A Model for Word Clustering

James A. Thom Justin Zobel

Department of Computer Science
Royal Melbourne Institute of Technology
GPO Box 2476V, Melbourne 3001, Australia

April 1992

Abstract

It is common to model the distribution of words in text by measures such as the Poisson approximation. However, these measures ignore effects such as clustering: our analysis of document collections demonstrates that the Poisson approximation can significantly overestimate the probability that a document contains a word. Based on our analysis, we propose a new model for distribution of words in text, and show how this model can be used to estimate the probability that a document contains a word and the number of distinct words in a document.

1 Introduction

Models for the distribution of words in text are used to estimate the performance of algorithms for text compression and for full text retrieval from databases. Many of these models, such as the Zipf and lognormal distributions, are used to predict the distribution of word frequencies (Carroll, 1967; Witten and Bell, 1990; Zipf, 1936; Zipf, 1949). However, such measures do not predict the probability that a given document in a document collection contains a particular word. Other models, such as the Poisson distribution, can be used to estimate this probability, but such models are inaccurate, as they ignore effects such as *word clustering*.

Clustering arises from the fact that words tend to be repeated a number of times in the same piece of text, even words that are (overall) quite rare. For example, the effectiveness of adaptive text compression techniques depends in part on clustering effects, where in general the number of bits used to encode a word decreases as the number of recent occurrences of the word increases (Cleary and Witten, 1984; Moffat, 1989). However, clustering has not been measured nor its effects quantified. Without some allowance for clustering, estimating the performance of information retrieval algorithms is difficult. For example, in full text retrieval systems, the cost of answering a query depends on the number of documents that satisfy the query, and the size of an index depends on the number of unique terms in each document.

To obtain accurate estimates of the probability that a document of a given size contains a given word, we analysed several document collections. Our analysis revealed that the Poisson estimate of this probability can be much greater than the observed probability. We propose as a new measure a *clustering model* and show that this model generally provides a good fit to observed data. Like other accurate models of text (Witten and Bell, 1990), this model is empirical.

We show that the clustering model can be used to estimate the average number of documents containing a word of a given probability and to estimate the number of distinct words in a document. There are several possible applications of this measure. It has already been used to estimate the costs of full text storage and retrieval (Zobel et al., 1991) and to estimate the importance of query terms in information retrieval (Wallis et al., 1991).

In Section 2 we describe the methods and results of our analysis of document collections. In Section 3 we discuss existing models for distribution of words in text. We present the clustering model in Section 4, and discuss how it can be verified and applied in Section 5. In Section 6 we discuss the limitations of the clustering model. Directions for further work are discussed in Section 7.

2 Analysis of document collections

In this section we discuss how we analysed our document collections and present the results of our analysis. For the purposes of our analysis, we assumed that a word is a series of alphabetic characters flanked by non-alphabetic characters. As in many information retrieval systems, uppercase and lowercase letters were converted to a single case and the resulting strings were stemmed using Lovins's algorithm (Lovins, 1968), thus reducing the number of distinct terms under consideration. We also assumed that a document is an entire, contiguous piece of text such as a book of the Bible, an Act of Parliament, or a speech, and that a document collection is a set of documents from a single source. The size of a document was measured by the number of words in the document rather than by the number of characters.

We had several document collections available for analysis. We chose to concentrate on three, the King James version of the Bible, the Commonwealth Acts of Australia from 1901 to 1988 (or *Comact*), and an extract from the 1989 Western Australian Hansard, the official transcript of that state's parliamentary proceedings. The sizes of these collections are shown in Table 1. As can

	Number of documents	Number of distinct words	Number of word occurrences	Average document length (in words)
Bible	66	8 892	791 448	11 992
Comact	3 459	21 740	16 548 025	4 784
Hansard	13 545	22 484	4 473 791	330

Table 1: Size of each document collection

be seen, the Bible is very much smaller than both Comact and Hansard, and was principally used to verify results obtained from the latter two collections. The presentation of our results in this paper is principally based on Comact. The three collections have very different characteristics. The Bible is somewhat representative of written English. Hansard is a collection of transcripts of spoken English, and each document (which is a debate on a single topic) tends to be repetitive and informal in structure. In contrast, the language in Comact

is highly formalised, the vocabulary limited, and the text available to us contained a large number of spelling errors, amounting to approximately one-third of all distinct terms. However, within each collection documents are written in a fairly homogeneous style.

Our principal aim was to discover a relationship between: the probability that a given word occurs; document size; and the probability that a document of that size contains that word. For most sizes, however, our document collections contained very few documents of or near that size. Thus for our analysis we chose to select a range of possible lengths and construct psuedo-documents (or *fragments*) of each length. Each pseudo-document was an excerpt of contiguous text from a real document. We generated the psuedo-documents by taking several fragments of each length from each real document. If the originating document was shorter than a given length, no fragment of that length was taken from it. This method of constructing psuedo-documents of given lengths from real documents allowed us to simulate collections of independent documents of a given size drawn from a large document space. We show in Section 5 that the model derived from psuedo-documents can be used to predict aspects of the behaviour of real documents.

The number of fragments of each length was chosen so that about 5% of each document was represented in fragments of that length. The starting point of each fragment was chosen at random, and we did not exclude the possibility of overlap. We considered 26 different fragment lengths between 2 and 100 000 words. For some of these lengths, the number of fragments generated for each of the document collections is shown in Table 2.

	Fragment length						
	40	100	400	1 000	4 000	10 000	40 000
Bible	1 021	431	138	80	37	27	2
Comact	22 530	10 362	3 822	2 191	979	417	46
Hansard	9 741	5 196	1 209	548	241	105	4

Table 2: Number of fragments in each document collection

Word occurrence probabilities for each document collection were computed by counting word occurrences over the whole of the document collection. The probabilities of fragments containing a word were given by counting, for each fragment size, the number of fragments of that size in which the word occurred. These probabilities were computed by maintaining a splay tree (Sleator and Tarjan, 1985) of distinct words with, for each word, several counters.

Throughout this paper, we use n to denote the number of words in a fragment, $p(w)$ or just p to denote the probability that a randomly chosen word in the document collection is the word w (that is, the *occurrence probability* of w), and $p_n(w)$ or just p_n to denote the probability that a fragment of length n contains word w. For all words in Comact, p and p_n (where $n = 100$, 1 000, and 10 000) are related as shown in Figure 1. We have omitted from Figure 1 the few points where $p > 0.005$.

One problem with our method of estimating probabilities is the difficulty of accurately deriving extreme p and p_n values. At very low p values only a few distinct p_n values occur, representing words occurring in only a few fragments; thus, at such p values, patterns such as simple lines appear in graphs. Moreover, since many of the words with very low p did not occur in any of the fragments we generated, our method derived $p_n = 0$, which is clearly an error since all words with some probability of occurrence must have non-zero p_n for all finite n. A similar but less severe problem occurs for very high p values, as words with such p values occur in nearly every fragment. Thus, estimated probabilities with very low or very high p or p_n were substantially inaccurate. In deriving our model in Section 4, we eliminate these extreme values for p and p_n.

