Relevance Weighting Using Distance Between Term Occurrences

David Hawking
Paul Thistlewaite[*]
Cooperative Research Centre For Advanced Computational Systems
Department of Computer Science
The Australian National University
Canberra ACT 0200 AUSTRALIA

{*dave,pbt*}*@cs.anu.edu.au*

January 25, 1996

[*]Equal joint authors.

Abstract

Recent work has achieved promising retrieval performance using distance between term occurrences as a primary estimator of document relevance. A major benefit of this approach is that relevance scoring does not rely on collection frequency statistics. A theoretical framework for lexical spans is now proposed which encompasses these approaches and suggests a number of important directions for future experimental work. Based on the formalism, approaches to issues such as scoring partial spans, treatment of repeated term occurrences within spans, and the importance of ordering are proposed. Consideration is given to the practical application of the formalism to both locating and scoring concept intersections and to locating phrases (with an estimate of confidence) despite intervening or substituted words.

1 Introduction

The idea that the relative positions of query terms within a document may supply information about relevance arose nearly forty years ago. As early as 1958, Luhn [6] wrote:

> *It is here proposed that the frequency of word occurrences in an article furnishes a useful measurement of word significance. It is further proposed that the relative position within a sentence of words having given values of significance furnishes a useful measurement for determining the significance of sentences. The significance factor of a sentence will therefore be based on a combination of these two measurements.*

Considerable research has been undertaken into the first of Luhn's criteria - term frequency - which has lead to numerous useful ideas, including representing queries and documents as term vectors and calculating the relevance of a document to a query using frequency-based similarity measures between vectors. However, until recently, scant attention has been paid to using the relative positions of terms within documents as part of the relevance calculation.

Recent experimentation by a number of groups does indicate however that measurements involving the distance between term occurrences can provide good relevance assessments, and that these relevance calculations exhibit other serendipitous properties for document retrieval that are not found with other document retrieval techniques. In particular, a very desirable property of distance-based relevance scores is their independence from collection statistics. This independence is highly desirable in applications such as routing and filtering. Furthermore, consistent document rankings can be trivially obtained across independent collections, even when those collections are distributed across the net.

In addition to assessing document relevance, we also propose that distance measures have application in identifying phrases with a greater degree of tolerance for word substitions and unexpected intervening words.

In the 1995 TREC conference [3] the University of Waterloo [2] achieved overall third-best average precision on the main ad hoc task and our own Australian National University submission [5] demonstrated that precision-recall results obtained in a merging task (in which sub-collections are processed independently) could be identical to those obtained over the full collection processed as a whole. Both groups used relevance measures based on distance between term occurrences. Even taking into account the uncertainty introduced by sampling errors and by uncontrolled variables such as query quality, it seems uncontroversial to conclude from the TREC-4 results that distance-based relevance estimation is worthy of more thorough study.

In this paper we:

1. Review some of the ways in which the concept of distance has been used in information retrieval in the past;
2. Describe and compare the recent experimental work of a number of groups (including our own) in using distance-based relevance calculations;
3. Propose a theoretical framework which encompasses past work and addresses a number of other significant questions; and
4. In the context of the framework, state a number of hypotheses to be tested in future empirical work.

2 What Is Meant By "Distance"

The term *distance* is used widely in the field of document retrieval and analysis, with different senses of the term arising depending on the type of objects and the space in which the distance between them is being measured.

The most familiar sense is that of *term-vector distance* where the objects are vectors of document or query term weights derived from term occurrence frequencies. The best-known example of use of distances of this sort is the SMART system from Cornell University [1] in which query-document similarity (the inverse of distance) is calculated by forming the inner product of the query and document vectors.

Another sense is that of *conceptual distance* where the objects are term nodes in a concept map or hierarchical thesaurus where the distance between terms is length of the (perhaps weighted) path between them. Richardson and Smeaton [8] describe experiments with this model.

A third sense, which can be called the *lexical distance*, takes terms from a query as objects and measures the distance between occurrences of these terms in a document. It is this sense with which we are principally concerned here.

