traditional clusters. Included are the number of terms in the cluster and the total number of terms from which the clusters were formed. In the columns headed by measure C1 and L1, we include the average rank of the clustered terms. For example, we see in the Hebrew Bible, the average rank of 649 is well below the midpoint rank of 1543, which would be expected if the clustered terms were in fact distributed at random. We carried out Mann-Whitney tests [15] to assess the likelihood that such small ranks might occur by chance. We include in the table the Z-values (deviations of the Mann-Whitney statistic from its expected value, measured in standard deviation units), with the significance level of the (one-sided) test in parentheses. For example, the ranks of the Hebrew cluster for L1 could have taken values at

least that extreme by chance with probability < 0.0003 . It is very reasonable to conclude that the association of the two types of clustering is not a consequence of chance in this case.

4. Final remarks

The task of identifying good index terms to use in IR requests, and assigning weights to them, is central to modern IR. Traditionally, such methods depended upon overall term statistics. With the availability of machine readable full text, the possibility arises of using topological as well as global statistical information.

One important topological characteristic of content-bearing words is the tendency of their occurrences to clump. In this paper, we developed a few measures of this clumping tendency, along with statistical tests of significance for these measures. We did not carry out any experiments to test whether our measures were effective in identifying documents relevant to a request. Rather, we are trying to develop an intuition regarding the types of words identified by the measures. The impression left by this research is that there is a real tendency for some words to clump, and that this tendency does seem to be associated with words that have meaning. This association is by no means perfect, and different clumping measures vary in their capacity to identify words that intuitively look good (much as is the case with global statistical measures). At our current state of knowledge, it seems that these measures do have a role to play in filtering words that show promise as index terms, and perhaps at introducing structural information into weighting schemes for IR experiments.

Acknowledgements: The authors wish to thank C.W. Sze and P. Lankinen for their help, and the *American and French Research on the Treasury of the French Language* (ARTFL) for providing us access to the TLF database.

Appendix I

Alternative derivation of equation (1)

The formula given for equation (1) was derived by combinatorial arguments that assume each term occurrence is distinguishable. The following dynamic argument confirms this result.

As before, we suppose we have D documents of identical size, W . Imagine composing each document, generating a term at a time. The probability is p that the term in question occurs at any opportunity. The number of occurrences of the specified term in any document is governed by the binomial distribution and approximated by the Poisson distribution with mean $\lambda = pW$. Thus, the probability that the first document has n_1 occurrences, the second n_2 , etc. is given by the product of the probabilities for each document:

$$\exp(-D\lambda) \frac{\lambda^{n_1+\dots+n_D}}{n_1!n_2!\dots n_D!}.$$

Letting $\Lambda \equiv D\lambda$, we have, for the probability that there are all told $T = \sum_i n_i$ occurrences of the term in question over the collection of D documents, the expression $\exp(-\Lambda)\Lambda^T/T!$. Thus, the probability of any distribution of terms conditional on T (that is, the probability that T terms will distribute themselves by n_1 terms going to the first document, etc.) is given by the ratio of these probabilities:

$$P(n_1, \dots, n_D) = \left(\frac{1}{D}\right)^T \binom{T}{n_1, n_2, \dots, n_D}.$$

We are seeking the probability that exactly N documents have terms, that is, the sum of $P(n_1, \dots, n_D)$ over all combinations of n 's satisfying the constraint that exactly N of the D values of n are greater than zero. In this sum, the factor $\left(\frac{1}{D}\right)^T$ is common to all the terms, and the sum is over $\binom{T}{n_1, n_2, \dots, n_D}$. But this sum is just the number ways to distribute T items over D cells in such a way that exactly N cells are occupied. We saw above that value was $N! \binom{D}{N} \left\{ \begin{matrix} T \\ N \end{matrix} \right\}$, and the formula reduces to the one given in the text.

It is interesting to compare this result with that obtained by assuming that terms are not distinguishable — that is, act like *bosons*, to use the terminology of physics. If terms are not distinguishable, then we must look at patterns of distributions. There are $\binom{T+D-1}{T}$ distinguishable patterns of placing T occurrences of a term into D documents. Of these, $\binom{D}{N} \binom{T-1}{N-1}$ will have terms in exactly N documents [10]. If the terms are in fact distributed independently, then each of these ways of distributing terms is equally likely and the probability that exactly N documents have at least one occurrence of the term is given by:

$$p_N \equiv \frac{\binom{D}{N} \binom{T-1}{N-1}}{\binom{T+D-1}{T}}.$$

Given the assumptions of equation (1), an approximate derivation is possible. Let us assume the terms are randomly distributed by “throwing” them at the documents one at a

time. Then the probability that any *single* occurrence of the term miss any given document is given by $1 - 1/D$, and the probability that *all* of the occurrences miss the document (that is, that the document not contain the term) is given by $q \equiv (1 - 1/D)^T \approx \exp(-T/D)$. The probability of at least a single occurrence is given by $p \equiv 1 - q$. We thus can approximate the probability that exactly n documents contain the term by the binomial distribution,

$$f_B(n|D, p) = \binom{D}{n} p^n (1 - p)^{D-n},$$

with expected value pD .

Then the probability, P_N , that N or fewer documents have at least one occurrence of the term is given by:

$$P_N = \sum_{n=1}^N f_B(n|D, p).$$

Appendix II

Relationship between expectation and probabilities

Using the notation introduced in Section 2.1, we show that if there are S different terms, then approximately $E_{P_0} = SP_0$ terms will have a given probability-value P_0 or less.