3 Existing models of word distributions

Models of word probabilities

Zipf's law is perhaps the best known model of word probabilities. It describes the fact that when words are ranked on frequency, from most to least frequent, plotting rank against frequency yields a hyperbolic curve (Zipf, 1936; Zipf, 1949). However, it has been argued that too much emphasis has been placed on this result: even words produced by a simple random generator conform to Zipf's law (Witten and Bell, 1990).

In any case, Zipf's law, or amendments to Zipf's law such as that proposed by Mandelbrot (Mandelbrot, 1952), do not apply to the problem we are considering: the number of word occurrences and number of distinct words in a document collection do not specify the parameters of the Zipf curve. Nor do these parameters, if known, help determine the probability that a document contains a given term. Although theoretically elegant, Zipf's law provides only a loose fit to actual text, and in practice must be modified by introduction of additional parameters (Witten and Bell, 1990). The lognormal distribution (Carroll, 1967) has similar limitations (Witten and Bell, 1990).

The Poisson approximation

For a sequence of trials in which the probability of each outcome is unchanged between trials (that is, the trials are *equivalent* and *independent*), the probability that exactly m of the trials have a particular outcome is given by the binomial distribution. Where the number of trials is large, the Poisson approximation can be used to estimate this probability using the formula for the Poisson random variable X:

$$P(X = m) = e^{-\lambda} \cdot \frac{\lambda^m}{m!}$$

where λ is the mean number of trials with the desired outcome (Feller, 1968). Thus the probability that at least one trial has a particular outcome is estimated by

$$P(X \geq 1) = 1 - e^{-\lambda}$$

In our context, the probability that a fragment contains one or more occurrences of a particular word can be estimated using the Poisson approximation by

$$pois_n = 1 - e^{-n \cdot p}$$

where $pois_n$ is the Poisson-estimated approximation to p_n. For fragments of sizes $n = 100$, $n = 1\,000$, and $n = 10\,000$, $pois_n$ is shown as the top curve (dashed line) in Figures 2, 3, and 4 respectively. Observed p_n values for each word are shown as points in these figures. These graphs show that the Poisson estimate $pois_n$ significantly overestimates p_n for most values of p. As can be seen, the error in the Poisson estimate becomes greater as n increases.

Although the Poisson approximation can be used to model distribution of words in text, choice of words when speaking or writing is not a sequence of independent trials. Usually, choice of a word is strongly limited by the words preceding it (Witten and Bell, 1990): for example, 'choice' is quite likely to be followed by 'of', but it is most unlikely that 'choice' would be followed by 'the'. For non-text data, it can be assumed that data values are randomly distributed (Yao, 1977), so that numbers of matching records can be estimated via the binomial distribution and hence the Poisson approximation. Christodoulakis has shown that, where this assumption is false, estimates based on the assumption over-estimate the number of matching records (Christodoulakis, 1984). Nonetheless, some authors have implicitly assumed that words in text data are randomly distributed (Kent et al., 1990; Sacks-Davis et al., 1987); our results indicate that this assumption is invalid.

Poisson estimates of the probability that a document contains a word are usually an overestimate, as they are based on the assumption that words are evenly distributed in text. Under this assumption, a rare word that has occurred in a document is very unlikely to occur elsewhere in that document. In practice, however, if any word occurs in a document, the probability that it will occur again, possibly several times, is relatively high. Thus, since a word is likely to occur several times in each document in which it occurs at all, Poisson estimates based on occurrence counts across a document collection will in general be too high.

4 The clustering model

If a word occurs in a document there is a relatively high probability that it will occur more than once in that document. We call this effect *word clustering*. In this section we describe an empirically-determined *clustering model* that relates the probability that a word occurs, document size, and the probability that a document of that size contains the word. This model is not based on a psychological or linguistic theory of language: rather, it is an approximation based on observations of the properties of actual text.

For most words w and for all but small document lengths n, the observed probability $p_n(w)$ is substantially less than $pois_n(w)$. The degree of difference between $p_n(w)$ and $pois_n(w)$ depends on w, since words with the same probability of occurrence can occur in different numbers of documents. The degree to

which the distribution of a word differs from a random distribution has been described by Wallis, Zobel, and Thom (Wallis et al., 1991) as the *topic specificity* of the word.

As can be seen from Figures 2, 3, and 4, the distribution of (p, p_n) values is similar in shape to that predicted by the Poisson approximation. We therefore assumed that a good fit would be given by models of similar form to Poisson. Thus we investigated models of the form

$$clus_n = 1 - e^{-\psi(n,p)}$$

where ψ is a function of n and p, and $clus_n$ is the cluster-estimated approximation to p_n. In order for $clus_n$ to be less than $pois_n$ we require $\psi(n,p)$ to be less than $n \cdot p$. Based on our observations of text, this relationship should hold for all but small documents.

We investigated several ψ functions. The simplest form of ψ that gave a good fit to our data is

$$\psi(n,p) = \alpha^{1-\beta} \cdot n^{\beta} \cdot p$$

where α and β are constants for a particular document collection.

In ψ, the parameter α is the document size above which clustering comes into effect. When $n = \alpha$, $clus_n$ and $pois_n$ are equal. Our formula predicts less clustering than the Poisson approximation when $n < \alpha$. Small documents or fragments, such as this slightly contrived sentence, usually exhibit little clustering and may actually show an opposite effect, because the authors' tend to avoid repeating words in any short piece of text.

The parameter β measures the degree of clustering; the smaller β is, the more tightly words are clustered. When $\beta = 1$, there is no clustering and our formula reduces to the Poisson approximation. In general $\beta < 1$, so if $n > \alpha$ then $(\alpha/n)^{1-\beta} < 1$ and $\psi(n,p) < n \cdot p$. The smaller β is, the greater the difference between $clus_n$ and $pois_n$. Therefore, for collections that are accurately modelled by $clus_n$ with small β, clustering is high: in such collections, words tend to occur frequently in a small number of documents, and are not evenly distributed throughout the collection.

The parameters α and β will vary between collections because they will have different points at which clustering begins to take effect, and because different styles of text will cluster to different degrees. For a particular document collection, α and β can be estimated as follows. Let $k = \alpha^{1-\beta} \cdot n^{\beta}$, which is a constant for given n. Inverting the formula for $clus_n$ and using observed values for p_n we get

$$k = -\frac{log_e(1 - p_n)}{p}$$

for $p > 0$ and $p_n < 1$. In Figure 5 we have graphed k against p_n for Comact fragments of size 1 000. For all but extreme values of p_n (which, as discussed in Section 2, are difficult to derive using our method), k values generally fall between 200 and 700 and are independent of p_n values.

In Figure 6, median k values for each fragment size n are plotted on a logarithmic scale, as suggested by Daniel and Wood's text (Daniel and Wood,

1980). As can be seen, the relationship between log_e(median k) and $log_e(n)$ is near linear, justifying our assumption that $k = \alpha^{1-\beta} \cdot n^{\beta}$ for some α and β. Using regression as described in this paper's Appendix, the solid, straight line can be fitted to the $(log_e(n), log_e(k))$ points. The equation of the line is

$$log_e(k) = \hat{\beta}_0 + \hat{\beta}_1 \cdot log_e(n)$$

Its intercept with the $log_e(k)$ axis is $\hat{\beta}_0$, and its slope is $\hat{\beta}_1$. Taking exponents we get

$$k = e^{\hat{\beta}_0} \cdot n^{\hat{\beta}_1}$$

Hence $\beta = \hat{\beta}_1$ and

$$\alpha = e^{\hat{\beta}_0/(1-\hat{\beta}_1)}$$

which is the intercept of the clustering and Poisson lines.

