# An Exploratory Analysis of Phrases in Text Retrieval

Jeremy Pickens and Bruce Croft
Department of Computer Science
University of Massachusetts
Amherst, MA 01002
{*jeremy, croft*}*@cs.umass.edu*

November 5, 1999

## Abstract

*Phrases are used in both commercial and experimental search engines. Despite the large amount of work in the area results remain mixed. It is still not clear whether phrases can be used to improve retrieval effectiveness. In this paper, we examine phrases and their properties independently of any specific retrieval approach. We explore phrase usage in text corpora and relevance patterns related to phrase usage. The result is not only a better understanding of phrases, but a better method by which phrases and phrase techniques may be evaluated. With this method we can directly determine the value of various phrase formulations for information retrieval.*

## 1 Introduction

Phrases in information retrieval have been used to aid effectiveness. They have gone by many names: multiword features, phrase terms, nominal compounds and so on. Despite their many implementations, retrieval experiments with phrases are often inconclusive. In some experiments phrases yield small improvements in retrieval, but most experiments show inconsistent results [CTL91, Fag87]. What is the source of this inconsistency?

The primary method for utilizing phrases in information retrieval has been to: (1) Create a technique that identifies phrases within queries[1], (2) Create a technique that identifies these query phrases within documents[2], (3) Create a retrieval algorithm that appropriately retrieves both document words and document phrases, and (4) Run retrieval experiments using this retrieval algorithm. If results improve, the entire phrase scheme (steps 1-3) is a good scheme; otherwise it is not.

The problem with this method is that broader phrase patterns become hidden by the retrieval experiment. When the entire measure of a phrase scheme is reduced to one statistic, that of retrieval effectiveness, one looses

[0]This material is based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623,and also supported in part by United States Patent and Trademark Office and Defense Advanced Research Projects Agency/ITO under ARPA order number D468, issued by ESC/AXS contract number F19628-95-C-0235. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor(s).

[1]Phrases identified within a query will be known as *query phrases*

[2]Phrases identified within a document will be known as *document phrases*

sight of important patterns. If one is trying to determine why a particular phrase scheme is bad, it is difficult to differentiate between a bad query phrase identification technique, document phrase indentification technique, or phrase retrieval algorithm. For example, both phrase query and document identification techniques might be ideal, but if the retrieval algorithm is poorly suited to phrases, the entire phrase scheme could produce bad retrieval results.

In this this paper we distance ourselves from direct retrieval experimentation and instead let the phrase and relevance patterns speak for themselves. We examine two document phrase identification techniques: In sections 4 and 5 we discuss phrase proximity[3] and in section 6 we discuss phrase anaphora[4]. We also examine a retrieval technique, inverse document frequency: Section 7 is an exploration of *idf* and how phrases behave with this weighting. Section 8 combines the structural technique of proximity with *idf* weighting.

A tool known as *weight of evidence* [Gre98] is the lens through which we examine these properties, through which we unearth interesting patterns. The value and contribution of this paper lie not only in these newly discovered patterns, but also in the method used to discover these patterns. Weight of evidence is an effective tool for measuring the value of various phrase formulations in information retrieval.

# 2 Background and Related Work

## 2.1 Definitions

In order to use phrases in retrieval experiments we must know whether phrases are found in our queries, and whether these query phrases are also found in our documents. We must also have some way of combining the query phrases and words with the document phrases and words.

We define *phrase identification* as the process by which words in a query are selected and recognized as phrases[5]. Techniques for identification include syntactic, statistical and manual [Fag87, Chu88, GC90, CTL91]. Syntactic techniques often tag words by their parts-of-speech and extract contiguous sequences of nouns. Statistical methods rely on frequency counts and measures of subterm relatedness such as co-occurrence. Much work has been done on indentifying phrases in text. The goal of this paper is not to add to this body of knowledge, but to make use of it. Before we can use phrases within a retrieval experiment or within an exploratory analysis we must identify the phrases under consideration.

We define *phrase structuring* as the process by which identified query phrases are recognized again within documents. Many phrase identification techniques only look at contiguous sequences of words; if two words are never adjacent, then they will never be identified as a phrase. Structuring techniques, on the other hand, are concerned with how identified phrases are actually used. For example, a phrase might be two contiguous words when it is identified, but twenty words apart when it is used within a document text. Structuring techniques such as proximity deal with these issues.

We define *phrase weighting* as the process by which phrases are combined with words: the retrieval algorithm. We must assign a score for every recognized phrase in a document. Techniques for phrase *weighting* include

[3] the lexical distance at which two subterms are found in a document
[4] an anaphor is a word or a phrase which references the same concept as another word or phrase
[5] The constituent words of a phrase will be known as phrase *subterms*

various formulations of *tf*[6] and *idf*[7], and higher scores for closer proximities [HT96].

Phrase identification, structuring and weighting techniques are heirarchically dependent in a retrieval experiment. A structuring technique depends on the identification technique; no query phrases that have not first been identified will be recognized in documents. A weighting technique also depends on the structuring technique; no document phrases that have not first been recognized can be weighted. The distintions drawn between these different techniques are not absolute boundaries. Some work on phrases does not differentiate between how phrases are identified and how they are structured. Other work on phrases [HT96] does not differentiate between how phrases are structured and how they are weighted. Nevertheless, analysis is easier if these distinctions are made.

The phrases identified in this paper are initially constructed using a Markov-based statistical phrase extractor, tempered by a number of heuristics which remove low quality phrases [FC99]. Although this is not a syntactic phrase extractor, the resulting phrases are mostly noun phrases. We will examine various structuring and weighting issues within the context of these identified phrases. There are numerous other identification, structuring and weighting methods which we do not explore, although one is certainly not limited by the possibilities mentioned here.

## 2.2 Word and Phrase Differences

The most apparent difference between phrases and words[8] is their structure. Even if methodological particulars differ, *identification* and *weighting* techniques exist for both words and phrases. But words, unlike phrases, have no need of *structuring* techniques; their existence as identified words provides their entire structure. Phrases, on the other hand, do not have this same intrinisic, unchanging structural form. It is not clear what phrase structural form should be and whether or not phrase weighting should change for different strutural forms.

[CTL91] provide an in-depth analysis of various phrase structuring issues. Much of the work in this paper uses their analysis as a framework. They also tackle the problem of weighting these various structural formulations for their retrieval model, the inference network [CCH92]. [GC90] show that phrases with two subterms are the most useful type, so we restrict ourselves to two-word phrases.

Hawking [HT96] explores phrase proximity. Not only does he treat this as a phrase structuring problem, but the manner in which he captures structure is the same manner in which he weights phrases. The two phrase subterms are weighted proportional to their lexical distance. [SB88] explore various weighting approaches for words, but conclude that judicious weighting of words is preferable to any weighting (or usage) of phrases. We examine the *idf* weighting approach for various phrase structures and compare it to the same weighting approach for words.

## 2.3 Exploratory Data Analysis

This paper examines phrase structuring and weighting issues by adopting an exploratory data analysis [HD79] approach similar to [Gre98, Gre99a, Gre99b]. We avoid relying on data summaries which possibly mask

[6] Term frequency measures the frequency of occurrence within one document text [Luh57] [SB88].
[7] Inverse document frequency measures the (inverse) frequency of occurrence within a document collection [Jon72] [SB88].
[8] Throughout this paper, *term* and *word* and *single term* are synonyms

important features of the data under examination. Visualization of the entire set of data allows patterns that otherwise might have remained hidden to be revealed. Thus, not only can one verify that expected trends do indeed exist, but one may also encounter unexpected trends, casting light on some aspect of the problem that might otherwise have gone unnoticed.

For example, when comparing retrieval experiments, entire recall-precision curves are often more useful than the single measure of precision averaged over a number of recall points [Kee92]. Average precision often hides important trends in the recall-precision data, such as which experiments are more precise at lower recall points. Similarly, recall-precision curves often hide important trends in phrase data because we cannot differentiate between various phrase identification, structuring or weighting techniques. Thus, we partition our exploratory analysis of phrases into these three areas and visually examine relevance trends within data gathered from each area.