The number of units in which the i -th term appears defines the value of the random variable \tilde{N}_i . It would be most simple if we could select several threshold values for N and check whether the expected number of terms had values of N less than these threshold values. Unfortunately, the distribution of \tilde{N}_i depends on T_i , the number of times the i -th term occurs in the database, so it is not possible to use the values of N directly to test the assumption of independent term distribution. In order to combine values from terms with different values of T , we create a probability metric as a substitute for N . For the i -th term, we transform N_i to a *probability-value*, $P(N_i, T_i) \equiv \Pr\{\tilde{N}_i \leq N_i | T_i\}$. This notation emphasizes that for each term, its value on the probability metric depends on the total number of tokens of the term in the database as well as on N_i . For example, for a term occurring T_i times, the probability-values it can possibly take are $P(1, T_i), P(2, T_i), \dots, P(T_i, T_i)$, where $P(n, T_i)$ is the probability that the i -th term occur in no more than n units. When i , and thus T_i , is fixed, we use the simpler notation, $P_i(n)$ for $P(n, T_i)$ and P_i for $P(N_i, T_i)$.

For a given probability-value, say P_0 , let $\tilde{\delta}_i(P_0)$ be the random indicator function of whether the probability-value of the i -th term is less than or equal to P_0 . Then $\tilde{I}_{P_0} \equiv \sum_{i=1}^S \tilde{\delta}_i(P_0)$ is the number of terms whose probability-values are less than P_0 ; its expected value is $E(\tilde{I}_{P_0}) = \sum E(\tilde{\delta}_i(P_0))$. Since $\tilde{\delta}$ is a binary random variable, its expected value is just the probability that the probability-value of the i -th term is less than P_0 —that is, the probability that it appear in N units, for a value of N for which $P_i(N) \leq P_0$. For each term, this will be the largest value of $P_i(N)$ not greater than P_0 ; for the i -th term we denote this by $P'_i(P_0)$.

Thus $E(\tilde{I}_{P_0}) = \sum_{i=1}^S P'_i(P_0)$. For a term with a large value of T , the probability $P'_i(P_0)$ will approximate P_0 and $E(\tilde{I}_{P_0}) \approx SP_0$, which we denote in the text by E_{P_0} . (This value is an approximation because of the discreteness of the values $P(n, T_i)$, which is especially

pronounced when T is small: it is unlikely that an N_i will exist for which $P'_i(P_0)$ exactly equals P_0 .)

References

- [1] **Bookstein A., Klein S.T.**, Compression of Correlated Bit-Vectors, *Information Systems*, **16**(4) (1991) 387–400. A shorter form appeared as: Construction of Optimal Graphs for Bit-Vector Compression, in *Proc. 13-th ACM-SIGIR Conf.*, Brussels (1990) 327–342.
- [2] **Bookstein A., Klein S.T.**, Information Retrieval Tools for Literary Analysis, in **Tjoa A.M., Wagner R.**, (eds), *Database and Expert Systems Applications: Proceedings of the DEXA 90 Conference, Vienna, Sept, 1990*, Vienna, Springer Verlag (1990) 1–7.
- [3] **Bookstein A., Klein S.T., Raita T.**, Detecting content-bearing words by serial clustering, *Proc. 18-th ACM-SIGIR Conf.*, Seattle (1995) 319–327.
- [4] **Bookstein A., Klein S.T., Raita T.**, Modeling word occurrences for the compression of concordances, in preparation.
- [5] **Bookstein A., Klein S.T., Raita T.**, Markov models for clusters in concordance compression, *Proc. Data Compression Conference DCC-94*, Snowbird, Utah (1994) 116–125.
- [6] **Bookstein A., Klein S.T.**, Using Bitmaps for Medium Sized Information Retrieval Systems, *Information Processing & Management* **26** (1990) 525–533.
- [7] **Bookstein A., Swanson D.**, A Decision Theoretic Foundation for Indexing, *J.ASIS* **26** (1975) 45–50.
- [8] **Choueka Y., Fraenkel A.S., Klein S.T., Segal E.**, Improved Techniques for Processing Queries in Full-Text Systems, *Proc. 10-th ACM-SIGIR Conf.*, New Orleans (1987) 306–315.
- [9] **Choueka Y., Fraenkel A.S., Klein S.T., Segal E.**, Improved Hierarchical Bit-Vector Compression in Document Retrieval Systems, *Proc. 9-th ACM-SIGIR Conf.*, Pisa (1986) 88–97.
- [10] **Feller W.**, *An Introduction to Probability Theory and its Applications*, Wiley, New York (1968).
- [11] **Knuth D.E.**, *The Art of Computer Programming, Vol I, Fundamental Algorithms*, Reading, MA, Addison-Wesley (1973).

- [12] **Lancaster F.W.**, *Vocabulary Control For Information Retrieval* (2nd ed), Arlington, Virginia, Information Resources Press (1986).
- [13] **Luhn H.P.**, A Statistical Approach to the Mechanized Encoding and Searching of Literary Information, *IBM Journal of Research and Development* **1**(4) (1957) 309–17.
- [14] **Salton G.**, *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley Publishing Company, Reading, MA (1989).
- [15] **Siegel S.**, *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill Book Co., New York (1976).
- [16] **Soergel D.**, *Organizing Information*, Orlando, Florida, Academic Press (1985).
- [17] **Van Rijsbergen C.J.**, *Information Retrieval*, London, Butterworths (1975).