Note that in Figure 6 the median is based on p_n values in the range $0.25 \leq p_n \leq 0.75$, since as discussed above estimates for p_n values outside this range are inaccurate. Also, this graph contains points for values of n not shown elsewhere in this paper, and points for $n < 60$ or $n > 40\,000$ have been discarded since at these points either there were less than twenty fragments, or there were less than twenty words with p_n values in the range 0.25 to 0.75.

On Comact, the above method for deriving α and β yields $\alpha = 40.8$ and $\beta = 0.734$. For these figures, the estimate $clus_n$ is graphed (continuous line) for Comact fragments of sizes $n = 100$, $n = 1\,000$, and $n = 10\,000$, in Figures 2, 3, and 4 respectively. A summary of α, β, and $\alpha^{1-\beta}$ values and confidence intervals for each document collection is given in Table 3. To derive this table,

	α	$\beta = \hat{\beta}_1$	$\alpha^{1-\beta} = e^{\hat{\beta}_0}$
Bible	2.74 (0.90 to 8.33)	0.920 ± 0.010	1.08 (0.99 to 1.18)
Comact	40.8 (25.2 to 66.0)	0.734 ± 0.014	2.68 (2.36 to 3.04)
Hansard	6.28 (1.89 to 20.8)	0.914 ± 0.012	1.17 (1.06 to 1.30)

Table 3: α and β values for each document collection

the techniques given in the Appendix have been used to give 95% confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$. The ranges of values for α derive from the confidence in $\hat{\beta}_0$ and $\hat{\beta}_1$. Although the α values have substantial variation, the value $\alpha^{1-\beta}$ used in ψ is much more tightly contained, indicating that significant errors in estimation of α only marginally affect the accuracy of the clustering model. In Figure 6 we have graphed the upper and lower 95% bounds, and as can be seen these lines enclose a small region.

There are marked differences between the α and β values for the different document collections; these probably correspond to variations in literary style in the different collections. We have not analysed a large enough number of collections to prescribe typical α and β values, but analysis of further collections has indicated that values near of those of Hansard appear to be common.

n	p	Mean p_n	$pois_n$	$clus_n$
100	10^{-6}	0.000075	0.000100	0.000079
	10^{-4}	0.0062	0.0100	0.0078
	10^{-2}	0.55	0.63	0.54
1 000	10^{-6}	0.00056	0.00100	0.00043
	10^{-4}	0.033	0.095	0.042
	10^{-2}	0.95	1.0	0.99
10 000	10^{-6}	0.0044	0.0100	0.0023
	10^{-4}	0.20	0.63	0.21
	10^{-2}	1.0	1.0	1.0

Table 4: Typical p, p_n, $pois_n$, and $clus_n$ values for Comact

A comparison of mean observed p_n values for Comact with the estimates $pois_n$ and $clus_n$ is shown in Table 4. As can be seen, cluster-model estimates of p_n are always closer to the observed values than are the Poisson estimates.

5 Verifying the clustering model

As can be seen in Figures 2, 3, and 4, the clustering model provides a reasonably close fit to observed data. Another way of examining the accuracy of the clustering model is to consider its predictive capability. In Comact, there are 2 191 fragments of 1 000 words. Therefore, according to the clustering model, on average a word with some probability p would be expected to occur in

$$\mu = (1 - e^{-2.02 \cdot 1000^{0.787} \cdot p}) \cdot 2\,191 = (1 - e^{-464 \cdot p}) \cdot\ 2191$$

fragments. For example, if $p = 10^{-6}$ then $\mu = 1.02$. Now consider all of the M words in Comact whose occurrence probability is near p; continuing the example, there are $M = 282$ words in Comact with p near to 10^{-6}. Some of these words occur in none of the 1 000-word fragments generated, some in one, some in two, and so on. Poisson methods can be used to estimate how many of the M words will occur in exactly m fragments using the formula

$$M \cdot P(Y = m) = M \cdot e^{-\mu} \cdot \frac{\mu^m}{m!}$$

These estimates can be compared with the number of words that occur in exactly m fragments. Graphs of these estimates for $p = 10^{-6}$ ($M = 282$) and $p = 10^{-5}$ ($M = 217$) are shown in Figures 7 and 8 (continuous line). Observed values are shown as points in these graphs. For comparison, we have also plotted estimates yielded by the Poisson approximation for word distribution (dashed line), for which μ is given by

$$\mu = (1 - e^{-1000 \cdot p}) \cdot 2\,191$$

As can be seen, for these probabilities the clustering model gives a fair prediction, for a set of words of a given probability, how many fragments each of the words will occur in; the Poisson model estimates these values very badly.

Similarly good results are given for $n = 100$ and $n = 10\,000$. Unfortunately, for probabilities greater than 10^{-5} there are not enough data points to use this method to verify the model.

These results also explain the wide scattering of k values at low probabilities shown in Figure 5. This scattering is a consequence of the fact that different words of a given probability will occur in different numbers of documents, in accordance with the Poisson distribution discussed above. Hence this technique can be used to give bounds to the number of fragments likely to contain a given word. For example, as can be seen in Figure 8, 95% of the words of $p \approx 10^{-5}$ occur in at least 3 and at most 20 of the 2 191 fragments of 1 000 words.

Another way to verify the clustering model is to use it to predict the number of distinct terms in a document of a given length. Consider a collection of N documents each of length n. Each word w will occur in $p_n(w) \cdot N$ documents. Thus, across the collection, there will be

$$\sum_{w \in W} p_n(w) \cdot N$$

distinct word-document occurrences, where W is the set of distinct terms in the collection. Thus each document should have

$$D_N = \sum_{w \in W} p_n(w)$$

distinct terms.

If p_n values are unknown, the Poisson approximation or the clustering model can be used to estimate these values, based on the p values of the words in the document collection. In Figure 9 we graph the number of distinct words in the actual documents of Comact, and also graph the clustering and Poisson estimates of D_N using $clus_n$ and $pois_n$ respectively to estimate p_n. As can be seen, the clustering model gives a good estimate of the number of distinct words in real documents.

If p values are also unknown, they can be estimated by measures such as the Zipf distribution (Zipf, 1936; Zipf, 1949) or amendments to the Zipf distribution such as that proposed by Mandelbrot (Mandelbrot, 1952). However, this additional level of approximation would increase the degree of error in the approximation.

This approach has been used by Zobel, Thom, and Sacks-Davis (Zobel et al., 1991) to estimate index sizes for different ways of storing text in databases; index sizes are dependent on the average number of distinct terms in each document. They also used the clustering model to approximate the probability that a set of terms appears in a document. This probability is useful because it can be applied to estimation of the number of answers to a query.