## 2.4 Weight of Evidence

The primary tool in our exploratory data analysis will be weight of evidence[9]. [Goo50] defines of weight of evidence as the evidence $e$ in favor of a hypothesis $h$, where $O$ is defined as *odds*:

$$woe(h:e) = \log \frac{O(h|e)}{O(h)} \tag{1}$$

When weight of evidence is applied to information retrieval, evidence is extracted from both a query and a document and relevance judgements are used to pair the queries and documents together. Therefore, to calculate the odds in equation 1, one need only count the number of documents, $\#d$, in which the hypothesis is true, $h(d)$, and divide that by the number of documents in which the hypothesis is not true, $\overline{h}(d)$, taking into account the evidence, $(e \in d)$, provided by the document. Thus:

$$\log \frac{O(h|e)}{O(h)} = \log \frac{\frac{\{\#d|h(d)\wedge(e\in d)\}}{\{\#d|\overline{h}(d)\wedge(e\in d)\}}}{\frac{\{\#d|h(d)\}}{\{\#d|\overline{h}(d)\}}} \tag{2}$$

Often when weight of evidence is used it is plotted against another measure. We obtain a smoother view of weight of evidence by binning the data points, averaging along the $woe$ axis at the same time we average along the other axis. However, $woe$ often takes on positive or negative infinite values, making averaging difficult. Thus, instead of binning on the actual $woe$ data points, we bin on the documents counts that produced each data point (see equation 2). Finally, $woe(h:e)$ is again calculated using these binned document counts [Gre98]. In this manner, the effects of infinite values are incorporated into normal data points.

## 2.5 Co-occurrence

One additional tool in this paper is a co-occurrence measure similar to the expected mutual information measure [CX95]. It is given by:

$$cooc(a,b) = \frac{n_{ab}}{n_a + n_b} \tag{3}$$

[9] Weight of evidence will sometimes be abbreviated as $woe$

Our initial experiments show that a window size of 100 is fairly robust for this statistic. Therefore, $n_a$ and $n_b$ are the number of occurrences of $a$ and $b$ while $n_{ab}$ is the number of co-occurrences of $a$ and $b$, all within the many hundred-word windows into which the collection is partitioned. An occurrence or co-occurence is only counted once for each window in which it appears, even if multiple instances are found within the same window.

The more $a$ and $b$ occur together, relative to the number of times they occur separately, the higher the co-occurrence score is going to be. For example, $a$ and $b$ might each only occur once, but if they also co-occur, then $cooc(a,b) = \frac{1}{1+1} = 0.5$. If, on the other hand, $a$ and $b$ co-occur a thousand times, but each occur in the collection five thousand times, then $cooc(a,b) = \frac{1000}{5000+5000} = 0.1$. Actual counts are not as important as the proportion of co-occurrences to total occurrences.

In this paper $a$ and $b$ will be defined a number of different ways. When exploring phrase proximity, they are defined as phrase subterms. When exploring phrase anaphora, they can either be subterms or the whole phrase. Regardless, the co-occurrence measure is the same, no matter what $a$ and $b$ are.

### 2.6 Topics and Collections

The phrases used in this paper are extracted from TREC topics [Har95]. In our initial exploratory analyses, phrases are extracted from the title and description fields of TREC topics 301-350. 35 phrases are identified in these topics, and relevance judgements from TREC volumes 4 and 5 are used to pair each topic with the documents in this collection. This paper refers to these data points as the *small EDA set*.

When more data points are necessary to further solidify an observed pattern, phrases are extracted from all fields of TREC topics 1-350 (seven sets). TREC volumes 1-5 are the document collections on which nine different sets of relevance judgements have been done. Seven topic sets and nine relevance judgement sets means that some topics are used twice. For example, topics 51-100 were judged once on TREC volumes 1 and 2 and once on TREC volume 3. Thus, this dataset contains 4641 unique phrases, where a phrase is not only extracted from a topic, but also paired with a document collection through a relevance judgement. This paper refers to these 4641 data points as the *large EDA set*.

## 3 Example

Phrases often have inconsistent structural behavior. In particular, the lexical distance, or proximity, at which the two constituent terms of an identified phrase are found varies not only from phrase to phrase on a single document collection, but from document collection to document collection for a single phrase.

What does this structural behavior look like? More importantly, what do patterns of relevance and non-relevance look like, for this structural behavior? The following exploratory data analysis answers this question for the phrase, *fiber optic*. Of course, because this is only one example, with no statistical significance, this indicative of a problem we see rather than an actual result.

The data for this exploration comes from two different TREC topics judged against three different collections. Our first topic, TREC topic 97, has the title `Fiber Optics Applications`. Our second topic, TREC topic 320, has the title `Undersea Fiber Optic Cable`. For topic 97, relevance judgements have been made

on two different document collections: TREC volumes 1 and 2 and TREC volume 3. For topic 320, the TREC volumes 4 and 5 collection were used.

In topic 97, the phrase *fiber optic* is the first concept mentioned in the <concept> field. Almost every other field in the topic mentions *fiber optic* in some form. In topic 320, the phrase *fiber optic* is also found in both the <desc> and <narr> fields. Clearly, both topics are "about" *fiber optic*. Or if one prefers, *fiber optic* is one of the most important phrases in both topics. We emphasize this point to show that inconsistencies appear even among "key" or "core" phrases, rather than among phrases which are not as germane to the topic(s) in question.

