A queuing-theory model of word frequency distributions

Robert Munro
University of Sydney
rmunro@it.usyd.edu.au

Abstract

This paper describes a novel model for term frequency distributions that is derived from queuing-theory. It is compared with Poisson distributions, in terms of how well the models describe the observed distributions of terms, and it is demonstrated that a model for term frequency distributions based on queue utilisation generally gives a better fit. It is further demonstrated that the ratio of the fit/error between the Poisson and queue utilisation distributions may be used as a good indication of how interesting a word is in relation to the topic of the discourse. A number of possible reasons for this are discussed.

1 Introduction

When a given term occurs in a document, what is the probability that it will occur again, and how will this probability change when you have seen 1, 2 or n instances of this term? While NLP is a vast field, the majority of research is devoted to looking for intelligent ways to automate and speed up processes such as grammatical and/or semantic parsing, which a human could complete with an equivalent or much higher accuracy. Where NLP can exceed human accuracy is in extraction of patterns from large scale corpora (Banko and Brill, 2001). Lexical systems, such as term collocations, co-occurrences and frequency distributions, are one such area.

Typically, a word's occurrence across documents have been assumed to be Poisson in distribution. When this hasn't worked, a combination of two Poisson mixtures has been attempted with suggestions that when this fails combinations of three or more are appropriate (Bookstein and Swanson, 1974; Church and Gale, 1995).

1.1 Contribution of this paper

It is demonstrated here that the distribution of a queue utilisation (QU) is a more accurate representation for word frequency distributions than a Poisson distribution. A Poisson distribution captures the 'burstiness' of the frequency of a given word, otherwise known as the clustered-ness or self-collocation of a word. Here, it is shown that while the underlying assumption of burstiness holds, there are better representations for capturing this.

While it has previously been observed that words that best singularly describe the topic (variously labeled the 'interest words' or 'content words' of a discourse) are the words that *least* obey a Poisson distribution (Church, 2000), this has typically been used only as an observed trend, and without the search for what method best fits modelling this other distribution (if any).

The significance of the relationship between the fits of a Poisson and QU distribution is explored, and it is demonstrated that the level of disparity between the fits of Poisson and QU distributions can be used to extract words that best singularly describe participants in news topics with a very high level of accuracy.

This paper reports some preliminary results

in using such a method on a large corpus of newswires, and discusses the results in relation to this register.

1.2 Outline

Section 2: background and related work.

Section 3: definitions of the equations use here.

Section 4: testing framework.

Section 5: results of testing.

Section 6: discussion & conclusions.

2 Background and Related Work

In terms of an attempt to describe the reasons for variance in the repetition of certain terms, the most recent work in this area is (Church, 2000) who defined the increased probability of a word occurring once it has already been seen 'positive adaptation'. It was the observation of the distributions described in this paper, along with the common trend of burstiness in network communications, that provoked the exploration of a queuing-theory model of word frequency distributions. There, and in (Church and Gale, 1995) the departure from a Poisson distribution were described as the result of 'hidden variables' relating to variation in register and author.

Word frequency distributions have been exploited in almost all tasks in computational linguistics and NLP, from subject boundary detection (Richmond et al., 1997), to information extraction/retrieval (Lynam et al., 2001; Pirkola et al., 2002; Jones et al., 2000).

The author's not aware of any previous word frequency distribution research using queuing theory models. Most previous research has used models based on information theory, presumably because it was to be used for information extraction/retrieval. Nonetheless, the corpus used here are newswires, so in the spirit of McLuhan, a model is investigated that is derived from the medium of communication.

3 Equations for probabilistic distributions

The goal of matching a word frequency distribution to a statistical equation is to find the optimal parameter value(s) for that equation, such that it fits the observed distribution with as little error as possible. There are different metrics and methodologies for calculating goodness of fit, including entropy measures, least squared regression and the practical effectiveness of the information when applied to some NLP task. Here, the chi-squared (χ^2) test is the metric used. This was appropriate for two reasons. It was the measure minimised by the optimisation algorithm (see below), and it is insensitive to error on the x axis, which is desirable as this represents an ordinal distribution and it is chiefly the divergence from the expected probability that is of interest.