6 Limitations of the clustering model

An obvious limitation of the clustering model is that, unlike the Poisson approximation, the parameters of the model vary between collections. However, this limitation also applies to other models of text, such as the Zipf distribution.

Furthermore, it is possible to choose typical α and β values and base estimates on these. Based on the data given in this paper, it is possible to choose values of α and β that are almost certain to exceed the actual values, yet would give values for $clus_n$ that are significantly less than $pois_n$. For example, if α is 60 and β is 0.9, then $clus_n < pois_n$ for all $n \geq 60$.

A more serious problem is that the model is not always a close fit to observed data. Although it is almost always closer than the Poisson approximation, the model is not very accurate for large p values. This inaccuracy is particularly noticeable for large n. This divergence arises because we have assumed that k is a constant for a given document collection and fragment size. In fact, as can be seen in Figure 5, k drops as p_n approaches 1. This effect is almost nonexistent for smaller n, but becomes evident (for Comact) for $n > 5\,000$. The most straightforward solution to this problem is to consider alternative forms of ψ. One form of ψ that seemed promising was

$$\psi(n, p) = 1 - e^{-\gamma \cdot p^{\sigma}}$$

where both γ and σ are dependent on n. Thus $clus_n$ would be defined by

$$clus_n = 1 - e^{-(1-e^{-\gamma \cdot p^{\sigma}})}$$

For any given n, values for γ and σ can be found by fitting curves to graphs of p against $-log_e(1 - p_n)$. The resulting $clus_n$ is a very good fit to the observed data for all p values. However, we could not identify any relationship between n and these parameters.

Another limitation is that the model only applies to a range of sizes within each collection. For example, in Comact α and β were computed by examining fragments between 60 and 40 000 words in length. However, $clus_n$ with these parameters does not provide as good a fit to the data for fragments of other lengths as it does for lengths in this range.

7 Conclusion

We have shown that existing techniques do not give a good estimate of the probability that a document of a given length contains a word of a given probability. We have proposed a new measure that allows for the tendency of words to cluster, the clustering model, and have shown that this model gives a much better estimate of the probability that the document contains the word than does the Poisson approximation. The parameters of this model vary between document collections, and indicate the degree to which words cluster in a collection. We have also shown that the clustering model can be used to give bounds to the number of documents likely to contain a given word, and to estimate the number of distinct words in a document.

There are some interesting problems that this model does not address. One is to model occurrences of compound terms such as word pairs within documents. Another is to model occurrences of documents containing each member of a set of words. Both of these problems are relevant to the problems of text compression and full text retrieval from databases.

Acknowledgements

We would like to thank Kotagiri Ramamohanarao, Ron Sacks-Davis, Peter Smith, Margaret Thom, and Ross Wilkinson for their assistance and comments, and Alan Kent for the excellent `graph+` package. We would also like to thank the anonymous referees for their helpful comments and suggestions.

This work was partly supported by the Australian Research Council and the Collaborative Information Technology Research Institute.

References

Carroll, J. (1967). On sampling from a lognormal model of word-frequency distribution. In Kŭcera, H. and Francis, W., editors, *Computational Analysis of Present-Day American English*, pages 406–424. Brown University Press, Providence, Rhode Island.

Christodoulakis, S. (1984). Implications of certain assumptions in database performance evaluations. *ACM Transactions on Database Systems*, 9(2):163–186.

Cleary, J. and Witten, I. (1984). Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, 32:396–402.

Daniel, C. and Wood, F. (1980). *Fitting Equations to Data*. John Wiley & Sons, Inc., New York.

Devor, J. L. (1982). *Probability and Statistics for Engineering and the Sciences*. Brooks/Cole, Monterey, California.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, volume I. John Wiley & Sons, Inc., New York, third edition.

Kent, A., Sacks-Davis, R., and Ramamohanarao, K. (1990). A signature file scheme based on multiple organisations for indexing very large text databases. *Journal of the American Society for Information Science*, 41(7):508–534.

Lovins, J. (1968). Development of a stemming algorithm. *Mechanical Translation and Computation*, 11(1-2):22–31.

Mandelbrot, B. (1952). An informational theory of the statistical structure of language. In *Proceedings of the Symposium on Applications of Communication Theory*, pages 486–500, London.

Moffat, A. (1989). Word based text compression. *Software Practice and Experience*, 19(2):185–198.

Sacks-Davis, R., Kent, A., and Ramamohanarao, K. (1987). Multi-key access methods based on superimposed coding techniques. *ACM Transactions on Database Systems*, 12(4):655–696.

Sleator, D. and Tarjan, R. (1985). Self-adjusting binary search trees. *Journal of the ACM*, 32:652–686.

Wallis, P., Zobel, J., and Thom, J. (1991). Document ranking, topic, and skew. In *Proceedings of the Second Japan-Australia Joint Symposium on Natural Language Processing*, pages 325–332, Iizuka City, Japan.

Witten, I. and Bell, T. (1990). Source models for natural language text. *International Journal of Man-Machine Studies*, 32:545–579.

Yao, S. (1977). Approximating block accesses in database organizations. *Communications of the ACM*, 20:260–261.

Zipf, G. (1936). *The Psycho-Biology of Language: An Introduction to Dynamic Philology.* George Routledge and Sons Ltd., London, England.

Zipf, G. (1949). *Human Behaviour and the Principle of Least Effort.* Addison-Wesley, Cambridge.

Zobel, J., Thom, J., and Sacks-Davis, R. (1991). Efficiency of nested relational document database systems. In *Proceedings of the Seventeenth International Conference on Very Large Data Bases*, pages 91–102, Barcelona, Spain.

Appendix: Regression

We first describe linear regression (or Gaussian least squares curve fitting) using the notation of Devor (Devor, 1982). Suppose we have: an independent variable x and a dependent variable y; and a series of x values $x_1, \ldots x_m$ and a corresponding series of y values $y_1, \ldots y_m$. For a probabilistic model in which there exists parameters β_0 and β_1, for any fixed value of the independent variable x

$$y = \beta_0 + \beta_1 \cdot x + \epsilon$$

where ϵ is a random variable with mean zero and variance σ^2. The equation

$$y = \beta_0 + \beta_1 \cdot x$$

is called the *true regression line*. We assume the pairs (x_i, y_i) are distributed about the true regression line in a random manner.

The principle of least squares states that among all straight lines $y = \beta_0 + \beta_1 \cdot x$ the least squares line or estimated regression line

$$y = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$$

is that line which minimizes the sum of the squared deviations

$$\sum (y_i - (\beta_0 + \beta_1 \cdot x_i))^2$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are point estimates for β_0 and β_1. The co-efficients $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimize the sum of the squared deviations are

$$\hat{\beta}_1 = \frac{m \cdot \sum x_i \cdot y_i - (\sum x_i) \cdot (\sum y_i)}{m \cdot \sum x_i^2 - (\sum x_i)^2}$$

$$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \cdot \sum x_i}{m}$$

To assign a level of confidence to the co-efficients, an estimate of the variance $\hat{\sigma}^2$ must be found, namely

$$\hat{\sigma}^2 = \frac{\sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i))^2}{m-2}$$

A confidence interval of $100 \cdot (1-\omega)\%$ for β_0 is given by

$$\hat{\beta}_0 \pm t_{\omega/2,m-2} \cdot \hat{\sigma} \cdot \sqrt{\left(\frac{1}{m} + (\frac{\sum x_i}{m})^2 \cdot \frac{1}{\sum x_i^2 - (\sum x_i)^2/m}\right)}$$

A confidence interval of $100 \cdot (1-\omega)\%$ for β_1 is given by

$$\hat{\beta}_1 \pm t_{\omega/2,m-2} \cdot \hat{\sigma} \cdot \frac{1}{\sqrt{(\sum x_i^2 - (\sum x_i)^2/m)}}$$

In these equations, $t_{\omega/2,m-2}$ is the student t distribution; for a confidence interval of 95%, $t_{\omega/2,m-2}$ is approximately 1.960 for large m.