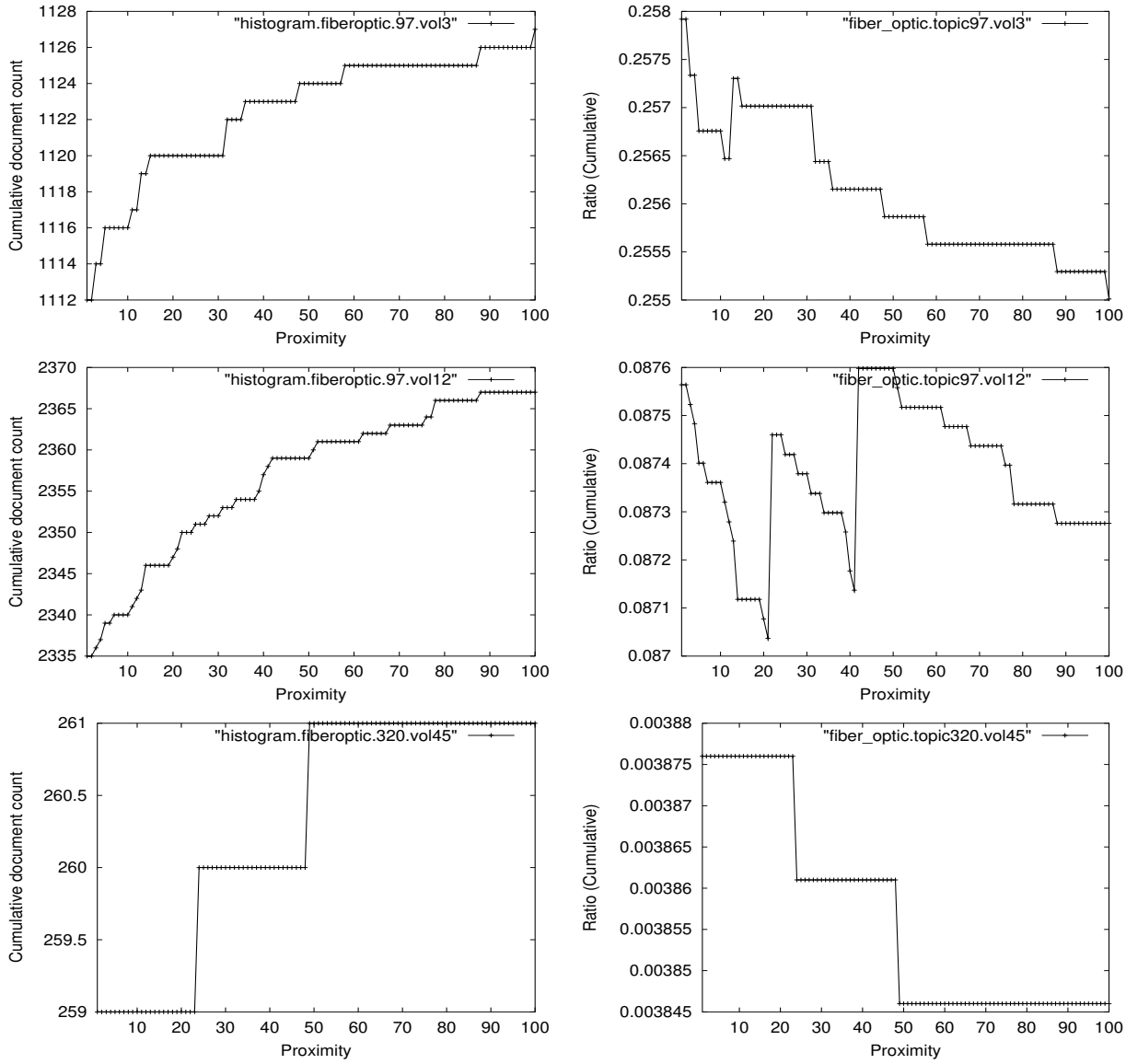


Figure 1: For the phrase "fiber optic", from proximities 1 to 100. *Left column:* Document count for proximity $\leq x$. *Right column:* Ratio of relevant to non-relevant documents for proximity $\leq x$. *Top row:* TREC topic 97, judged on volumes 1 and 2. *Middle row:* TREC topic 97, judged on volume 3. *Bottom row:* TREC topic 320, judged on volumes 4 and 5.

Data is gathered separately for each collection. The number of documents that contain both *fiber* and *optic* are counted and the minimum proximity for each document is calculated. A histogram of the number of documents at each proximity is created. Furthermore, relevance judgements between the topics and the collections are used to partition our histogram counts into a ratio histogram of relevant-to-non-relevant documents at each proximity.

**Summary:** The results of the EDA are shown in Figure 1. These graphs are cumulative; proximity $= x$ includes all documents with minimum proximities less than or equal to $x$. We observe the following trends:

- The majority of occurrences of *fiber optic* are adjacent occurrences. The slope of the document count histograms also indicates that fewer and fewer additional occurrences are accumulated as proximity widens.
- Relevant documents do not just occur at low proximities. In fact, in one collection, there are actually more relevant than non-relevant documents at a lexical distance $\leq 50$ than $\leq 1$.
- Large drops or increases in relevance occur, in all three examples, at proximities often associated with word, sentence, and paragraph-level boundaries [Cal94].

Again, this is only one example, rather than an entire analysis, but it unearthed a number of patterns that we otherwise would not have seen if we had only done retrieval experiments. The rest of this paper adopts this approach.

# 4 Phrase Structuring: Proximity

Our example in section 3 leads us to believe that smaller phrase subterm proximity is better than larger proximity, based on our observations for the phrase *fiber optic*. Now we ask the question, for all phrases, if adjacency[10] is better than Boolean[11], as a structuring technique for phrases. This is not a new concept in information retrieval. Our approach, however, is.

Traditionally, structures which take advantage of the idea of proximity, or lexical distance, have done so in one of the three following manners:

- Boolean: Recognize the presence of a phrase if and only if the two phrase subterms are present in the document, at any lexical distance.
- Adjacent: Recognize the presence of a phrase if and only if the two phrase subterms are present in the document, and they are adjacent to each other; proximity equals 1.
- Somewhere in the middle: Recognize the presence of a phrase if and only if the two phrase subterms are present in the document, at some lexical distance $x$ or less, where $1 < x < \textit{document length}$.

[10] Phrases whose constituent subterms appear in adjacent form will be known as *adjacent phrases*. Documents in which an adjacent phrase is found will be known as *adjacent documents*

[11] Phrases whose constituent subterms appear in Boolean form will be known as *Boolean phrases*. Documents in which a Boolean phrase is found will be known as *Boolean documents*. Note that all adjacent documents are also Boolean documents, but not vice versa

Recall that the weight of evidence $woe(h : e)$ is the weight in favor of the hypothesis $h$, provided by the evidence $e$. Our hypothesis now is that of document relevance: $h(d) = rel(d)$. We start with the evidence, $e_1 = t_1 \wedge t_2$, that both of our phrase subterms are already in the document. What weight in favor of the relevance hypothesis does the additional evidence, $e_2 = t_1 t_2$, of phrase subterm adjacency provide? In other words, if both phrase subterms already appear anywhere in the document, adjacent or not, does knowing that they are adjacent give better indication of relevance? This is expressed as $woe(h : e_2|e_1)$:

$$woe(h : e_2|e_1) = \log \frac{\frac{\{\#d|rel(d) \wedge (t_1 t_2 \in d) \wedge (t_1 \wedge t_2 \in d)\}}{\{\#d|\overline{rel}(d) \wedge (t_1 t_2 \in d) \wedge (t_1 \wedge t_2 \in d)\}}}{\frac{\{\#d|rel(d) \wedge (t_1 \wedge t_2 \in d)\}}{\{\#d|\overline{rel}(d) \wedge (t_1 \wedge t_2 \in d)\}}} \tag{4}$$

**Results** The following table is constructed from the small EDA set, sorted by ascending weight of evidence. The higher the weight of evidence value, the better adjacency is over Boolean. The identified phrase that is being compared in both adjacent and Boolean form is also listed, along with the TREC topic number in which this phrase was originally identified.

| WOE | Phrase | Topic # |
|---|---|---|
| -$\infty$ | chemical reactions | 349 |
| -$\infty$ | drug treatment | 339 |
| -$\infty$ | economic factors | 345 |
| -$\infty$ | u.s. efforts | 347 |
| -$\infty$ | young people | 309 |
| -0.320009 | detailed description | 307 |
| -0.116783 | alzheimer's disease | 339 |
| -0.051049 | air pollution | 329 |
| -0.003636 | trade secrets | 311 |
| -0.000149 | united states | 318 |
| 0.0082728 | endangered species | 304 |
| 0.032556 | spotted owl | 347 |
| 0.0375164 | fiber optic | 320 |
| 0.103242 | cold war | 342 |
| 0.120605 | third world | 321 |
| 0.177065 | mexico city | 329 |
| 0.194261 | organized crime | 301 |
| 0.226074 | international relations | 324 |
| 0.239375 | economic impact | 345 |
| 0.244002 | living conditions | 318 |
| 0.244038 | computer terminals | 350 |
| 0.266284 | criminal activity | 301 |
| 0.271946 | adverse effects | 338 |
| 0.367057 | school systems | 346 |
| 0.367349 | income tax | 332 |
| 0.37449 | tobacco companies | 345 |
| 0.405927 | political power | 321 |
| 0.543098 | world bank | 331 |
| 0.601034 | export controls | 334 |
| 0.646323 | international flights | 341 |
| 0.833559 | e mail | 344 |
| 0.951343 | trade shows | 311 |
| 1.24294 | public places | 348 |
| 1.41229 | daily basis | 350 |
| $\infty$ | negative effect | 309 |

Figure 2: Results of equation 4, $woe(adjacency|Boolean)$, on the small EDA set

**Summary** As we suspected, and much more often than not, adjacency is positive evidence in favor of relevance. There are proportionally more relevant documents within an adjacent proximity than within a Boolean proximity. However, it is not an absolute rule. Almost a third of the time adjacency has negative weight. In fact, there are five phrases where weight of evidence is $-\infty$. We cannot simply discount these examples as anomalies. These phrases have no relevant documents with the adjacent form and at least one relevant document with the Boolean form. We lose all of these relevant documents, no matter how few or many, by restricting the proximity to adjacency.

# 5 Proximity and Co-occurrence

Can we tell, a priori, which phrases are better candidates for adjacency and which are not? One common measure for phrase adjacency is phrase subterm co-occurrence [CTL91]. Intuition says that increasingly higher co-occurrence scores between phrase subterms means the adjacent phrase form will be increasingly better for information retrieval. Is this intuition correct?

Using the phrases from the large EDA set, we determine $woe(h : t_1 t_2 \mid t_1 \ \wedge \ t_2)$ for each phrase and plot these weights against co-occurrence of the two phrase subterms (see equation 4).