Many document retrieval systems, such as freeWAIS-sf [7] and PADRE [4] support proximity search operators (based on lexical distance) as part of their query language. Typically, proximity operators allow a user to specify a complex query term $T_q =< W, D >$ as a set of words $W = [W_1, \ldots, W_n]$, perhaps ordered, and a number D in some metric (eg., characters, words, paragraphs). The query term T_q will then identify, as a separate individual matching document term T_d, any contiguous span of words in a document that contains all words in W (perhaps in order, and typically including other words) provided that the absolute length of the span (called the *absolute distance*) is less than or equal to D units. (Alternatively, the proximity operator can take D not as the absolute length of the span, but rather as the maximum permitted interval between occurrences of members of W in the span, called the *interval distance*.)

The start and end points of a span T_d are always considered to align on a member of W. Typically, query languages that support proximity operators do not specify whether the extent of the document term T_d is the shortest or longest span that satisfies the constraints of T_q, nor do they otherwise constrain the extent, as this information is irrelevant to the way they computate query-document similarity. Indeed such systems effectively treat lexical distance as a variable with only two possible values: *close* and *not close*.

Lexical distance should not be confused with query language constructs that permit users to confine the scope of term occurrences to particular tagged sections of documents. Many document retrieval systems support section scoping, wherein a term from the query is only treated as a document term if it occurs in the named section of the document. Section scoping is not a relation between term occurrences in the way that lexical distance is. However, intervening section boundaries may need to be taken into account in measuring lexical distance as will be seen later.

A related idea is that of *passage level retrieval* (Salton, Allan and Buckley [9]). Underlying this idea is the fact that large documents can often contain considerably smaller passages of text that are relevant to a given query, but documents containing these passages are ranked low by many frequency-based relevance ranking procedures because of the relatively higher concentration of irrelevant terms in the document, and the fact that the contained relevant passages are not independently visible to the ranking process. Passage level retrieval attempts to overcome this problem by making intra-document *passages* visible to the ranking process, where a passage is defined either as a contiguous group of paragraphs or sections of a document (identified through markup in the document), or as a block of words of fixed length, or as a *window* over the document which can move in fixed-length jumps, or as a non-contiguous but semantically linked thread of paragraphs through the document.

Although passage level retrieval, like section scoping, is not a relation between term occurrences, it is often motivated by the intuition that semantically related term occurrences often occur closer to each other, for example in passages or ideally in the same sentence or subsentence, than unrelated term occurrences. This intuition motivates the notion of *local text similarity* measures (Salton, Allan and Buckley [9]) as distinct from global document similarity measures. A similar motivation can be given for exploring lexical distance as a means of calculating relevance.

To illustrate the point, an article was posted to the network news three years ago (and retrieved by

an essentially boolean query) which contained widely separated occurences of the terms `Hawking`, `text` and `retrieval`. Its content bore no connection with one of the present authors and none with the field of IR. This particular false hit could have been avoided had a suitably-chosen proximity constraint been imposed. In general, however, any fixed proximity threshold is likely to cause both misses and false hits. It is our intuition that some continuous function related inversely to the lexical distance separating the term occurrences could be usefully used to estimate the probability that together they indicate relevance.

Based on intuition and the results of initial experiments, we hypothesize that the probability that a document D is relevant to a topic T:

1. Increases as the lexical distance between appropriately chosen groups of query term occurrences in a document (spans) decreases; and
2. Increases as the number of such spans found in a document increases.

3 Reported Uses of Lexical Distance in Scoring Relevance

3.1 Z-Mode - Australian National University

One of ANU's submissions (`padreZ`) in the 1995 TREC task [3] scored relevance entirely on the basis of these two heuristics. No singleton term counted anything directly toward the relevance of documents. Sets of related terms (words, phrases and regular expressions, including alternative spellings, plurals, different parts of speech and even mis-spellings) were grouped together to represent *concepts* from the retrieval topic. Next, queries were composed which located spans including at least one representative of each of a number of intersecting concepts. The inverse square root of the length of span was used to calculate the contribution to the document's relevance score.

For example, in locating documents relevant to the topic, "*What is the economic impact of recycling tires?*", three concepts were identified:

1. economic impact
2. recycling
3. tires

The text fragment

```
 ... reported huge profits to be made from recycling discarded automobile
tires. ...
```

would have contributed to the relevance score of a document containing it, because the span from `profits` to `tires` includes representatives of each concept. Applying the formula used in the `padreZ` run, the size of the contribution would be $1/\sqrt{8}$.