Where there is a non-linear relationship between the independent and dependent variables, such as between n and k in Section 4, we need to use non-linear regression. The linear model can be transformed by setting $x = log_e(n)$ and $y = log_e(k)$. The parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ can be estimated by substituting transformed values x_i's and y_i's into the above formulae, and approximate confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$ can be derived in a similar fashion.

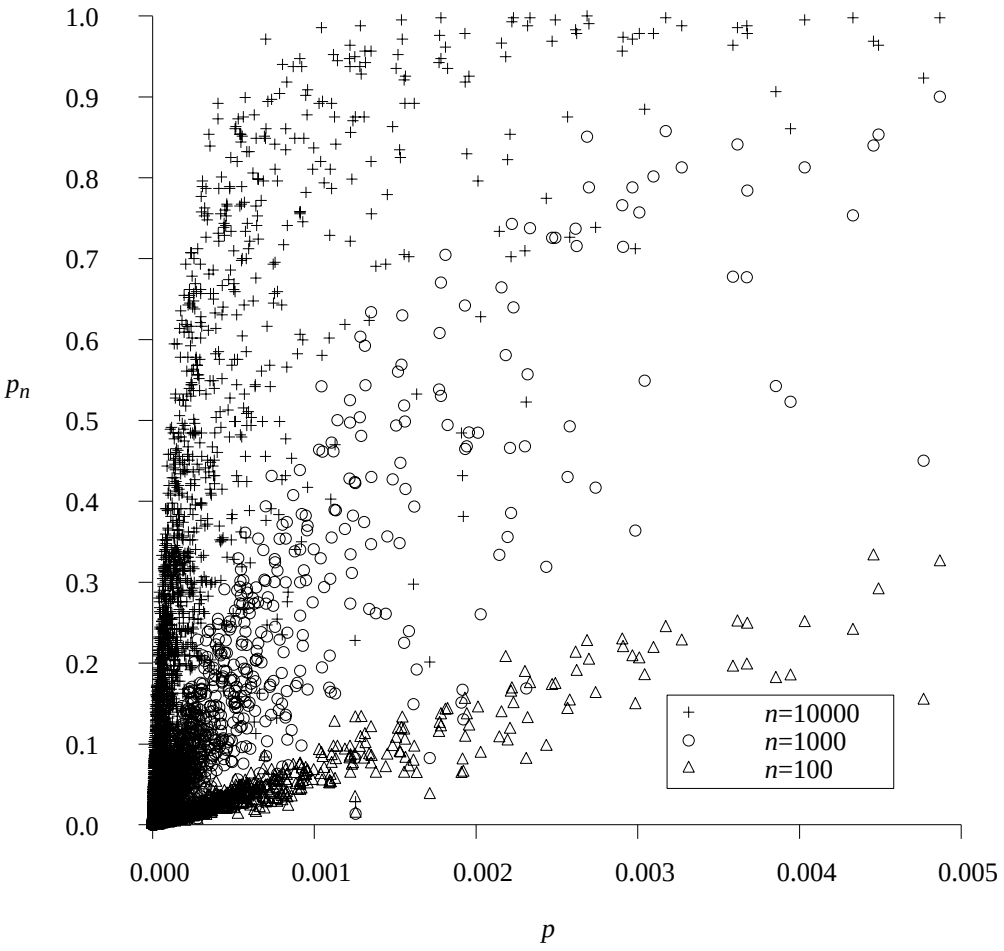

Figure 1: Relationship between p and p_n for Comact

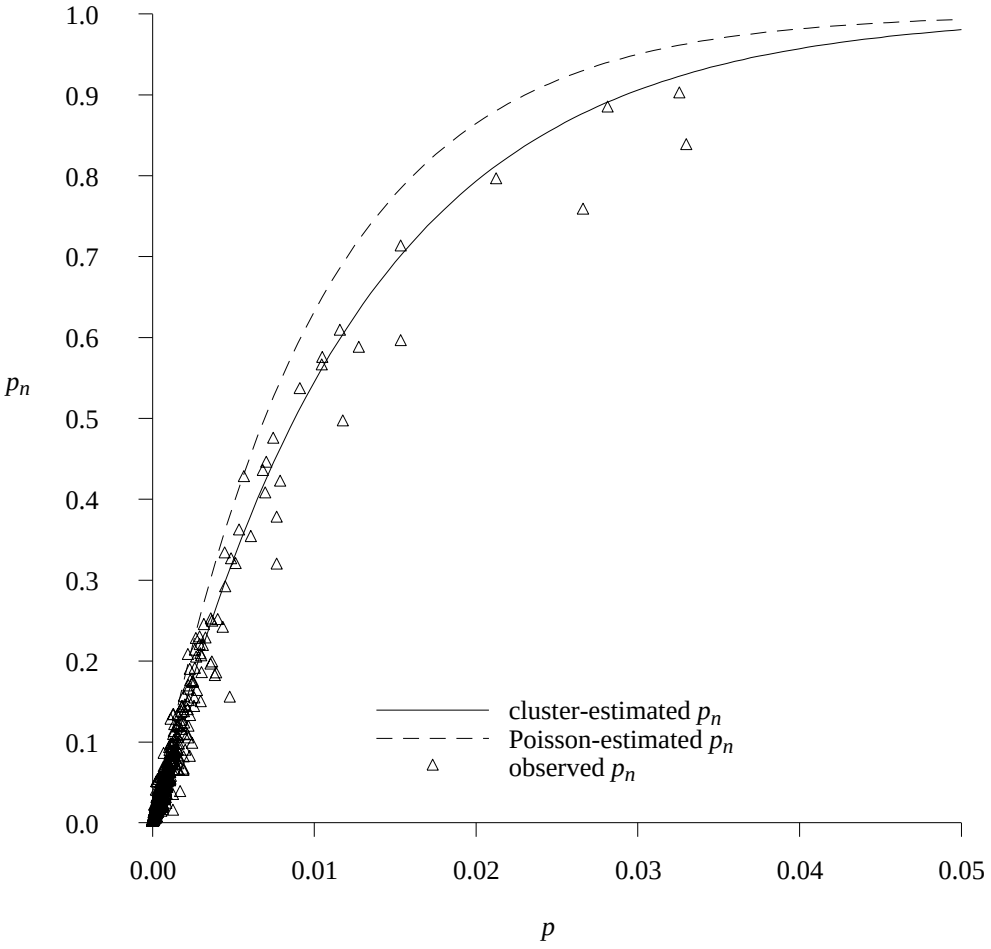

Figure 2: Observed, Poisson-estimated, and cluster-estimated p_n for $n = 100$ for Comact