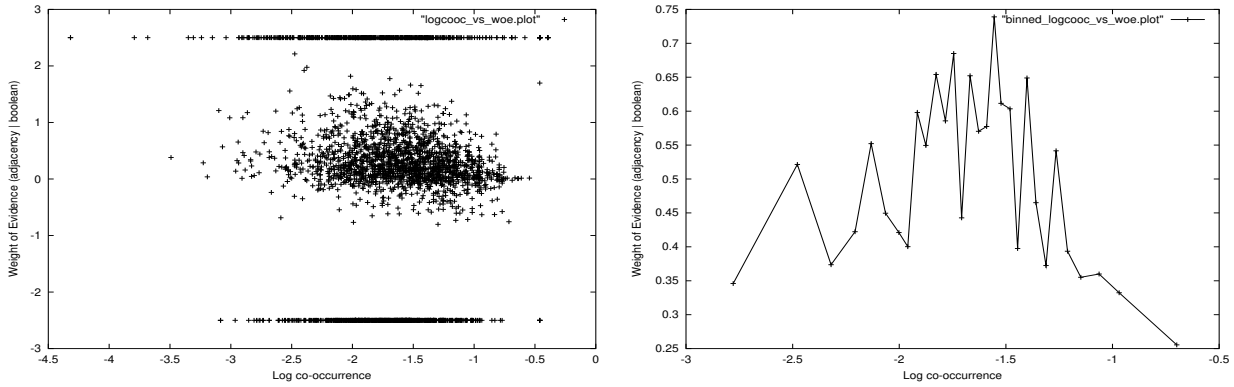


Figure 3: Phrases taken from the large EDA set. Log co-occurrence of phrase subterms is plotted against the weight of evidence of adjacency given Boolean. Note that $+\infty$ and $-\infty$ are plotted as +2.5 and -2.5.

**Results** Figure 3 contains the resulting plots, binned and unbinned. Of our original 4641 phrases, 1628 phrases extracted from the topics were not found in the various collections associated with those topics. We discard these phrases. Additionally, 549 phrases that were found in the collections had no relevant documents associated with them at any proximity. These phrases appear as infinite evidence in our unbinned data. We ignore them for now, but do take them into account when binning. Of the remaining phrases, adjacency was positive weight of evidence in favor of relevance in 1412 phrases while adjacency was negative weight (including $-\infty$) in 1052 phrases.

Not only are there significantly more positive than negative instances, the magnitudes of the positive data points are larger than the negative. This could be misleading, however, because of the large number of positive and negative infinite values. We bin the data, taking into account all document counts, and nevertheless see that at all co-occurrence scores, adjacency is positive evidence in favor of relevance.

Once again, the adjacency structural form is a useful form. We expected this. What bears more discussion, however, is the shape of the curve. It appears to suggest that phrases with high and low co-occurrence scores are not as valuable, structured as adjacent phrases, as are those with mid-range co-occurrence scores. Our original intuition was misleading. Further discussion is in order.[12]

[12] Figures 4 through 8 support this discussion. These plots were created using the same large EDA set as our original experiment. The raw data was generated, but in the interest of space we only show the binned versions. Also notice that we often plot along the log scale to provide a clearer view of all data points.

**Probability of Occurrence in Relevant and Non-Relevant Documents** The first possible explanation for this pattern is that the probability of occurrence in either relevant or non-relevant documents (or both) is different for adjacent and Boolean phrases. A difference in these probabilities could account for variation in weight of evidence. For example, if high-probability Boolean phrases have a much lower probability of occurrence in the relevant documents than adjacent phrases at the same probability of occurrence, all else equal, then we know that adjacent phrases are better phrases simply by virtue of their likelihood of appearance in relevant documents.

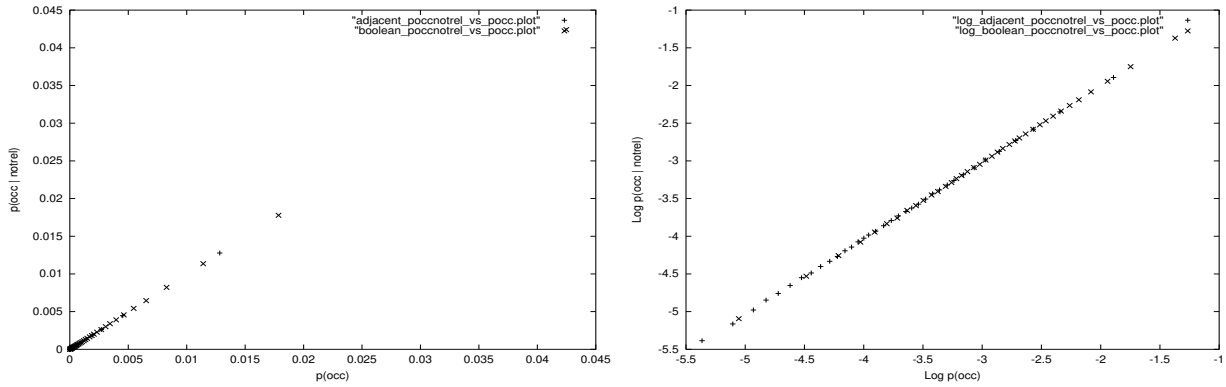


Figure 4: $p(occ|\overline{rel})$ as a function of $p(occ)$ for both adjacent and Boolean phrases *Right:* Log scale

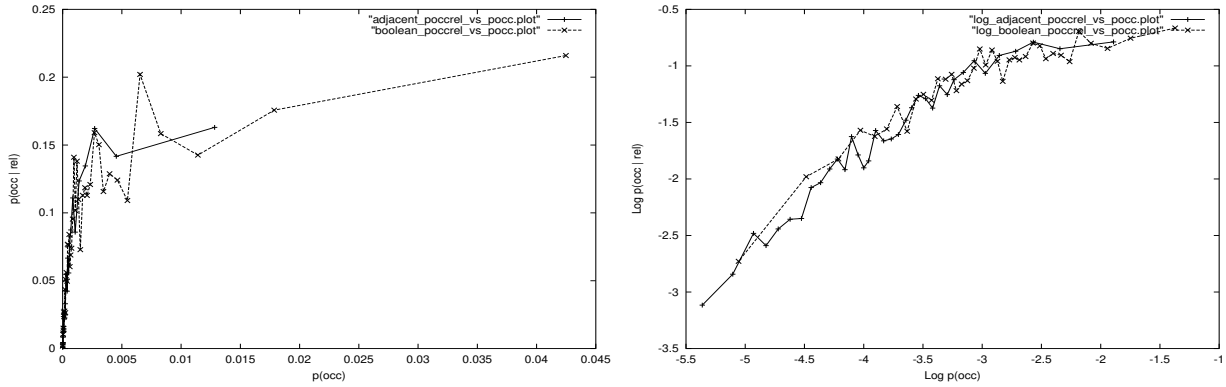


Figure 5: $p(occ|rel)$ as a function of $p(occ)$ for both adjacent and Boolean phrases. *Right:* Log scale

Figure 4 plots both adjacent and Boolean phrases. As the probability of occurrence in the collection, $p(occ)$, increases, the probability of occurrence in the set of non-relevant documents, $p(occ|\overline{rel})$, also increases. Not only is this increase linear, but the slope and position is the same for both adjacent and Boolean phrases. If the reader cannot distinguish between the two curves, this is because of the tight overlap.

Figure 5 also plots both adjacent and Boolean phrases. As the probability of occurrence in the collection, $p(occ)$, increases, the probability of occurrence in the set of relevant documents, $p(occ|rel)$, also increases. There is not the same perfect overlap betweeen the adjacent and Boolean phrases we see in the non-relevant document plots. Nevertheless, the plots appear very similar.

