Global Statistics in Proximity Weighting Models

Craig Macdonald, Iadh Ounis
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ, UK
{craigm,ounis}@dcs.gla.ac.uk

ABSTRACT

Information retrieval systems often use proximity or term dependence models to increase the effectiveness of document retrieval. Many of the existing proximity models examine document-level *local statistics*, such as the frequencies that pairs of query terms occur within fixed-size windows of each document, before applying standard or adapted weighting functions – for instance Markov Random Fields. Term weighting models use Inverse Document Frequency (IDF) to control the influence of occurrences of different query terms in documents. Similarly, some proximity models also take into account the frequency of pairs of query terms in the entire corpus of documents. However, pair frequency is an expensive statistic to pre-compute at indexing time, or to compute at retrieval time before scoring documents. In this work, we examine in a uniform setting, the importance of such *global statistics* for proximity weighting. We investigate two sources of global statistics, namely the target corpus, and the entire Web. Experiments are conducted using the TREC GOV2 and ClueWeb09 test collections. Our results show that local statistics alone are sufficient for effective retrieval, and global statistics usually do not bring any significant improvement in effectiveness, compared to the same proximity approaches that do not use these global statistics.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*

General Terms

Performance, Experimentation

Keywords

Proximity, Term Dependence, Global Statistics

1. INTRODUCTION

In Information Retrieval (IR) systems, documents are primarily matched and ranked using the presence and frequency information of query terms occurring in these documents. Term weighting models, from TF-IDF, BM25 [22], Language Modelling [28] or Divergence from Randomness [1] are typically employed to score the documents. Each model has some notion of the discriminating power or importance of each term (as modelled by Inverse Document Frequency (IDF) [24] or smoothing [28]). Indeed, Fang & Zhai identified the Term Discrimination Constraint as a heuristic that should be present in all weighting models. This constraint ensures that given a fixed number of occurrences of query terms, a document that has more occurrences of discriminative terms will be favoured [10].

In recent years, proximity (or term dependence) models have made a significant impact on the retrieval performance of IR systems [6, 15, 19], particularly for retrieval from very large corpora. In such models, the proximity of query terms in documents are taken into account, and documents where pairs of query terms occur in close proximity are favoured. By doing so, precision is often enhanced [17].

Similar to their term-weighting counterparts, some proximity models take into account the frequency of each pair of query terms within the entire corpus of documents. This represents some measure on the importance of a pair of query terms, based on their frequency of occurrence in the entire collection. However, such global pair frequency statistics are expensive to obtain [17]. Three options are commonly considered: (i) Using two passes of the inverted posting lists to complete scoring is a costly proposition, and incompatible with other techniques which can markedly reduce retrieval time, such as dynamic pruning [26]. (ii) Instead, some authors suggest the maintenance of a dedicated lexicon or inverted file for pairs of posting lists at indexing time – however, with billions of possible pairs of query terms, the problem becomes how to identify which particular pairs of terms should have special pair posting lists built. (iii) Lastly, [17] discussed how the global statistics could be approximated, but did not examine their actual importance. In contrast to these options, other proximity models make no use of global statistics during scoring. This leads us to naturally question the importance of such *global statistics* for effective proximity scoring, in addition to document-level *local statistics*.

In this work, we contribute an empirical examination of the importance of global statistics during proximity scoring. Indeed, we train and evaluate several state-of-the-art proximity weighting models with and without global statistics, and conclude on the resulting effectiveness. Moreover, we investigate if using a larger collection to obtain the global statistics has any impact on effectiveness. In particular, we use an index of the Web (provided by the Bing search engine using the Microsoft Web N-gram Service API) to obtain the global statistics. The remainder of this paper is structured as follows: Section 2 describes several proximity weighting models that we experiment with in this work; Section 3 discusses the use of global statistics in proximity weighting; Our research questions and experimental setup are described in Section 4; Results and analysis follow in Section 5; Concluding remarks are made in Section 6.

SIGIR 2010 WEB N-GRAM Workshop, July 23, 2010, Geneva, Switzerland.

2. PROXIMITY MODELS

There are many queries where the relevant documents contain occurrences of the query terms in close proximity. Hence, modern retrieval systems apply not just single-term weighting models when ranking documents. Instead, proximity weighting models are applied, which highly score the co-occurrence of pairs of query terms in close proximity to each other in documents [8]. Some proximity (or term dependence) models have recently been proposed that integrate single term and proximity scores for ranking documents [15, 19]. In this manner, the ranking model of an IR system for a query Q can be expressed as:

$$\begin{aligned} \text{score}_Q(d,Q) &= \omega\, S(d) + \sum_{t \in Q} \big(\text{score}(tf, *_d, t)\big) \\ &+ \phi\, \text{prox}(d,Q) \end{aligned} \quad (1)$$

where $S(d)$ is the combination of some query independent features of the document d (e.g., PageRank, URL length), and $\text{score}(tf, *_d, t)$ is the application of a weighting model to score tf occurrences of query term t in document d. $*_d$ denotes any other document statistics required by a particular weighting model, such as document length. $\text{prox}(d,Q)$ represents some proximity document scoring function. Control over the influence of the various features is achieved using weights ω and ϕ.