3.1 Equations Used

These are the four equations that were tested here:

The Poisson distribution ($P(x)$):

$$\frac{e^{-\mu}\mu^x}{x!} \tag{1}$$

The Double Poisson distribution ($P_2(x)$):

$$\alpha\frac{e^{-\mu_1}\mu_1^x}{x!} + (1-\alpha)\frac{e^{-\mu_2}\mu_2^x}{x!} \tag{2}$$

The Queue Utilisation ($QU(x)$):

$$\rho^x(1-\rho) \tag{3}$$

The Double Queue Utilisation ($QU_2(x)$):

$$\alpha\rho_1^x(1-\rho_1) + (1-\alpha)\rho_2^x(1-\rho_2) \tag{4}$$

A weighted combination of a QU and Poisson distributions was also considered. It is desirable as only a Poisson distribution initially allows $P(x) < P(x+1)$ (for values of $\mu > 1$). Some brief testing confirmed it gave accurate fits, but it is not included in the testing and analysis here as:

1. It is difficult to find a theoretical justification for combining the two in this manner. If the only advantage is that a Poisson distribution allows $P(x) < P(x+1)$, then perhaps there exists some other equation that can model this in manner that is not so ad-hoc. Figures 3 and 6, which do possess this property, actually had less error in QU distributions, although it may take a close analysis of the graphs to see this. What is easier to see is that a combination of the two would definitely give a much better fit.

2. For an accurate comparison with the combinations of a single distribution type, it would be necessary to calculate the extent of overfitting that occurs, increasing and complicating the work described here.

3.2 Poisson

A Poisson distribution is often used to model systems where the *probability* of an observation is very low, but the *possibility* of an observation is very high.

In information retrieval, a double Poisson distribution is often favoured (Robertson and Walker, 1994). It is simply the weighted combination of two Poisson distributions.

Even with the additional parameters of the second μ and the weighting factor, a double Poisson distribution still fails to accurately capture word frequency distributions, but extending the equations to the combination of three or more Poisson mixtures will have the additional problem of the extra parameters to optimise. The most obvious disadvantage of this, other than the computational expense, is the potential for the resultant curve to overfit the data.

Church and Gale addresses this by suggesting Poisson mixtures (Church and Gale, 1995), which are an arbitrary number of Poisson distributions, allowing multiple values for μ, but varying according to a single density function, and therefore only adding one extra parameter. The assumption maintained here is that the underlying distribution *is* Poisson in nature, which is later noted as problematic (Church, 2000).

3.3 Utilisation Queue

In queuing-theory, a utilisation queue is an equation representing the probability of some queue having a given number of entities 'waiting' in it.

A queuing system is often described by the general M/G/1 model. Where:

ρ = The occupancy
μ = the average service rate
σ = the variance of the service time
c = $\sigma\mu$

The system is expressed by:

$$M/G/1(x) = \frac{\rho^x(1+c^2)}{x(1-\rho)} \tag{5}$$

Any queue may be described in terms of this equation, such as people waiting in line to buy tickets at a movie cinema, or the traffic on a computer network. The second of these, and its most common usage, is interesting in this context as it also represents a form of communication, albeit one a little more abstracted.

Within this, the model is often specialised to the M/M/1 model. It makes two assumptions about the data: both the queue arrival rate and the service rate are Poisson. Its relationship to a Poisson distribution is actually a little more involved, but the description of this is outside the scope of this paper and not directly relevant. This is given by:

$$M/M/1 = \frac{\rho^x}{x(1-\rho)} \tag{6}$$

The equation used here, the Queue Utilisation ($QU(x)$) model, is an equation describing the probability that an M/M/1 system has x items in it at a given point in time.

There are several immediate aspects of the QU equation that are appealing:

1. Simplicity. By Occam's Razor, if this were to model as accurately as a more complicated equation, then it would be the more desirable choice.

2. Extensibility. As there already exist, within queuing theory, extensions to the given equations to deal with various phenomena relating to density and limitations on queue size, these might be applied to the given situation.

3. Slower tail-off than Poisson. For large values of x, the tail-off is slower than for Poison, explicitly capturing the phenomena described both in, and by the title of, (Church, 2000).

A related equation called M/D/1 was tested but results indicated it was not appropriate for this task, but other variations of M/G/1 may have desirable properties.

Following a double Poisson distribution, a weighted combination of a two Queue Utilisation distributions was also considered.