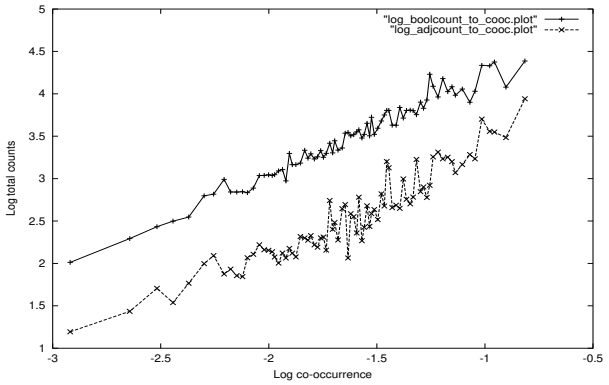


Figure 6: Boolean document count (*top curve*) and adjacent document count (*bottom curve*) as functions of co-occurrence.

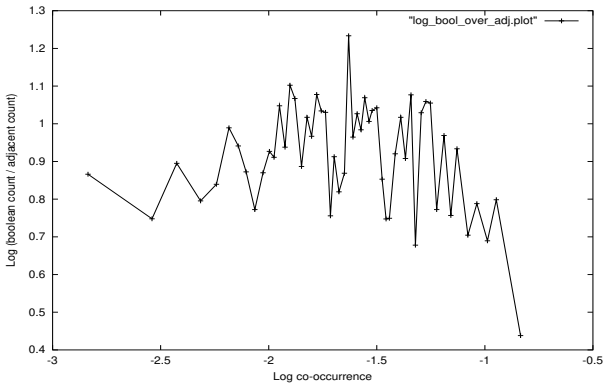


Figure 7: Proportion of Boolean to adjacent document counts, $\frac{\{\#d|bool \in d\}}{\{\#d|adj \in d\}}$, as a function of co-occurrence.

These graphs show that there is no significant different between probability of occurrence in relevant and non-relevant documents for adjacent and Boolean documents. At a given $p(occ)$, both an adjacent phrase and a Boolean phrase appear in the same number of relevant and non-relevant documents. Adjacent phrases are not automatically better, as one might expect. Something else must account for the weight of evidence pattern evident in Figure 3.

**Probability of Occurrence for Adjacent and Boolean Phrases** This leads to the conjecture that varying weight of evidence is best expressed as a function of different probabilities of occurrence in the collection for adjacent and Boolean phrases.

Figure 6 shows adjacent and Boolean document counts as a function of co-occurrence. The document count of a phrase is simply the probability of occurrence, $p(occ)$, of the phrase multiplied by the size of the collection. This graph shows that as co-occurrence increases, $p(occ)$ of both adjacent and Boolean phrases also increases.

Figure 7 shows the proportion of Boolean to adjacent documents as a function of co-occurrence. Also note that $\frac{Boolean\ p(occ)}{adjacent\ p(occ)} = \frac{\{\#d|bool \in d\}}{\{\#d|adj \in d\}}$. In other words, since the collection size is the same for both an adjacent and Boolean phrases taken from the same collection, the proportion of Boolean to adjacent phrases is dependent only on their total counts within that collection.

The first graph, Figure 6, was an attempt to find variation between the probability of occurrence for Boolean and adjacent phrases as co-occurrence increases. We see that, in general, the probability of occurrence increases as co-occurrence increases for both types of phrases. Both plots appear to have the same slope, but we want to make sure. There could be subtle, important differences. Figure 7 better depicts the two slope differences, plotting the ratio of the probabilities of occurrence for Boolean to adjacent phrases against co-occurrence.

This latter plot is what we hoped to see. It shows us that when co-occurrence is small, the proportion of Boolean to adjacent phrases is also small. When co-occurrence is in the mid-range, the proportion is larger. Finally, when co-occurrence is large, the proportion of Boolean to adjacent phrases goes down again. Admittedly, the plot is rough and very noisy, but those three general trends are there. Even more interesting, however, is the similarity of the shape of this plot to Figure 3, weight of evidence as a function of co-occurrence.

It is upon this similarity that we can capitalize.

**Proportion and Weight of Evidence** Based on the observation that Figure 7 looks so similar to Figure 3, we make the following supposition: Weight of evidence of adjacent phrases given the Boolean phrase is better expressed as a function of the proportion of Boolean to adjacent phrases than as a function of co-occurrence.

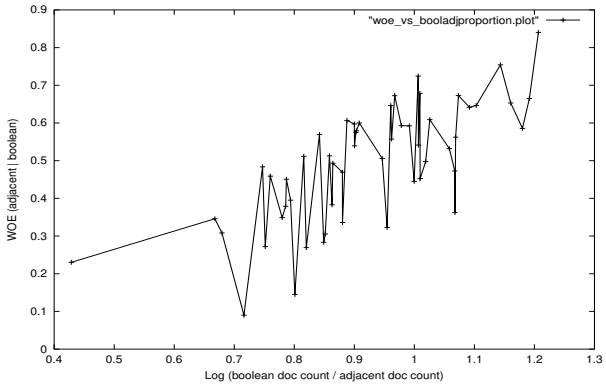


Figure 8: Weight of evidence of adjacent phrases given Boolean phrases as a function of the proportion of Boolean phrase document counts to adjacent phrase document counts, $\frac{\{\#d|bool \in d\}}{\{\#d|adj \in d\}}$

Figure 8 is a plot of this weight of evidence against this proportion. We see that, indeed, weight of evidence is linearly correllated with proportion. The more Boolean documents there are, relative to the number of adjacent documents, the more valuable adjacent phrases are. When making the decision about whether to use a phrase operator which uses the adjacent or the Boolean form of a phrase, one need only look at the collection-based proportion of documents with both forms of the phrase. Rather than calculate co-occurrence, we calculate proportion.

**Conclusion** On one hand, this result appears to be obvious. When there is very little difference between the sizes of the Boolean and adjacent document sets, the value of Boolean over adjacent phrases will of course be small. For example, if the Boolean and the adjacent document sets were exactly the same size (every time both subterms of a phrase co-appeared in a document, they co-appeared adjacent), then the weight of evidence of adjacency given Boolean could only be zero.

On the other hand, it is less obvious that as the proportion gets larger, the adjacent documents will have higher relevance ratios than the Boolean relevance ratios. For example, suppose a phrase is found in hundred Boolean documents and in only one adjacent document. Why should this one adjacent document be a relevant document? Our results nevertheless show that this is the trend.

Our exploratory data analysis using weight of evidence verified our conjecture that the adjacent structural form is a better indicator of document relevance than the Boolean structural form. The analysis also revealed a pattern we did not expect. We did not know that increasing levels of co-occurrence was actually not an indicator of increasing effectiveness of adjacency. Instead, the proportion of Boolean to adjacent documents is directly related to the value of adjacency for information retrieval.

# 6 Phrase Structuring: Anaphora

Recall that the problem of phrase structuring is analogous to the problem of, already having identified a phrase, how to recognize the phrase again. Proximity was one way of dealing with this recognition. Another way which we now examine is phrase anaphora.

An anaphor is a word or phrase which references the same object as another word or phrase. In this paper we examine phrase subterms. We want to find those subterms, those possible anaphora, which co-reference the same object as the original phrase to which the subterm belongs. For example, the phrase *Great Britain* and the subterm *Britain* seem to co-reference the same object. However, the subterm *Great* does not share this same co-reference.

Our conjecture is that subterms and their parent phrases are robust co-references when the subterm and the parent phrase share the same document context. In other words, if the subterm and the original phrase co-reference the same object, they will also co-occur within the same document. For example, it is clear that *Alzheimer's* and *Alzheimer's disease* are strong co-references, and that *Parkinson's* and *Parkinson's disease* are strong co-references. But what about *disease*? Are we as confident that this latter subterm references the same object as either parent phrase? If so, which of the two aforementioned phrases? Document context provides the answer. If the document mentions *Alzheimer's disease*, for example, then later on mentions *disease*, then *disease* likely references *Alzheimer's disease* and not *Parkinson's disease*.

Rather than only using a single document, however, we will calculate the contextual strength between a subterm and its parent phrase using all the documents in the collection. We use co-occurrence (see section 2.5) between the subterm and its parent phrase as a measure of overall collection context.