All proximity approaches examine the proximity of occurrences of query terms within documents, which we refer to as local statistics. In particular, some approaches examine the minimum or average distance between occurrences of query terms [6, 9, 25], and score highly documents where low distances occur. Meanwhile, in [3], the extent that a phrase occurring in a document matches the query is measured. Instead of examining distances, some other approaches examine the number of windows where more than one of query terms occur [15, 16, 19].

With such a plethora of alternative approaches for proximity ranking, it is not surprising that proximity weighting is increasingly being treated as a learning problem, with various proximity 'feature' functions being combined using machine learning techniques [9, 25]. However, in this paper, we focus on two proximity models that are theoretically founded in that they model proximity using windows, but using statistically similar methods to those used in term weighting. In particular, in the following we introduce two proximity models that define $\text{prox}(d,Q)$ in terms of pairs of query terms, namely the Markov Random Fields [15], and Divergence from Randomness proximity [19] approaches.

2.1 Markov Random Fields Proximity

In [15], Metzler & Croft defined the Markov Random Fields (MRF) approach to term dependence. In particular, two approaches were modelled, namely sequential dependence – where the pairs of sequentially adjacent query terms are considered – and full dependence – where all possible pairs of query terms are considered. In both cases, prox(d,Q) is calculated as:

$$\text{prox(d,Q)} = \sum_{p \in \text{pairs}(Q)} \text{score}\big(pf(t_i, t_{i+1}, d, k), l_d, p\big) \quad (2)$$

where $pf(t_i, t_{i+1}, d, k)$ represents the number of occurrences of the pair p of query terms (t_i, t_{i+1}) occurring in document d in windows of size k (abbreviated as pair frequency pf). The function $pairs(Q)$ defines how pairs of query terms are derived from the query. In particular, for sequential dependence (SD), all adjacent pairs of query terms are considered, while for full dependence (FD) all pairs are considered:

$$\begin{aligned} \text{pairs}_{SD}(Q) &= \{(t_i, t_{i+1}) \in Q\} \\ \text{pairs}_{FD}(Q) &= \{(t_i \in Q, t_j \in Q), i \neq j\} \end{aligned}$$

Typically in MRF, prox(d,Q) is instantiated twice, with $k = 2$ (to account for proximity of two terms as a phrase) and $k = 8$ (to account for proximity at approximately sentence level). In this work, for simplicity, we separate the two instantiations of Equation (2) such that we can examine the effectiveness of proximity at different window sizes. Following [15], $\text{score}(pf, l_d, p)$ is implemented using Dirichlet language modelling [28], but where pair frequency takes the role of term frequency, as follows:

$$\text{score}_{\text{Dirichlet}}\big(pf, l_d, p\big) = \log\left((1-\lambda_{LM})\frac{pf}{l_d} + \lambda_{LM}\frac{F}{T}\right) \quad (3)$$

where F is the frequency of pair p in the entire corpus, and T is the size of the entire corpus. $\lambda_{LM} = \frac{\mu}{\mu + l_d}$ is the Dirichlet smoothing (where $\mu \gg 0$ is a smoothing parameter). However, a disadvantage of MRF, which is not discussed in [15] is the presence of F, which is needed before scoring can commence. Indeed, for pairs of query terms, F is expensive to calculate, unless it has been precalculated at indexing time for particular pairs. We will return to this important fact later in this paper.

2.2 DFR Proximity

Divergence From Randomness (DFR) [1] models can also be used for proximity weighting in a similar fashion to the manner in which MRF uses Dirichlet language modelling [2, 19]. In general, DFR models for term weighting follow the form:

$$\begin{aligned} &score(tf, l_d, t) = Inf_1 \cdot Inf_2 \\ &= -\log_2(prob_1(tfn|Collection)) \cdot (1 - prob_2(tfn|E_t)) \end{aligned} \quad (4)$$

where $prob_1(tfn|Collection)$ is the probability of a term occurring with normalised frequency tfn in a document by chance, according to a given model of randomness. $prob_2$ is some function that calculates the information gain by considering if a term is informative (i.e., important) in a document. E_t is the *elite* set of documents – the set of documents containing t. The normalised term frequency tfn is obtained by normalising term frequency tf with respect to length of the document l_d and the average length of documents in the collection (avg_l), according to *Normalisation 2* [1]:

$$tfn = tf \cdot \log_2(1 + c \cdot \frac{avg\_l}{l_d}) \quad (5)$$

where $c > 0$ is a hyper-parameter.