Only spans containing at least one representative of each concept counted towards relevance. Multiple representatives of a single concept in the absence of the other concepts counted nothing.

Some `padreZ` queries attempted to locate more than one concept intersection. The relative importance of one such intersection to the others could be specified at query-generation time in the form of a factor by which component relevance contributions were multiplied prior to final summation.

3.2 Shortest Substring Ranking - University of Waterloo, Canada

The method of scoring *extent*s (equivalent in meaning to *spans* as used here) in the GCL system used by Clarke, Cormack and Burkowski at the University of Waterloo [2] (TREC run `uwgcl1`) is remarkably similar to ours, with some differences:

1. Both methods allow overlapping spans, and both seek spans of minimum length. However, PADRE accepts a span even if it contains another span within it, provided that the two have distinct starting points, whereas GCL never accepts a span which contains another.

2. The `padreZ` and `uwgcl1` submissions also differed in the way in which relevance scores decayed with increasing span length. Both groups found that a simple reciprocal function decayed too rapidly. ANU used a square root function to slow the decay whereas Waterloo chose a threshold length A and scored 1.0 for lengths of less than A and $A/length$ for greater lengths.

3. Hawking and Thistlewaite applied a distance cutoff of 1,000 characters, after which spans were not recognised, whereas the Waterloo group do not appear to have done so.

The way in which distance-based operators was used by the two groups was also very similar. Both groups used un-scored fixed proximity operators to identify phrases and, based on an incomplete study of the Waterloo TREC queries, both groups used distance-based relevance calculation only with the intent of locating concept intersections.

4 A Theoretical Framework for Lexical Spans

Until now we have been intentionally vague (as has the literature generally) about the structure, identity and individuation of "spans".

In discussing the approaches implicit in the ANU and Waterloo strategies, it will help if we have more formal machinery for talking about spans and distance measures.

More importantly, while the experimental results look quite promising, it is desirable that there be a space of alternative hypotheses to direct future experimentation, and to provide a basis for explaining experimental results.

4.1 Definitions

Let $W = [W_1, ..., W_n]$ be a contiguous sequence of words (or word surrogates, such as stems) in a document and subquery $Q_k = \{Q_{k1}, ..., Q_{kj}\}$ be a set of query words (or word surrogates), such that $W_1 \in Q_k$, $W_n \in Q$ and $n >= 1$[1]. W is what we have been calling a *span*.

If $W_i \in W$ and $W_i \in Q_k$, then W_i is termed a *pivotal point* in W w.r.t. Q_k. We call Q'_k such that $Q'_k \subseteq Q_k$ and all members of $Q'_k \in W$, the *set of pivots* in Q_k w.r.t. W. Note that there may be more pivotal points in W than the cardinality of the largest set of pivots, as members of Q_k may have multiple instances in W. If W_i is W_1 or W_n, then it is termed a *boundary pivotal point*, otherwise it is termed an *interval pivotal point*. Note that for some W and Q_k, there may be no interval pivotal points.

For convenience, we will define an *intervening sequence* V of W w.r.t Q_k to be all the words in W, except the first occurrence of each pivot, in the same order in which they occurred in W.

A span W is termed *saturated* w.r.t. Q_k if all members of W are in Q_k. A span W is termed *complete* w.r.t. Q_k if all members of Q_k are in W. A span W is termed *a partial span of degree n* w.r.t. Q_k if all members except n of Q_k are in W. Due to the possibility of repeated pivots, a saturated and complete span may have a non-empty intervening sequence.

We define the *absolute inner span* of W to be $[W_2, ..., W_{n-1}]$, the *absolute left-anchored span* of W to be $[W_1, ..., W_{n-1}]$, and the *absolute outer span* of W to be $[W_1, ..., W_n]$.