The question is whether this context measure, co-occurrence, indicates co-references which are useful for text retrieval. Weight of evidence again proves useful. Now that we have a question to ask, we can ask it directly of the documents, queries, and relevance judgements between the two. We want to explore the weight of evidence in favor of relevance provided by a phrase's subterm $(e_1 = t_n)$[13] conditioned upon already having seen the parent phrase $(e_2 = t_1 t_2)$ in an adjacent form somewhere in the document.[14] This is given by Equation 5:

$$woe(h : e_2|e_1) = \log \frac{\frac{\{\#d|rel(d) \wedge (t_n \in d) \wedge (t_1 t_2 \in d)\}}{\{\#d|\overline{rel}(d) \wedge (t_n \in d) \wedge (t_1 t_2 \in d)\}}}{\frac{\{\#d|rel(d) \wedge (t_1 t_2 \in d)\}}{\{\#d|\overline{rel}(d) \wedge (t_1 t_2 \in d)\}}} \tag{5}$$

Once we have calculated co-occurrence and weight of evidence we are ready for an exploratory analysis. We plot weight of evidence against the co-occurrence score between phrase subterm and its adjacent parent phrase using the small EDA set.

**Results** The results of this exploration are found in Figure 9. Note that there are twice as many data points as there are phrases in the small EDA set. Each phrase has two possible anaphora: *fiber* with *fiber optic* as well as *optic* with *fiber optic*.

[13] Note that $n = 1$ and $n = 2$. In other words, we calculate *woe* for both subterms, $t_1$ and $t_2$, separately

[14] Note that we draw a distinction between the subterm that we find as part of the original phrase, and the subterm that we find standing alone, apart from the original phrase. All free-standing occurrences of a subterm are counted separately from occurrences which have already been swallowed up into an adjacent version of the original phrase.

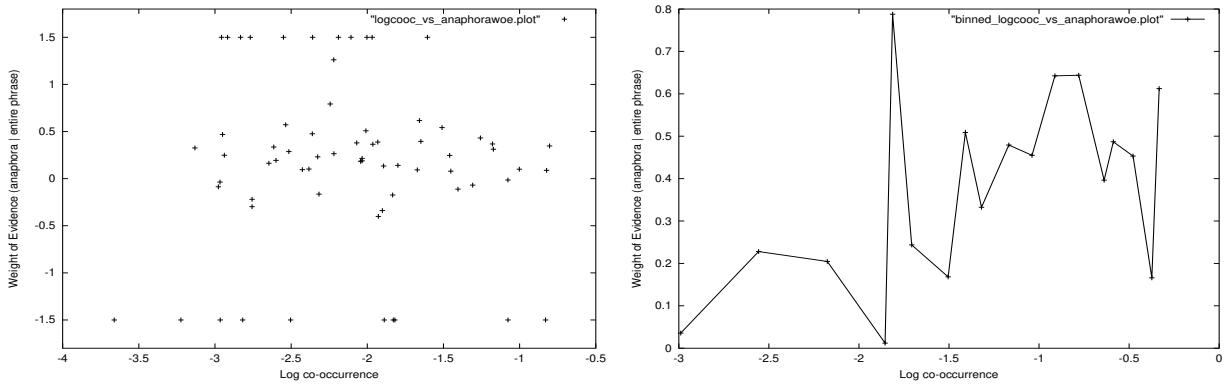


Figure 9: For the various phrase anaphora $t_n$ taken from TREC topics 301-350 judged on volumes 4+5. Log co-occurrence of the phrase and each of its possible anaphora is plotted against the weight of evidence of that anaphora, given the original phrase. *Left:* Unbinned/raw data. Note that $+\infty$ and $-\infty$ are plotted as +1.5 and -1.5. *Right:* Binned data.

Though the plots are rough, a pattern begins to emerge. As co-occurrence between the subterm and the original phrase increases, the weight of evidence in favor of relevance that subterm provides over the original term also seems to increase. The number of data points used, however, is still relatively small. A larger study should be undertaken before a more confident claim is made. But what is valuable here is not just this result (or lack thereof), but the method by which we come to this result. It is through this method that one may directly explore the value of phrase anaphora.

# 7 Phrase Weighting: *idf*

Among the more well-known approaches for term weighting is inverse document frequency [Jon72]. The terms that have been explored, however, are words. To our knowledge, no work has been done on whether document frequency is a useful discriminator for phrases, and if so, whether its discrimination works the same for phrases as for terms. If *idf* indeed behaves differently for phrases, it is not advisable to use the same *idf* weighting function for both.

[Gre98] provides a theory about why *idf* is effective a weighting mechanism for retrieval. The theory decribes how the relationship between $\log \frac{p(occ|rel)}{p(occ)}$ and document frequency shows how *idf* is a good term weighting feature. An exploratory analysis in that paper gives us a good visualization of an *idf* curve for words.

In this section we reconstruct these experiments using adjacent phrases rather than words. The basic theory is unchanged; we want to see if phrases exhibit the same behavior as words. We present both the binned and unbinned versions of the data for better visualization as well as for comparison against the word plots. As we begin exploring this weighting issue, it is important to keep in mind that not only have we fixed our phrase identification technique, but now also our structuring technique as well. This exploratory analysis of *idf* involves only the adjacent form of those phrases in our large EDA set.

**Occurrence in Non-relevant Documents** Figure 4 shows the $p(occ|\overline{rel})$ as a function of $p(occ)$ for adjacent phrases. As expected, this relationship is linear, just like it is for words. The one different we note here is

that $p(occ)$ is much smaller for phrases than it is for words. $p(occ)$ ranges from $[0 - 0.6]$ for words, while it ranges from $[0 - 0.045]$ for phrases.

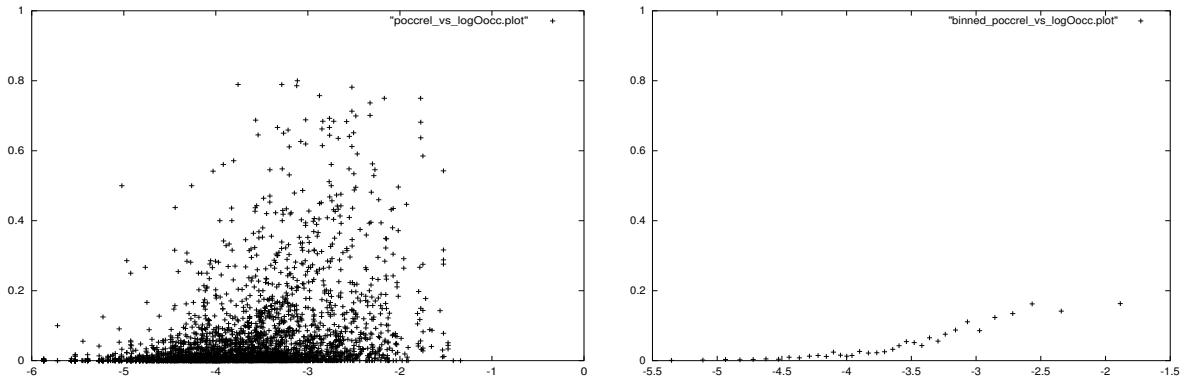


Figure 10: $p(occ|rel)$ as function of $\log O(occ)$

**Occurrence in Relevant Documents** Figure 10 is a plot of $p(occ|rel)$ as function of $\log O(occ)$, where $O$ is defined as *odds*. The resulting binned graph shows that increase in $\log O(occ)$ is accompanied by an increase in $p(occ|rel)$. The more a phrase appears in the collection, the more it appears in relevant documents. This is the same pattern observed for words.

We plot against $\log O(occ)$ because phrases, like words, typically have such low document frequencies, relative to the size of the collection, that there is little difference between $\log p(occ)$ and $\log O(occ)$. Furthermore, as we mentioned above, $p(occ)$ for phrases is an entire order of magnitude less than terms. So the log scale is even more useful for providing a better view of those low frequency data points which otherwise remain hidden.

Despite the positive relationship between $p(occ|rel)$ and $\log O(occ)$, there are two differences between phrases and words. The first concerns this observation that phrases are rarer than words. Phrase $\log O(occ)$ ranges from about $-6$ to $-1.5$, whereas words range from about $-5$ to 0.5. Despite the difference in ranges, in those document frequencies where phrases and words overlap ($-5$ to $-1.5$) they share similar $p(occ|rel)$ values. There is a strong similarity between the phrase curve and the word curve not only in the shape of the curve, but in the actual values of points along the curve. (The reader is again referred to [Gre98] for these word plots.)