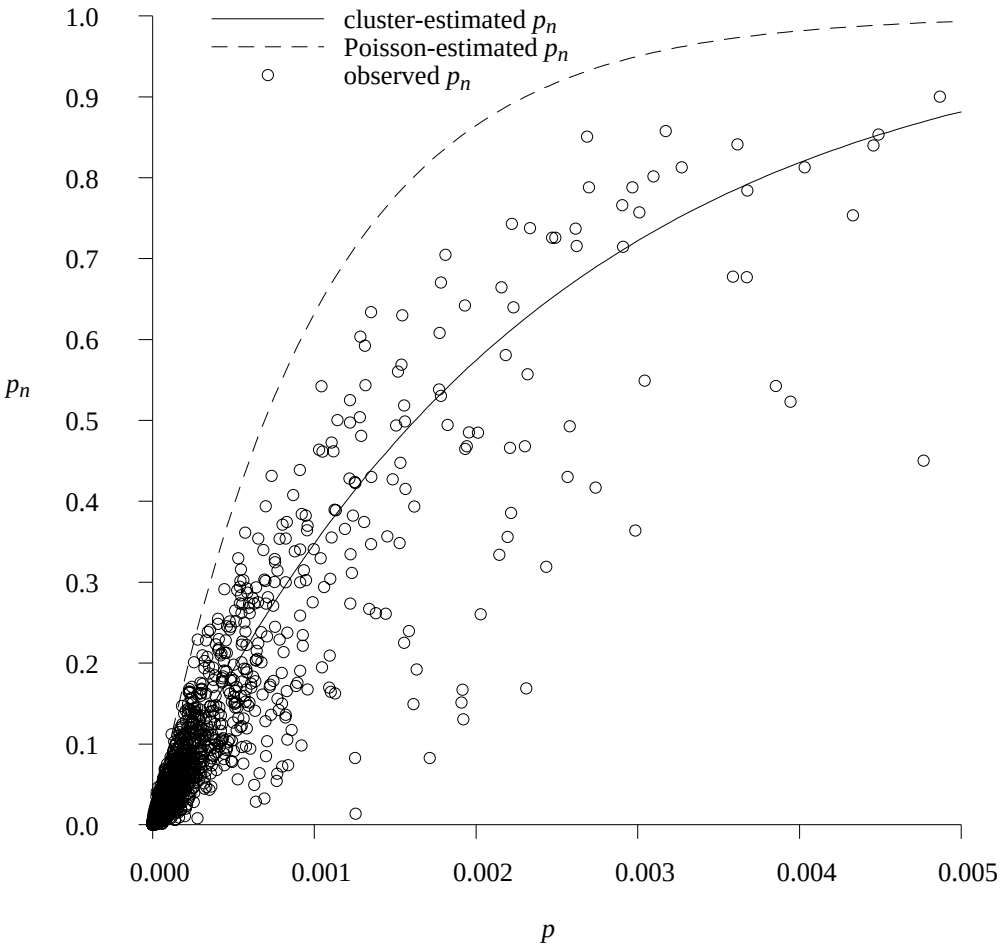

Figure 3: Observed, Poisson-estimated, and cluster-estimated p_n for $n = 1\,000$ for Comact

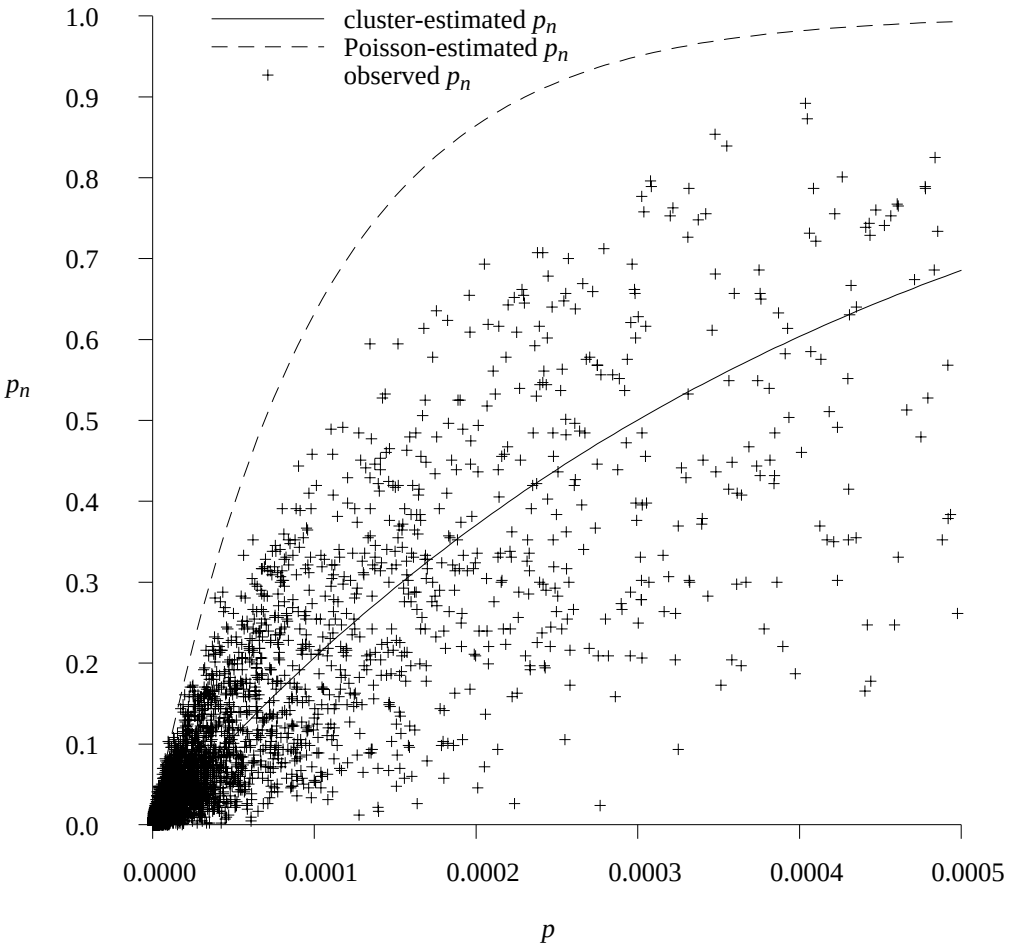

Figure 4: Observed, Poisson-estimated, and cluster-estimated p_n for $n = 10\,000$ for Comact