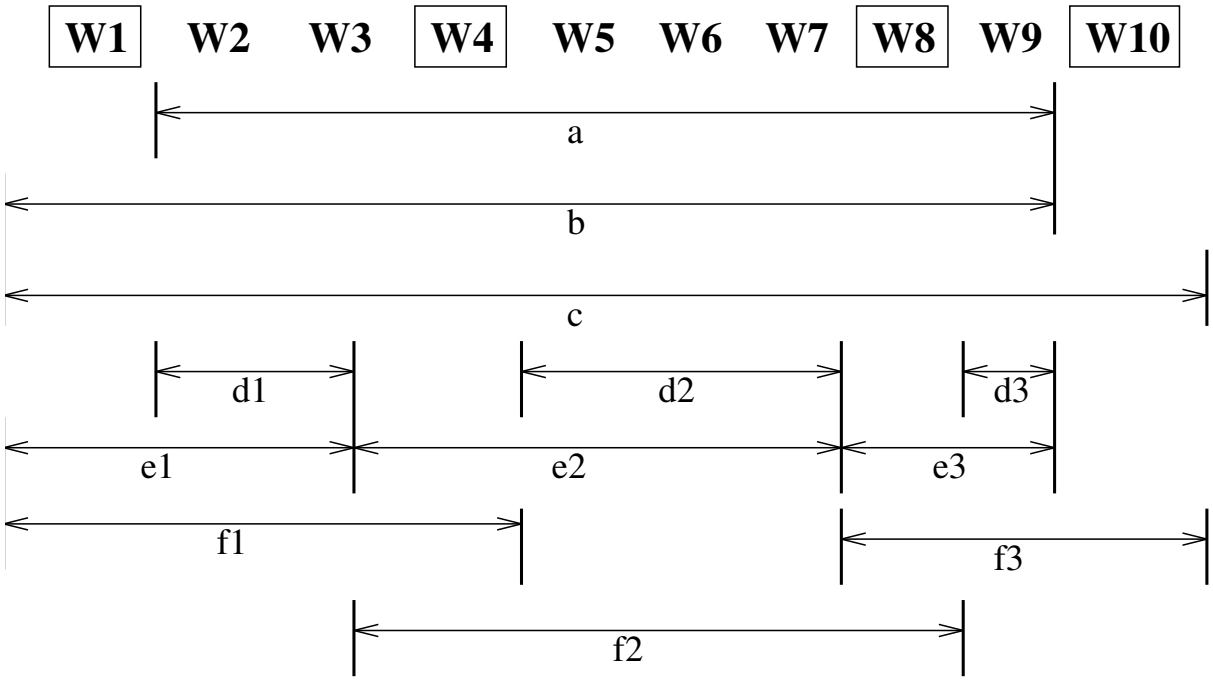
Let $P = [P_1, ..., P_m]$ be the sequence (in order of occurrence in W) of the pivotal points in W w.r.t. Q_k (so, given the definitions, P_1 will be W_1 and P_m will be W_n). Then we can define an *interval inner span* of W to be $[W_{i+1}, ..., W_j]$ such that $W_i = P_l$ and $W_{j+1} = P_{l+1}$; an *interval left-anchored span* of W to be $[W_i, ..., W_j]$ such that $W_i = P_l$ and $W_{j+1} = P_{l+1}$; and an *interval outer span* of W to be $[W_i, ..., W_j]$ such that $W_i = P_l$ and $W_j = P_{l+1}$.

These definitions are illustrated in Figure 1.

Note that if there are i pivotal points in W, then there are exactly $(i-1)$ interval spans of a particular type (inner, left-anchored or outer).

Assuming an absolute or interval span $W = [W_1, ..., W_n]$, then the length of its inner span is $(n-2)$; the length of its left-anchored span is $(n-1)$; and of its outer span is n. Provided that there are no repeated pivots, the length of the intervening sequence of a span is identical to the sum of its interval inner spans.

[1] One might have expected $n > 1$ but we wish to treat a single term as a special case of a span.



(a) is the absolute inner span of W;

(b) is the absolute left-anchored span of W;

(c) is the absolute outer span of W;

(d1) (d2) and (d3) are (all of) the interval inner spans of W;

(e1) (e2) and (e3) are (all of) the interval left-anchored spans of W;

(f1) (f2) and (f3) are (all of) the interval outer spans of W.