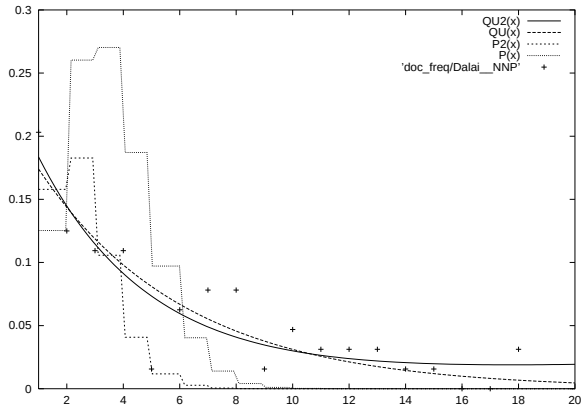

Figure 1: Fits for the term ‘Dalai’

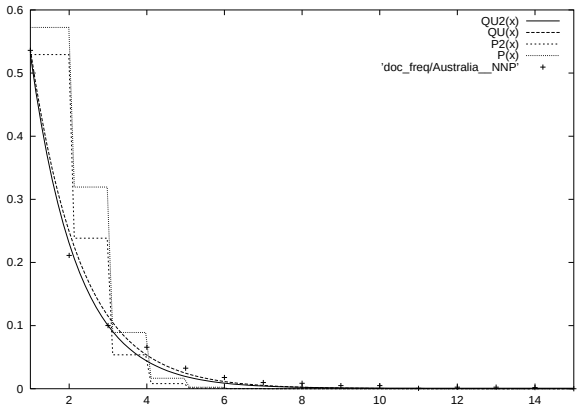

Figure 4: Fits for the term ‘Australia’

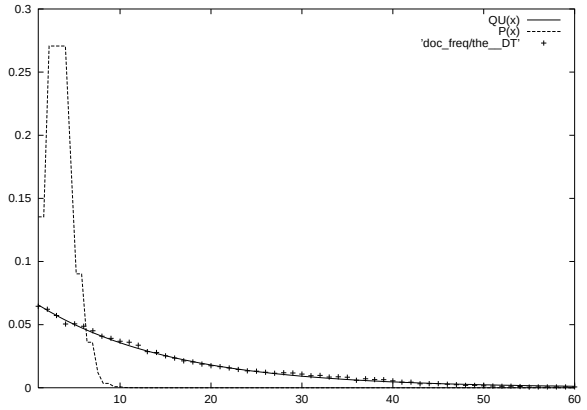

Figure 2: Fits for the term ‘the’

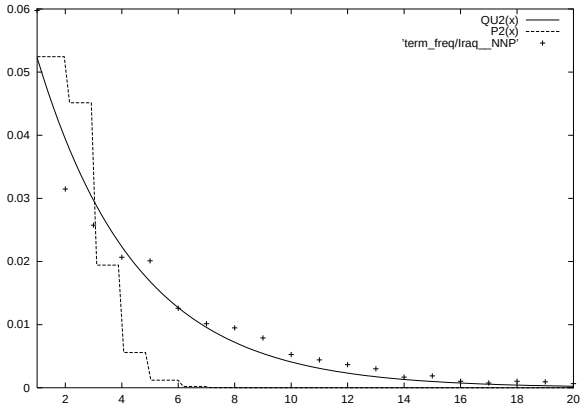

Figure 5: Fits for the term ‘Iraq’

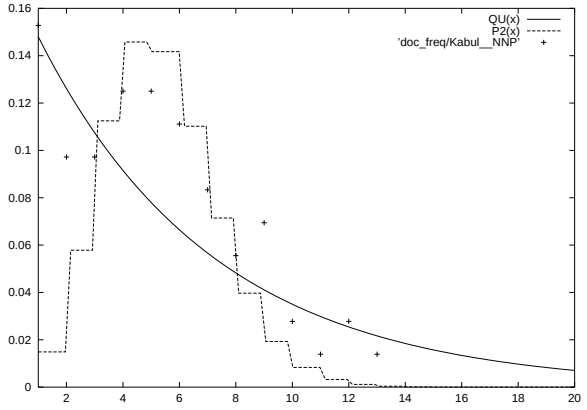

Figure 3: Fits for the term ‘Kabul’