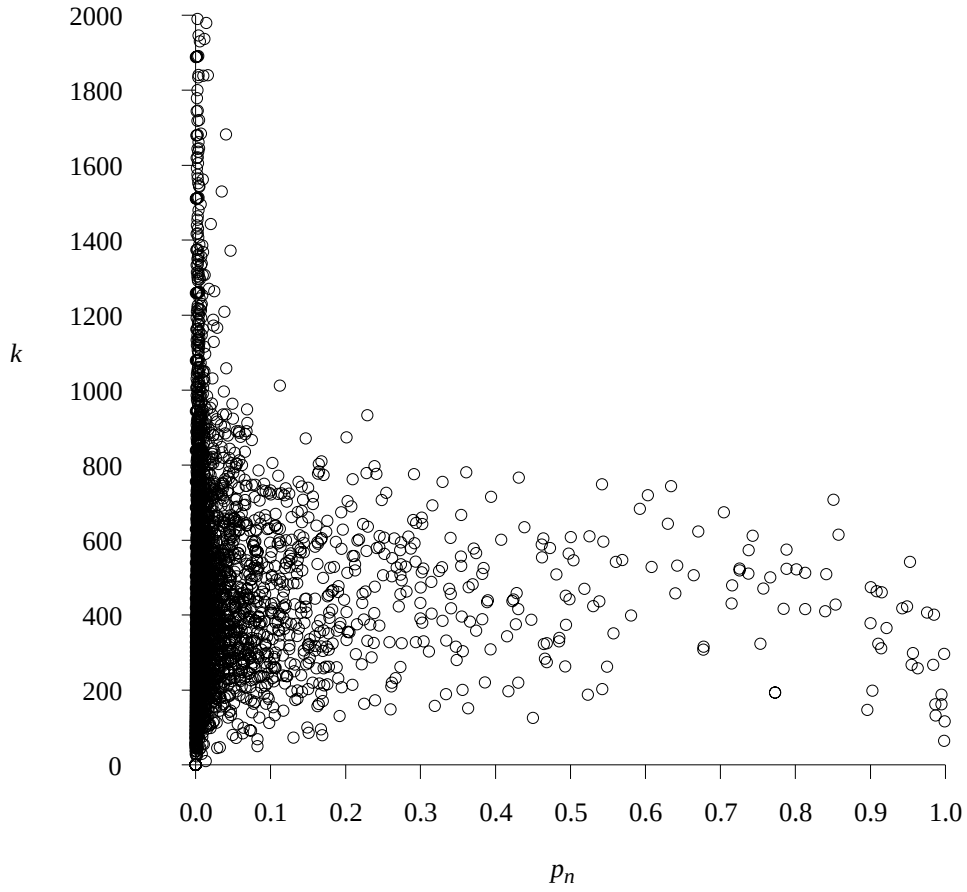

Figure 5: $k = -\frac{log_e(1-p_n)}{p}$ against p_n for $n = 1\,000$ in Comact

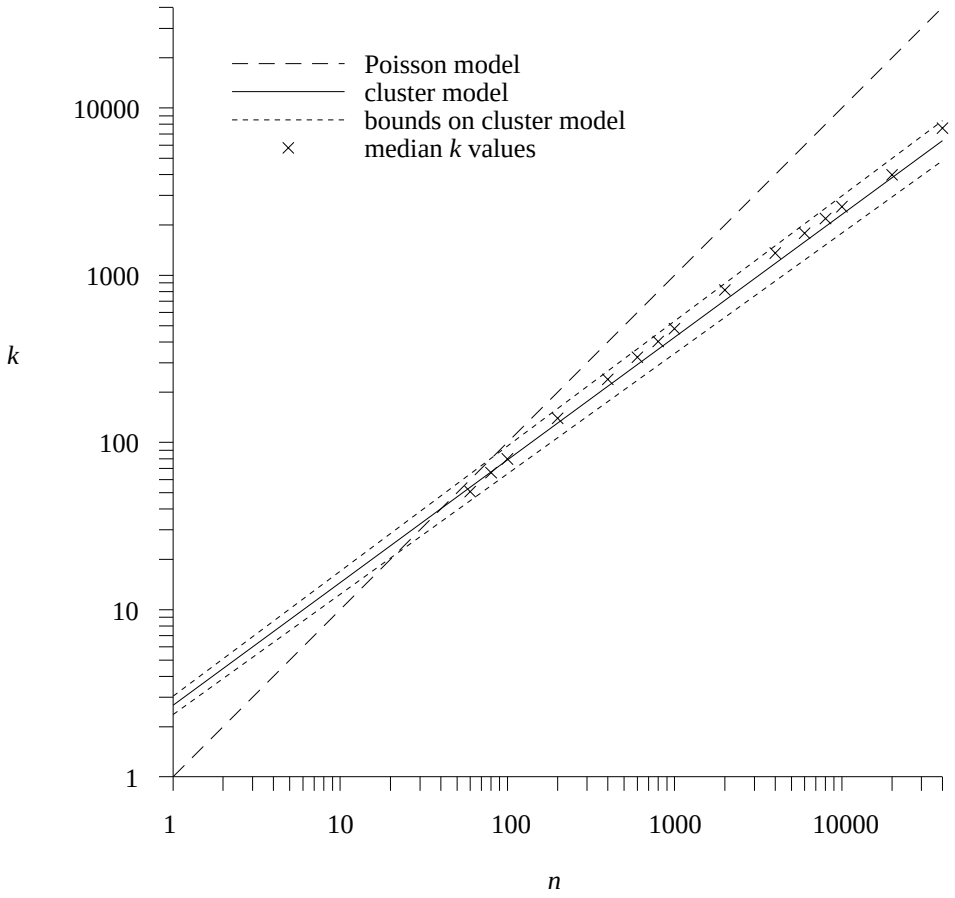

Figure 6: Median $k = -\frac{log_e(1-p_n)}{p}$ against n for Comact

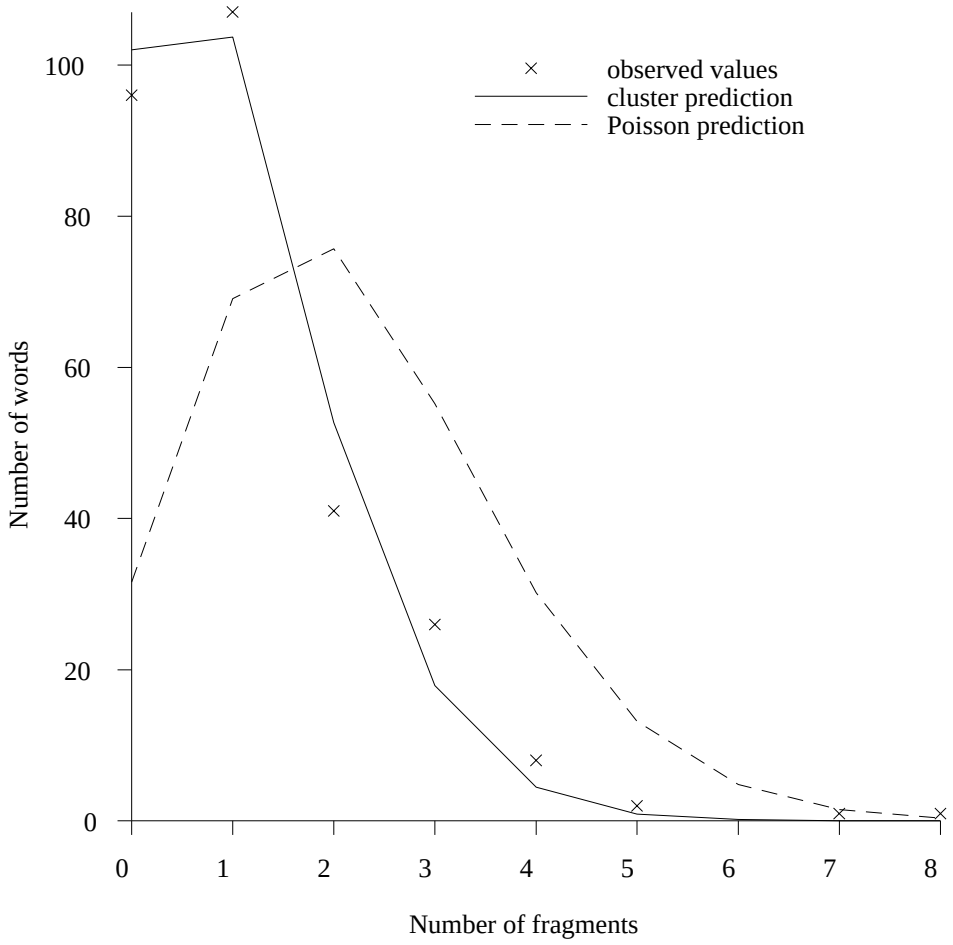