Figure 1: A hypothetical ten-word sequence containing four query words, illustrating the various ways in which spans may be defined.

Note that where there are no interval pivotal points, the interval (inner, left-anchored, or outer) span is just the corresponding absolute span.

4.2 Intuitions About Completeness, Overlap and Containment of Spans

In both relevance estimation and phrase identification, it seems that complete spans are of more interest than partial spans even if the latter are shorter. Neither PADRE (ANU) nor GCL (U Waterloo) in their TREC-4 versions contained automatic mechanisms for scoring partial spans. Because of the danger that a fully specified sub-query might be too restrictive and cause excessively low recall, both groups designed sequences of subqueries layered according to increasing weakness. What we hope will be a better, fully automatic method is proposed below.

Both Waterloo and ANU submissions at TREC-4 allowed overlapping spans; at this point we see no reason to exclude them.

In TREC-4, Waterloo excluded all spans which contain a complete span within them. ANU's `padreZ` submission chose the shortest complete span starting at a particular point in the text. At this point however, we are not aware of any evidence to justify excluding longer complete spans. Our model is capable of allowing containment and, until the necessary experiments can be conducted, we will continue to allow this possibility. Naturally longer spans would score lower than the complete spans they contain.

4.3 The Significance of the Cardinality of Q_k

The number of query terms required to specify a topic with a given degree of specificity varies greatly according to the topic and to the particular terms. For example, a single query term such as `autohypnosis` may convey meaning just as precisely as a sequence of terms such as `self induced hypnotic trance`.

In general, queries are likely to comprise a number of independent sub-queries $Q_1, ...Q_m$ in which the significance of each Q_k is not determined by its cardinality but rather by a significance weight S_k assigned during query generation by a process which does not concern us here.

Consequently, will attempt to devise relevance functions such that all saturated complete spans will score equally, prior to adjustment by S_k.

4.3.1 The Special Case of Singleton Sub-queries

The example of autohypnosis shows a realistic case in which the cardinality of a subquery may be one. Do such singletons require a completely separate treatment or can they be treated as a special case of the distance-based framework?

We propose that occurrences within a document of instances of a one-word sub-query should be treated as saturated complete spans. Naturally, in these cases, distance is always as short as it can possibly be. Furthermore, successful query generation in this model will almost certainly require that singleton sub-queries be restricted to those with very high information content or that they be subject to very low significance weighting.

4.4 Measuring Span Length

In section 4.3 we argued that saturated complete spans should score equally. When measuring saturated complete spans, lengths calculated using absolute outer, absolute left-anchored, absolute inner, interval outer and interval left-anchored spans all increase with the cardinality of Q_k. We therefore rule out these methods of defining length, leaving two possibilities:

1. the length of the intervening sequence, and
2. the sum of the lengths of the interval inner spans.

In the absence of repeated pivots, both these methods will yield the same result. Our intuition is that repeated internal pivots do not effectively increase the distance between the boundary pivotal points. For example, when looking for

$$Q_k = \{\texttt{Australian}, \texttt{communist}|\texttt{marxist}, \texttt{party}\}$$

we consider that

```
... Australian communist party ...
```

and

```
... Australian marxist, communist party ...
```

should be treated as having equal lengths.

Accordingly, we choose to define *span length* as the sum of the interval inner spans.

4.4.1 Treatment of Repeated Pivots

Many complete spans contain multiple occurrences of a term from the sub-query, particularly when the term is in fact a long list of alternative words. It is arguable that such repetitions increase the indication of relevance. Already, under our definition of span length, the length of a span containing repetitions of pivots is less than it would have been had the repetitions been replaced by non-pivot words. However, it is possible that this adjustment does not adequately reflect the additional contribution to relevance.

Further adjustment could be applied by multiplying the relevance contribution of a span by a factor derived somehow by the number of repetitions within the span. Initial experiments in this direction have not shown a benefit but more extensive experimentation is needed in order to be sure.

4.4.2 Non-Linearities in Span Length

Intuition suggests strongly that a span which crosses a sentence, paragraph or section boundary is less likely to indicate relevance than a similar span which does not. Representing this in our model may be achieved in two ways:

1. Disallowing spans which cross major section boundaries, and/or
2. Increasing the effective length of the span by an increment chosen according to the type of section boundary crossed.