One of the most popular DFR models is PL2 [1], which is particularly effective at high precision tasks. PL2 deploys the Poisson randomness model (denoted P in the DFR framework), which assumes that the occurrences of a term are distributed according to a binomial model. Then, the probability of observing tfn occurrences of a term in a document is given by the probability of tfn successes in a sequence of F Bernoulli trials with N possible outcomes:

$$prob_1(tfn|Collection) = \binom{F}{tfn} p^{tfn} q^{F-tfn} \quad (6)$$

where F is the frequency of term t in the collection of N documents, $p = \frac{1}{N}$ and $q = 1 - p$. If the maximum likelihood estimator $\lambda = \frac{F}{N}$ of the frequency of a term in this collection is very low, or in other words $F \ll N$, then the Poisson distribution can be used to approximate the binomial model described above, making use of a

Stirling series to expand the factorials. In this case, the informative content of $prob_1$ is given as follows:

$$-\log_2\left(prob_1(tfn|Collection)\right) = \quad (7)$$
$$tfn \cdot \log_2 \frac{tfn}{\lambda} + (\lambda - tfn) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tfn)$$

The Inf_2 component of PL2 uses the Laplace law of succession (denoted L in the DFR framework), which corresponds to the conditional probability of having one more occurrence of a term in a document, where the term appeared tfn times already:

$$1 - prob_2(tfn|E_t) = 1 - \frac{tfn}{tfn+1} = \frac{1}{tfn+1} \quad (8)$$

DFR models can also be applied for proximity scoring [2, 19], by substituting tfn with pfn, and counting l_d in terms of windows of size k, instead of tokens. Hence, the PL2 proximity score for a pair of query terms p for document d is:

$$\text{score}_{\text{PL2}}(pf, l_d, p) = \frac{1}{pfn+1}\big(pfn \cdot \log_2 \frac{pfn}{\lambda}$$
$$+(\lambda - pfn) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot pfn)\big) \quad (9)$$

where λ is the mean and variance of a Poisson distribution, given by $\lambda = \frac{F}{N}$.

Similar to language modelling, PL2 relies on F – the frequency of the term in the whole collection – to provide IDF-like discrimination between query terms. However, as noted above, in the proximity setting, F is expensive to calculate. Instead, as an alternative, the BiL2 model was proposed [13, 19], which does not consider F. In contrast to PL2, in BiL2, the randomness model measures the probability of tfn successes in a sequence of avg_l Bernoulli trials with only avg_l possible outcomes:

$$prob_1(tfn|Collection) = \binom{avg\_l}{tfn} p^{tf} q_p^{avg\_l - tfn} \quad (10)$$

where avg_l is the average length of all documents, $p = \frac{1}{avg\_l - 1}$ and $q = 1 - p_p$. Once again, by applying the Poisson approximation (using the Lanzcos approximation of the Γ function [21, p. 213], which results in lower error than the Stirling series [20]), together with Laplace and Normalisation 2, the final formula for BiL2 is as follows:

$$\text{score}_{\text{BiL2}}(pf, l_d, p) = \frac{1}{pfn+1} \cdot \quad (11)$$
$$\Big(\quad - \quad \log_2 (avg\_w - 1)! + \log_2 pfn!$$
$$+ \quad \log_2(avg\_w - 1 - pfn)!$$
$$- \quad pfn \log_2(p_p)$$
$$- \quad (avg\_w - 1 - pfn)\log_2(q_p)\Big) \quad (12)$$

where $avg\_w = \frac{T - N(k-1)}{N}$ is the average number of windows of size k tokens in each document in the collection, N is the number of documents and T is the total number of tokens in the collection. $p_p = \frac{1}{avg\_w - 1}$, $q_p = 1 - p_p$, and pfn is the normalised frequency of the pair of query terms p, as given by applying Normalisation 2 (Equation (5)) to pf.

Apart from the use of different approximations, BiL2 is a very similar model to PL2. In particular, instead of measuring the probability of tfn with a prior of F, it measures pfn occurrences with a prior of avg_w. This means that it no longer relies on global statistics, namely F, the frequency of pair of query terms p in the entire collection. However, in contrast to PL2 and Dirichlet language modelling, BiL2 would not be a useful model for term weighting as it performs no discrimination between query terms.

3. GLOBAL STATISTICS IN PROXIMITY WEIGHTING

With term weighting models, the use of an IDF-like component, as identified by Spärck-Jones [24], is essential to effectiveness. IDF is based on the intuition that query terms which are frequent in the entire collection are unimportant, and therefore unlikely to cause much discrimination between relevant and irrelevant documents. Indeed, Fang and Zhai [10] identified a heuristic that they named the Term Discrimination Constraint, that all weighting models should encompass, whereby given a fixed number of occurrences of query terms, a document that has more occurrences of discriminative terms will be favoured.