Figure 7: Distribution of p_n values for $p = 10^{-6}$ and $n = 1\,000$ in Comact

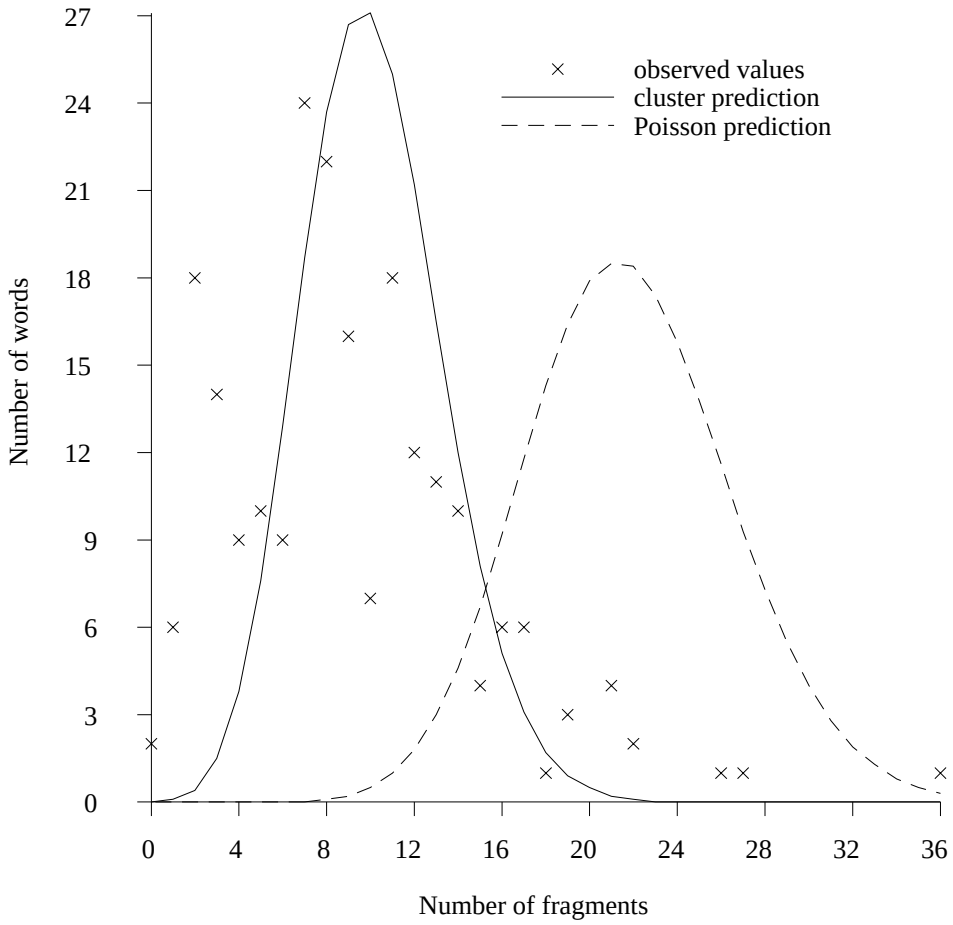

Figure 8: Distribution of p_n values for $p = 10^{-5}$ and $n = 1\,000$ in Comact

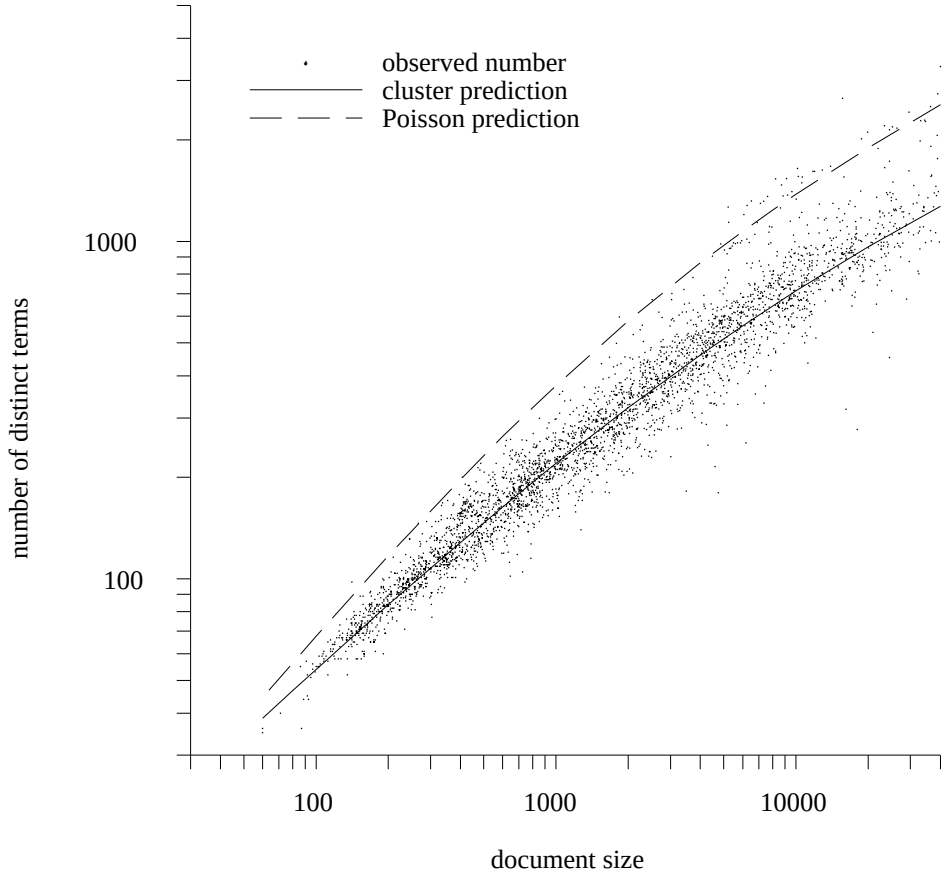

Figure 9: Number of distinct terms in each document of Comact