The second difference is observable in the unbinned data. Of the 4641 phrases used to contruct these plots, not a single phrase had $p(occ|rel)$ greater than 80%. This is somewhat unexpected. There were some words with $p(occ|rel)$ between and including 80% to 100%, even at lower document frequencies. This suggests that phrases are not as useful as words. Nevertheless, once the data points are binned, these differences disappear into the larger pattern.

**Log of the Ratio of $p(occ|rel)$ to $p(occ)$** Finally we normalize $p(occ|rel)$ by the $p(occ)$ for that phrase in the collection, also using the log scale. Figure 11 is a plot of $\log \frac{p(occ|rel)}{p(occ)}$ against $\log O(occ)$.

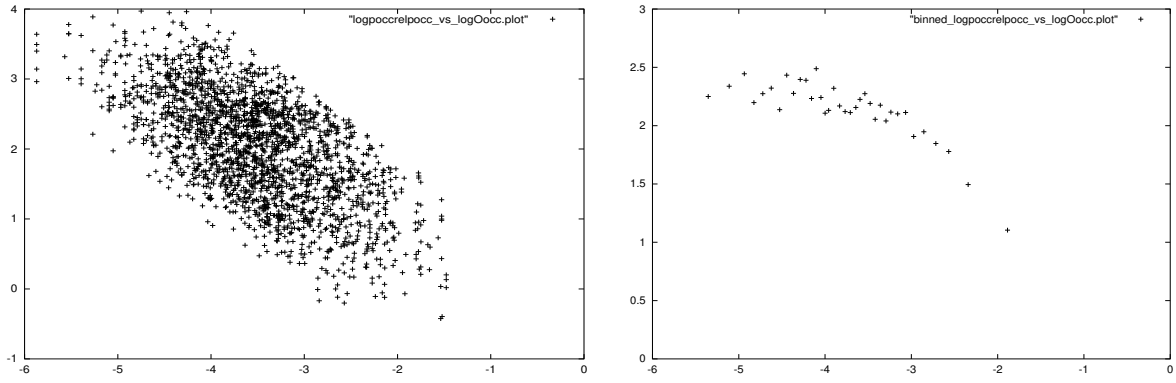


Figure 11: $\log \frac{p(occ|rel)}{p(occ)}$ as function of $\log O(occ)$

This normalized view of $p(occ|rel)$ accounts for the $p(occ)$ of the phrase as a whole. If higher frequency phrases indeed appear in more relevant documents, is the proportion of relevant documents in which they occur nevertheless larger than we would expect to see, given their likelihood of occurrence in the collection? Similarly, even though low frequency phrases appear in fewer relevant documents, is this appearance also disproportionate to their appearance in the collection?

Figure 11 shows that query phrases, like query words, are more likely to appear in relevant documents. Furthermore, this likelihood is a function of document frequency. What we see, in fact, is the same *idf* curve observed for words.

Phrases are still rarer than words, so there are almost no data points higher than $\log O(occ)$ of about $-1.5$. If we concentrate instead on the remaining data points, we see that the the *idf* curve for phrases looks almost exactly the same as that for words. We observe in our binned plot a roughly linear negatively sloped relationship between $\log \frac{p(occ|rel)}{p(occ)}$ and $\log O(occ)$: The linear relationship is stronger in the mid-range of document frequencies and flattens at the low-range frequencies.

Not only are the phrase and word *idf* curves similar in shape, but in actual values as well. The *idf* value for phrases flattens out around $\log \frac{p(occ|rel)}{p(occ)} = 2.5$, and also begins this flattening process at $\log O(occ)$ of about $-3$. These are the same values observed for words.

There are differences, however. The first difference is the distribution of the data points. In the word plot, most words are found in the mid-to-high range frequencies, whereas with phrases, most phrases are found in the mid-to-low range. The second difference is that when the phrase values do flatten out at low frequencies, they become somewhat erratic, and more so than words.

Even with these differences, it is remarkable that, on average, phrases and words have similar relevance values at similar *idf* values. This indicates that a phrase's value as a discriminator is indeed a function of document frequency alone, rather than a function of document frequency relative to the *df* of the most common phrase in the collection, or some other measure. Phrases are much rarer than words, so it could have been possible that the entire *idf* curve for phrases, while similar in shape to the curve for words, might be shifted in its entirety along the $\log O(occ)$ axis. But this is not the case: Phrases simply behave like medium-to-rare words, giving similar discriminatory value at similar collection *idf*.

# 8 Phrase Proximity and *idf*

Salton and Buckley [SB88] explain the difficulty with incorporating phrases into retrieval experiments which already use words. The problem is that the techniques used for identification and/or structuring have historically either been too strict in their requirements, yielding too few new identifiers beyond words to be useful, or they are too lax in their requirements, yielding too many identifiers, which produce too many non-relevant documents[15].

In section 4 we explored the tradeoff between phrase structuring that was quite strict (adjacency) and quite loose (Boolean). We concluded that strict was better on average because the ratio of relevant to non-relevant documents is higher at adjacency than it is at Boolean. That does not, however, say that Boolean by itself is a useless construct. Our weight of evidence says nothing about the actual document counts at adjacency. We discarded many non-relevant documents going from Boolean to adjacency, but we also discarded many relevant documents. Adjacency is precise, but it sacrifices recall. Boolean gives us higher recall, but at the cost of the precision of adjacency. Perhaps there is a balance between the two.

We already suspect that looser phrase structures are going to be, on average, less precise than their stricter forms. Does this imprecision affect how we weight looser phrases, in an *idf* weighting scheme? Do strict and loose phrases have similar *idf* values at similar document frequencies, or does the imprecision of looser phrase make stricter and looser phrases incomparable? If looser phrases can indeed be weighted the same as stricter phrases, then future work on structuring may freely experiment with different proximities without concern that each proximity requires a completely different weighting scheme.

Once again we do an exploratory analysis and examine *idf* for these looser phrases. Our phrase identification technique is still fixed, but we explore now a number of different structuring techniques, ones which accommodate looser phrases. We pick as our test subjects the looser phrase structures $prox20$, $prox50$ and $prox100$, where the value $n$ in $proxn$ refers to all occurrences of two phrase subterms at a lexical distance *less than or equal to* $n$. We duplicate the experiments from section 7 with these looser structures again using the large EDA set. We create the same intermediate graphs, but for this paper only show our final result, the plot of $\log \frac{p(occ|rel)}{p(occ)}$ as function of $\log O(occ)$. Furthermore, we do not show the plot for $prox100$ because it is so similar to the plot for $prox50$.

**Discussion** Figure 12 shows our results, with the $prox1$ (adjacent) phrase structure in the background as reference. There are a number of things happening here. As we expand the proximity of our phrases from 1 to 20 to 50, we increase our document frequency. All documents which contain a phrase at $prox1$ also contain this phrase at $prox20$, so naturally we expect to see a general right shift along the $\log O(occ)$ axis as proximity expands. There will never be fewer documents at a larger proximity, and chances are there will actually be more. Close examination of the curve shows that, indeed, the $prox20$ points have higher df than the $prox1$ points, while the $prox50$ points have higher document frequency than the $prox20$ points.

It might appear that the shift does not happen on the rare side of the *idf* curve, with the leftmost $prox50$ point to the left of the leftmost $prox20$ point. But note that $prox50$ now includes a number of phrases whose

[15] In this paper, *strict* and *loose* are informal terms. *Strict* refers to a phrase proximity structure where the lexical distance between phrase subterms is small; *loose* refers to a structure where the distance is permitted to be large

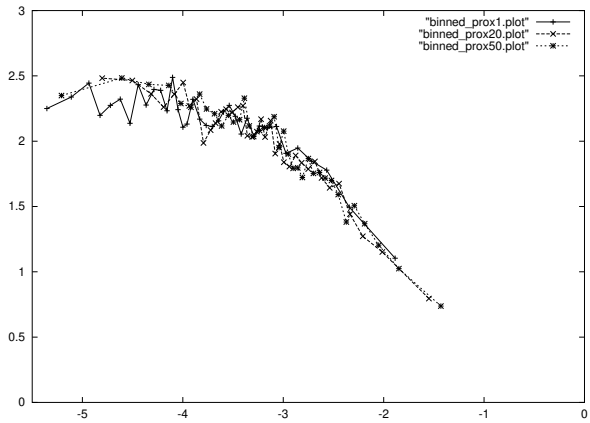


Figure 12: $\log \frac{p(occ|rel)}{p(occ)}$ as function of $\log O(occ)$, for phrases at cumulative proximities 1, 20 and 50.