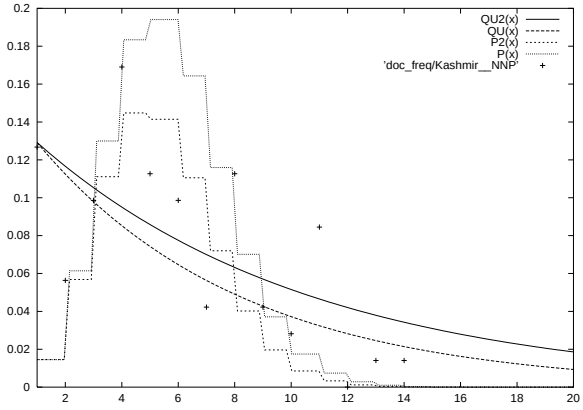

Figure 6: Fits for the term ‘Kashmir’

4 Testing Framework

4.1 Corpus

The corpus used was a set of 50,000 *Reuters* newswires from the year 1996. This constituted approximately 15 million words. Newswires make an interesting register for this sort of analysis as each article typically addresses only one main topic or event, which is probably a better delimiter than using the raw number of words or sentences.

The words were marked up for part-of-speech by the brill-tagger (Brill, 1992). The terms considered were the word/POS pairs, so that a rough division between uses occurred. After some consideration, it was decided that the words should *not* be stemmed or lemmatised - either the distributions of all cases/tenses of a word are similar, or there are differences which may themselves be interesting to investigate.

To reduce noise, words that occurred with a frequency of less then 100 were omitted. Words that never occurred in a document with a frequency of 4 or greater were also omitted, as the differing curves of the equations used would not be emergent.

This resulted in 4935 word/POS tokens. The total number and frequencies of the occurrences of each of these were collated across all documents.

The recorded distributions were then normalised to be tested in two ways:

1. *document frequency*. All document frequencies were divided by the total number of documents containing the given frequency of the word. This is a 'true' normalisation, in the sense that the exact distributions are maintained.

2. *term frequency*. All document frequencies were divided by the total number instances of the term occurring with the given frequency. This will make the tokens with a relatively high level of self-collocation (positive adaptation) more emergent, relative to the document frequency, and is similar to the $tf.idf$ measure common to information extraction/retrieval and document classification.

POS	doc freq	term freq
noun	95.9%	99.9%
prop n.	93.5%	99.8%
verb	90.5%	100.0%
adj	91.8%	100.0%
adverb	90.9%	100.0%
all	94.3%	99.9%

Table 1: Percent of terms for which $QU(x)$ was a better fit than $P(x)$

POS	doc freq	term freq
noun	94.9%	94.6%
prop n.	91.6%	90.9%
verb	90.5%	89.9%
adj	89.6%	89.0%
adverb	89.8%	87.5%
all	93.1%	92.6%

Table 2: Percent of terms for which $QU_2(x)$ was a better fit than $P_2(x)$

4.2 Best Fits

The optimal models were found using gnuplot (Williams and Kelley, 1991), which implements the Marquardt-Levenberg algorithm for parameter optimisation (Press et al., 1992) . The equations were prevented from allowing $\mu < 0$ or $\rho < 0$, as this would give negative probabilities not appropriate here.

The success of the fits were gauged by the final sum of squares of residuals or the χ^2 result on the same term across the different equations.

Once set up, the entire process, from the single scan of the corpus, to the fitting of the equations, took only a couple of minutes to run on a current PC, the majority of this being the time taken for the 'double' distributions to converge.

5 Results

Results are reported for the relative success of the different distributions, given by the percentage of terms for which a given equation has a lower χ^2. These are given as totals and for the divisions into noun, proper noun, verb adjective and adverb parts-of-speech.

For unknown reasons, gnuplot was unable to converge the double Poisson for about 20 terms.

POS	doc freq	term freq
noun	68.7%	0.8%
prop n.	60.0%	1.5%
verb	60.1%	1.9%
adj	64.0%	0.8%
adverb	61.2%	1.1%
all	64.9%	1.1%

Table 3: Percent of terms for which $QU(x)$ was a better fit than $P_2(x)$

Comparisons were made between the fits of all Poisson / queue utilisation combinations. There were no significant results for which the single Poisson was more accurate than a double QU. The other results are given in Tables 1, 2 and 3.

The results demonstrate clearly that a queue utilisation model of word frequency distributions is a more appropriate representation for modelling term frequencies. The most impressive results are in Table 3, showing that even a single QU, with only one free variable, can provide a more accurate representation of the data than a double Poisson, which has two free variables plus the weighting factor.