Empirical study is required to determine whether either or both of these modifications are effective and what sized increments are most effective. Further study is necessary to determine whether "noise words" should be ignored when measuring length.

4.4.3 Limiting Span Length

It is clear that spurious indications of relevance will arise if spans are allowed to cross document boundaries within a collection and it is also clear that beyond some span length L_{max}, the chance that the pivots are connected is negligible. Implementations of distance-based relevance scoring may ignore spans longer than L_{max} for efficiency or other practical reasons.

4.4.4 Significance of the Distribution of Internal Pivots

At present it is not known whether the distribution of internal pivots in a span W, is related to relevance or not. In the example spans below, where the Ps are pivotal points and the Ns are not, is E1 more indicative of relevance than E2, or vice versa, or is there no difference?

```
(E1)  P1 N N N N P2 N N N N P3

(E2)  P1 N P2 N N N N N N N P3
```

Thus far, our proposed measure of span length is blind to different distributions. We believe that an alternative approach based on separate relevance calculations for each interval could lead to different relevance scores when subsets of the pivots are clumped together (for example E2) compared to the case where all of them are uniformly spread (for example E3). However, devising a suitable scoring method along these lines is significantly complicated by repetitions and must remain the subject of future work.

4.4.5 Scoring of Partial Spans

Let us imagine a sub-query Q_k comprising j query terms and a complete span W with respect to it. Now let us imagine a nearly identical partial span W' in which one of the members of Q_k, say Q_{ki} is not present. Its absence may be because:

1. W' is not indicative of relevance, or
2. W' is less strongly indicative of relevance, or
3. Q_{ki} is actually present as a synonym or synonymous phrase which was not anticipated during query construction.

It is impossible to distinguish which of these cases applies. Therefore, we may conclude that the presence of W' in a document boosts the probability that it is relevant but not by as much as W would have.

Intuitively, as stated earlier, complete spans should score more highly than partial spans. The more pivots missing, the lower should be the score. In general, we hypothesize that a saturated partial span of degree n should be given a relevance score equivalent to that which would have been assigned to a partial span of degree $(n+1)$ of length $L_{max}+1$. In other words, for a given W and Q_k, the highest scoring partial span of degree 1 will score lower than the lowest scoring span of degree 0 (complete), but only just. Similarly, the best possible one of degree 2 will score lower than the worst of degree 1, but only just.

To avoid distorting this constructive hierarchy of partial spans, partial spans which are contained within another span of lower degree are ignored.

4.4.6 Significance of Pivot Ordering

The English language provides many alternate forms in which the order of words is reversed. For example, `brass doorknobs` and `doorknobs of brass`. Since the relevance implied by such alternates is often the same it is useful to consider order-independent spans. However, in less-frequent cases, it may be useful to define ordered spans in which the order of the pivotal points in W must match a specified ordering of the words in the sub-query Q_k. For example, `concrete proposal` is not equivalent to a `proposal for concrete`.

4.5 A Family of Distance-Based Relevance Formulae

The relevance score contribution RC for a document D against a sub-query Q_k derived from a complete span W is proposed to be:

$$RC(D, W, Q_k) = \frac{C \times S_k}{F((Length(W)+1) \times (L_{max}+1)^n)} \quad (1)$$

where C is a constant, usually unity but possibly related to the number of repeated pivotal points within complete spans (as mentioned above), S_k is the significance assigned at query generation time to Q_k and n indicates the degree of a partial span (zero for a complete span).

The function F may be the identity function but could be another function designed to alter the rate at which the relevance contribution decays with length. One is added to the length before the function is applied to avoid potential difficulties of division by zero.

For each document D the contributions for each W and Q_k are summed to give an estimate of the relevance of D to the overall query Q.

4.6 Applications of Distance-Based Relevance Formulae

We hypothesize that the relevance of documents may be scored using formula 1, using sub-queries chosen to represent concept intersections as described in our TREC-4 paper [5], and choosing a function F such as square-root rather than the identity function in order to slow down the decay of the relevance function

with increasing span length. However, formula 1 does not take into account the fact that the members of sub-query Q_k may be "loose phrases" (phrases occurring with intervening words).