$prox20$ and $prox1$ $\log O(occ)$ was $-\infty$: the phrases simply didn't appear in the collection at lower proximities, whereas they do with higher. So the leftmost $prox50$ point is actually a right shift from $-\infty$.

Even more interesting than the document frequency shift is the pattern of downward shifts along the $\log \frac{p(occ|rel)}{p(occ)}$ axis. If $prox20$ and $prox50$ phrases were different entities than adjacent phrases or single words, we would expect to see deviation from the values along this axis. Instead, every right shift along the $\log O(occ)$ axis is accompanied by a downward shift along the $\log \frac{p(occ|rel)}{p(occ)}$ axis *in a proportion seemingly governed only by document frequency.*

In other words, rare $prox20$ phrases simply behave like slightly less rare $prox1$ phrases. Medium-frequency $prox50$ phrases behave like medium-to-rare $prox20$ phrases. We examined this data in a number of ways, but visually it tells its own story best. We already know that adjacent phrases follow the *idf* curve, but this plot shows us that $prox20$ and $prox50$ phrases appear to follow the *idf* curve as well. $prox100$ phrases were so similar to $prox50$ phrases that we didn't even plot the curve; medium frequency $prox100$ phrases behave like medium frequency $prox50$ phrases.

Our intuition was that since larger proximities are more imprecise than smaller ones we could not use larger proximity phrases the same way we used adjacent phrases. Instead, we found that both follow the same *idf* curve. A weighting scheme which uses *idf* will need to make no special provisions when expanding a phrase's proximity structure. If the phrase's document frequency changes due to a different structural form, then *idf* will automatically downweight the looser phrase or upweight the stricter phrase appropriately. Words, adjacent phrases, and looser phrases, at least those we examined, are all comparable under an *idf* weighting scheme.

# 9 Conclusion and Future Work

This paper explored a number of different phrase patterns. We found that adjacent phrases tended to be better than Boolean phrases. When we asked how much better, we found that the proportion of Boolean to adjacent document counts was a clearer indicator than co-occurrence. We introduced a method for examining

phrase anaphora. Finally, we showed that not only do adjacent phrases follow the same *idf* weighting as words, but looser $prox20$, $prox50$ and $prox100$ phrases do as well.

The value and contribution of the paper lies not only in uncovering interesting properties about phrases, but in providing a tool for the future exploration of other phrase patterns. Exploratory data analysis using weight of evidence was this tool. By splitting our analysis of phrases into three components, indentification, structuring and weighting, we were further able to to identify where, if at all, phrase usage has its problems. A valuable structuring technique should not be hidden by a useless weighting technique, or vice versa. There is much to be gained by analyzing phrase patterns directly, rather than relying on the summary statistics of a retrieval experiment.

We hope to ask more questions about various phrase structuring and weighting techniques, possibly even identification techniques as well. With an exploratory data analysis mindset and weight of evidence as a tool, we can explore old questions at the same time we discover new ones. We have already seen patterns that we didn't expect while looking for patterns we did expect. The more phrase patterns we look at the clearer our picture of phrases will become.

One possibility is to look at "proximity residuals" much in the same way [Gre99a] has looked at term frequency residuals. What is the weight of evidence of a phrase being $n+1$ or less words apart, given that it's already $n$ or less words apart, for all values of $n$? In other words, are there any interesting distances in between adjacency and Boolean? Perhaps a plot of these residuals would yield an interesting and useful pattern.

We could develop more detailed weighting model for phrases by looking at phrase frequency, an analogue to term frequency. We've already seen that *idf* is no different for some phrases and single terms, but what about phrase frequency as opposed to term frequency? Is phrase frequency equally useful?

Finally, another future direction would be to look at phrase stemming. Word stemming is fairly widespread, with algorithms such as the Porter stemmer [Por80] in common use. Do these stemmers also work for phrases? If so, how? Some phrases create their stems with the last constituent subterm: *fiber optic* $\rightarrow$ *fiber optics*. Other phrases create their stems with the first constituent subterm: *notary public* $\rightarrow$ *notaries public*. Still other phrases don't use any stemming: *third world* $\not\rightarrow$ *third worlds*. We can evaluate a phrase stemming algorithm by looking at the weight of evidence that a variant phrase stem provides over the original phrase. If this weight is positive, then the stem helps us. If it is zero, then at least it doesn't hurt us. If it is negative, we turn off stemming or look for a better algorithm. In this manner, we again explore the value of our phrase stemming algorithm directly, one stem at a time, rather than indirectly though retrieval experiments.

## 10 Acknowledgements

We would like to thank Warren Greiff for many hours of discussion; this paper could not have been written without his help. We would also like to thank Fang-fang Feng for the use of his phrase identifier. In addition, we would like to thank Don Byrd, Anton Leouski and Victor Lavrenko for their insightful comments.

## References


[Cal94] J. P. Callan. Passage-level evidence in document retrieval. In *Proceedings of ACM SIGIR*, pages 302–310, 1994.

[CCH92] J.P. Callan, W.B. Croft, and S.M. Harding. The inquery retrieval system. In *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, pages 78–83, 1992.

[Chu88] K. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143, 1988.

[CTL91] W.B. Croft, H. Turtle, and D. Lewis. The use of phrases and structured queries in information retrieval. In *Proceedings of ACM SIGIR*, pages 32–45, 1991.

[CX95] W.B. Croft and J. Xu. Corpus-specific stemming using word form co-occurence. In *Proceedings for the Fourth Annual Symposium on Document Analysis Information Retrieval*, pages 147–159, April 24-26 1995.

[Fag87] J.L. Fagan. *Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods.* PhD thesis, Cornell University, 1987.

[FC99] F. Feng and W.B. Croft. Probabilistic techniques for phrase extraction. Technical report, CIIR, Department of Computer Science, University of Massachusetts, Amherst, 1999.

[GC90] L.S. Gay and W.B. Croft. Interpreting nominal compounds for information retrieval. *Information Processing and Management*, 26(1):21–38, 1990.

[Goo50] I.J. Good. *Probability and the Weighing of Evidence.* Charles Griffen, 1950.

[Gre98] W.R. Greiff. A theory of term weighting based on exploratory data analysis. In *Proceedings of ACM SIGIR*, pages 11–19, 1998.

[Gre99a] W.R. Greiff. Empirical studies of query/document characteristics as evidence in favor of relevance. Technical report, CIIR, Department of Computer Science, University of Massachusetts, Amherst, 1999.

[Gre99b] W.R. Greiff. *The Maximum Entropy Approach and Probabilistic IR Models.* PhD thesis, University of Massachusetts, Amherst, 1999.

[Har95] D.K. Harman. The trec conferences. In R. Kuhlen and M. Rittberger, editors, *Hypertext - Information Retrieval - Multimedia; Synergieeffekte Elektronischer Informationssysteme, Proceedings of HIM '95*, pages 9–28. Universitaetsforlag Konstanz, 1995.

[HD79] F. Hartwig and B.E. Dearing. *Exploratory Data Analysis.* Sage Publications, 1979.

[HT96] D. Hawking and P. Thistlewaite. Relevance weighting using distance between term occurrences. Technical Report TR-CS-96-08, Department of Computer Science, Australian National University, 1996.

[Jon72] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[Kee92] E.M. Keen. Presenting results of experimental retrieval comparisons. *Information Processing and Management*, 28:491–502, 1992.

[Luh57] H.P. Luhn. A statistical approach to the mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317, October 1957.

[Por80] M.F. Porter. An algorithm for suffix stripping. *Program*, 14:130–137, 1980.

[SB88] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523, 1988.