A cursory glance at the accuracies in Table 3 would seem to indicate that modelling term frequency is less desirable than word frequency, despite the Table 1 accuracies being significantly better. However, looking at the best represented words by each, given by the largest ratio of a QU to a Poisson distribution shows the opposite. The top 1% of words for which $QU(x)$ was a better fit than $P_2(x)$ for term frequency is given in Table 4 (*my groupings*). These distributions are not emergent for the top 1% of words for document frequency, given in Table 5.

The trend for the term frequency distribution continued, with the next 1% of containing terms such as *Ansett, Chechnya, Iraq, Mother Theresa, Pope, Tibet* and *Tyson*, the one outlier being the determiner *the*. By contrast, the bottom 0.5% of terms, given in Table 6, are relatively mild and not particularly indicative of what the topics may be.

The ordering of terms for which a single $QU(x)$ was a better fit than a single Poisson corresponded well to the ordering of terms for the double Poisson, with the majority of terms in the top 1% also appearing in Table 4. Those that aren't in Table 4 are thematically related and include *Harridine, Holbrooke, Izetbegovic, rupiah, Srinagar* and *Zavgayev*.

Category	**Terms**
civil war/unrest:	Pol Pot, Ieng Sary, Rouge, Taleban, Kabul, Jalalabad, Afghan, Bossi, Kashmir, Albania, Liberia, Dalai, Hutu, Burundi
financial terms:	Grade, Prop, Select, asset-backed, min, m0, Jan-July, Level
(troubled) people:	Murtaza Bhutto, Portillo, Summers, Pen, Dutroux, Espy,
(troubled) currencies:	tolars, dinars, kroons, dirhams, bolivars, drachmas,
(troubled) companies:	Loewen, Michelin, Tractebel, Olivetti, Erste, Truck, DWT,
other:	inning, homer, vs, Manitoba, canola, Chernobyl

Table 4: Terms for which the $QU(x)$ gave the best fit, as compared to $P_2(x)$ for term frequency

Similarly, the other document frequency ratios gave orderings of ratios as relatively uninteresting as the set of words in Table 5.

Although the QU's best described terms were proper nouns and nouns, and the worst more of a mix of adverbs and verbs, the various parts-of-speech were represented about equally as well as each other. A minor exception can be seen in Tables 2 and 3 where the double Poisson was a little less able to improve the representation of some nouns.

In brief, while a document frequency model more accurately described the distribution of terms, a term frequency model would seem to be much more useful for comparing with Poisson to extract topics from a very large volumes of information.

6 Discussion and Conclusion

There are some proper nouns that serve as more static exophoric referents than others. The ones in Table 4 are among the most strict, with the notable absence of even the most common colloca-

from, which, with, it, an, then, spokesman, was, to, in, and, of, a, by, the, seek, not, but, for, bullish, discounts, before, we, trader, added, at, day, will, as, casualties, been, up, because, durum, cruise, futures, its, on, between, sharply, firmer, be, after, hand, this, have, wheat, while, nation, now

Table 5: Terms for which the $QU(x)$ gave the best fit, as compared to $P_2(x)$ for document frequency

Nigerian, believed, strength, steady, beyond, treasury, none, midday, critics, relatively, approved, thought, began, begin, making, delayed, expect, getting, accept, finished, possibly, failed, mainly, once, recently

Table 6: Terms for which the $QU(x)$ gave the worst fit, as compared to $P_2(x)$ for term frequency

tions of first name or title. This means that the terms in Table 4 are also among the least polysemous, and therefore the most pure in their distributions in this regard. The same can be said for the demonstrative *the*, also an explicit referent although typically as an anaphoric reference in written text.

The 'financial' words in Table 4 can be explained simply in terms of their being generic economic terms that are commonly tabulated or used in lists. They are not discussed further.

It is telling that in the top 1% of best represented terms, all the named entities are of persons, places or organisations involved in some sort of conflict. As a topic extraction task this is quite an achievement, but to a certain extent it can also be seen as a product of the register.

Assuming a word is a good indication of topic, there are two broad reasons that the word may also exhibit high 'adaptation':

1. The nature of the subject matter is complex, requiring a longer article and the introduction of more themes/participants.

2. The word represents one of multiple participants, and the explicit repetition of the word is chosen instead of the use of anaphora, in order to maintain cohesion.