We further hypothesize that lexical distance can be used to assign a probability that a particular span W in fact represents the phrase specified by a particular Q_k. The formula used is very closely related to formula 1:

$$p = \frac{1}{F((Length(W)+1) \times (L_{max}+1)^n)} \tag{2}$$

In the case of loose phrases, the probability should fall off quickly as the number of intervening words increases. Accordingly, F may well be the identity function or even a power function and L_{max} may be reduced.

Now, formula 1 must be modified to take into account the probabilities that the pivotal points in W are actually occurrences of their corresponding pivot. For literal terms, the probability is unity; for phrases, the probability is calculated using formula 2. The probability P that all pivots are in fact present is the product of the individual probabilities p. If a pivot is repeated, it seems intuitive that the probability that the pivot is present is the maximum of the probabilities for the pivotal points for that pivot.

Finally, we have:

$$RC(D, W, Q_k) = \frac{CS_k P}{F((Length(W)+1) \times (L_{max}+1)^n)} \tag{3}$$

5 Conclusions and Future Work

We offer the model described above as a formalism for describing and manipulating spans. As has been seen, it encompasses overlapping, component, partial and singleton spans and allows for the use of spans in both relevance calculation and identification of loose phrases. Further extension may be needed to determine whether the distribution of internal pivots is important and what significance should be attached to repeated internal pivots.

At present, we have not yet fully implemented the model. There is a pressing need to empirically test the hypotheses listed above, to determine optimal choices of constants and decay functions and to confirm that top-ranking precision-recall performance can be reliably obtained using distance-based measures alone. Specifically, further work is needed to determine:

1. Whether spans containg other complete spans should be included in relevance assessment;
2. Whether the length adjustment for repeated pivotal points implied by span length as defined, appropriately reflects the additional relevance contribution (if any);
3. How span length should be adjusted to take into account sentence, paragraph, section and other boundaries;
4. Whether "noise words" should be counted when measuring span length;
5. What settings should be used for L_{max} in relevance calculation and loose phrase finding; and
6. What are the best choices of F in relevance calculation and loose phrase finding.

Perhaps the most important experimental task outstanding is to confirm that document retrieval based on this model is capable of achieving top-flight precision-recall performance. Without this, the advantage of distance-based methods, that of independence of collection statistics, is reduced.

Acknowledgments

Thanks to Paul Kantor of Rutgers University for stressing the importance of handling partial spans and suggesting possible ways in which this could be done.

References

[1] Chris Buckley, Gerard Salton, and James Allan. Automatic retrieval with locality information using SMART. In D. K. Harman, editor, *Proceedings of TREC-1*, pages 59–72, Gaithersburg MD, November 1992. NIST special publication 500-207.

[2] C. L. A. Clarke, G. V. Cormack, and F. J. Burkowski. Shortest substring ranking (MultiText experiments for TREC-4). In Harman [3], pages 295–304. NIST special publication 500-236.

[3] D. K. Harman, editor. *Proceedings of TREC-4*, Gaithersburg MD, November 1995. NIST special publication 500-236.

[4] David Hawking and Peter Bailey. *PADRE v. 2.4 User Manual.* Department of Computer Science, The Australian National University, Canberra, `/http://cap.anu.edu.au/cap/projects/text\_retrieval/`, 1996.

[5] David Hawking and Paul Thistlewaite. Proximity operators - so near and yet so far. In Harman [3], pages 131–143. NIST special publication 500-236.

[6] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2:159–168, 1958.

[7] Ulrich Pfeifer. *freeWAIS-sf*, 1995. `http://ls6-www.informatik.uni-dortmund/freeWAIS-sf/`.

[8] R. Richardson and A. F. Smeaton. Using WordNet in a knowledge-based approach to information retrieval. Technical Report CA-0395, Dublin City University School of Computer Applications, `http://www.compapp.dcu.ie/CA\_Working\_Papers/WP\_home.html`, 1995.

[9] Gerard Salton, James Allan, and Chris Buckley. Approaches to passage retrieval in full text information systems. Technical Report TR93-1334, Department of Computer Science, Cornell University, Ithaca NY, 1993.