A third reason, that is worth speculating about but is hard to confirm in a study necessarily abstracted from the text itself by scale, is that the phenomenon is capturing an underlying subjectivity. In news reporting, the use of narratives of conflict are not limited to articles describing military conflicts, often being used in sports reporting. Here, only the terms *inning* and *homer* evidence this. This may simply be because sports reporting is known to favour the continuity of Subject, and therefore *is* able to make more use of anaphora, while maintaining Subject in the reporting of a military conflict is more likely to be viewed as Subjectivity and is therefore avoided, at least grammatically. But perhaps there is something about a term like *Taleban militia*, which, independent of context, makes a news reporter desire to repeat it more than a term like *St Louis Cardinals*, and that this is, in part, a measure of attitude.

Relating to the first reason above, in the reporting of conflict, the explicit attributing of actions and possessions to entities is common, and so repetition will not just be as group/phrase heads but as possessives, *Burundi's army*, and Classifiers, *Burundi radio*.

While these properties will not always hold, it is the emergence of these patterns over large scales that gives these words greater positive adaptation. Even if the results indicated by Table 4 are mostly a product of the lack of anaphora and ellipses, rather than their 'content' or newsworthiness, they are still quite an accurate set of terms for describing topic for any technique that *doesn't* look at grammatical systems.

As stated, a Poisson distribution is often used to model systems where the probability of an observation is very low, but the possibility of an observation is very high. In reality, if a word relates closely to the topic, then the probability of observation is high, but the possibility of (coherent) uses becomes low (in a report on Afghanistan, the term *Kabul* cannot not easily be made a referent to anything other than a specific city). Assuming that for a word to be commonly repeated more than three or four times in a document indicates that it is likely to relate closely to the topic, then for all such words it is to be expected that a Poisson distribution is inadequate.

A brief manual inspection of the terms that occur with a total frequency greater than 100, but never with a document frequency of four or greater (there were about 1700 additional words meeting this criteria) showed that, beneath the noise, these were disproportionately likely to be Epithets, realised by either an adjective or a gerund. This is not surprising, as when a term is functioning as the property or evaluation of an entity, not part of the entity itself, there is rarely the need to remind the reader/listener of this property/evaluation in a short discourse. An additional experiment with words occurring with a maximum document frequency of three was also conducted, giving a further small advantage in overall fit to the QU distribution, but no significant additions to the word lists already given.

The term *Noriega* (Church, 2000) appeared with frequency less than 100, and was therefore omitted from the experiments. This shows the other known 'burstiness' property of term frequencies, that they are also temporal. The articles used here were deliberately taken from a small window of time to negate this, but it would be interesting to see if the same model could be used for modeling temporal distributions independently of, or in addition to, the frequency distributions.

As conflict discourses feature multiple participants, the investigation of co-occurrence would be interesting, as too would be the investigation of collocations/n-grams.

For the most part, the part-of-speech tags alone would have been enough to cluster the various words into the groupings given in Table 4, but with information about the words, such as the error and discovered optimal ρ, it would be interesting to see what additional tendencies could be discovered and/or exploited in a variety of NLP tasks.

Acknowledgements

Thankyou to the ongoing support of my supervisors, Geoff Williams and Sanjay Chawla.

References

M. Banko and E. Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Meeting of the Association for Computational Linguistics*.

A. Bookstein and D. Swanson. 1974. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 25(5):312–318.

E. Brill. 1992. A simple rule-based part-of-speech tagger. In *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing*.

K. Church and W. Gale. 1995. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190.

K. W. Church. 2000. Empirical estimates of adaptation: The chance of two Noriegas is closer to $p/2$ than p^2. In *Proceedings of The 18th International Conference on Computational Linguistics (COLING-2001)*.

K. S. Jones, S. Walker, and S. E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments - part 2. *Information Processing and Management*, 36(6):809–840.

T. R. Lynam, C. L. A. Clarke, and G. V. Cormack. 2001. Information extraction with term frequencies. In *Proceedings of the Human Language Technology Conference (HTL'01)*.

A. Pirkola, E. Leppänen, and K. Järvelin. 2002. The RATF formula (Kwok's formula): exploiting average term frequency in cross-language retrieval. *Information Research*, 7(2).

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 1992. *Numerical Recipes in C*. Cambridge University Press, 2 edition.

K. Richmond, A. Smith, and E. Amitay. 1997. Detecting subject boundaries within text: A language independent statistical approach. In Claire Cardie and Ralph Weischedel, editors, *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*.

S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*.

T. Williams and C. Kelley. 1991. Gnuplot - an interactive plotting program.