However, for proximity weighting models that use global statistics – for instance, the number of documents containing a pair of query terms, or the frequency of the pair in the corpus (F) – there is a problem that these can be expensive to calculate [17]. In particular, for a standard inverted index containing postings lists with position information for each term, three options are possible:

(i) For a small index, all postings for all query terms may be maintained in memory, with a first scan to calculate global statistics and a second to perform the document scoring.

(ii) For a larger index, where it is not possible to hold all query term postings in memory, two passes of the inverted index postings are required – a costly proposition.

(iii) In contrast, other works have proposed approximations of the global statistics for phrases, based on the statistics of the constituent terms [4, 17].

Other approaches [23, 29, 30] to proximity weighting have relied on maintaining separate posting lists for pairs of query terms. However, with billions of possible pairs of terms, the problem becomes how to identify which particular pairs of terms should have special pair posting lists built.

It is intuitive that occurrences of query terms with low discriminatory power should not be given as much emphasis as query terms with higher discriminatory power. For proximity models that apply an IDF-like component in the same manner as term weighting, the term discrimination constraint [10] can be paraphrased as follows:

Pair Discrimination Constraint: if two documents each contains different pairs of query terms with the same frequency, then the document that contains the more discriminative pair, as suggested by the global statistics of the pairs, should be favoured.

However, in proximity weighting, the occurrence of any pair of the query terms is likely to positively impact on the likelihood of the document's relevance, because the occurrence of pairs are comparatively rare events. Hence, it is less likely that the importance of a pair of query terms will be over-estimated.

In addition, in a term weighting model, if there is only one query term, then the IDF component can have no impact on the effectiveness of the weighting model, as no discrimination between query terms is required. Similarly, in proximity weighting, as queries tend to be short, the number of pairs of query terms can be very few. Indeed, the performance of the proximity models may vary significantly with respect to the length of the query. For instance, for obvious reasons single term queries do not benefit from proximity. Moreover, if only one pair is present, then global

statistics will have no impact on the effectiveness of the proximity weighting function alone. Nevertheless, the global statistics may still have an important role in estimating the importance of proximity for the entire retrieval system of Equation (1). Indeed, it follows from the pair discrimination constraint that:

Corollary: for two different queries each of one pair and fixed ϕ, the global statistics have a role in indicating the importance of $\text{prox}(d, Q)$ for each query – for a query with a highly discriminating pair, $\text{prox}(d, Q)$ should be higher than for a high frequency pair.

In this paper, we empirically investigate the importance of global statistics for proximity. In particular, we examine their benefit to retrieval effectiveness. Moreover, we experiment with using different corpora in the calculation of the global statistics, to determine if using a larger "Web-scale" corpus has any bearing on our conclusions. This work differs from that of [17], which only compared different approximations of global statistics, but did not question their actual necessity in the first place.

Metzler [14] notes the lack of a study into the importance of global statistics in proximity models. Indeed, in his own implementation of MRF in the Ivory retrieval system[1], a constant frequency of $F = \frac{N}{50}$ is assumed for *all* pairs of query terms. In contrast, the BiL2 model makes no use of global statistics when calculating the importance of a pair of query terms in a document.

4. EXPERIMENTAL SETUP

In this work, we aim to address the following research questions:

1. To what extent do global statistics matter for proximity scoring in addition to local statistics? (Section 5.1)
2. Are longer queries benefited differently by global statistics during proximity scoring? (Section 5.2)
3. Does using larger corpora for global statistics impact on the resulting effectiveness? (Section 5.3)

To address these research questions, we perform experiments using two large-scale TREC test collections, namely .GOV2 and ClueWeb09, with corresponding adhoc retrieval tasks. In particular, .GOV2 consists of 25 million documents crawled from the `.gov` domain of the Web, while we use the first 50 million English documents of the ClueWeb09 general Web crawl (commonly denoted CW09B). We index both corpora using Terrier[2] [18], applying Porter stemming (unless otherwise noted) and removing standard stopwords. During retrieval, documents from each corpus are initially ranked by BM25 [22], before the proximity weighting models are applied to the top 1000 scored documents.

For each corpus, we train on 100 queries, and test on 50. In particular, Table 1 details the queries used in our experiments. For .GOV2, our test queries correspond to the TREC setting from the 2006 Terabyte track [5]). For CW09B, we select training and testing queries from the 2009 Million Query track [7][3]. Moreover, as proximity scoring may benefit differently queries of different lengths, Table 1 provides a breakdown on the number of queries for each length. Finally, the number of pairs of query terms identified for both sequential dependence (SD) and full dependence (FD) are also shown.

[1] `http://www.umiacs.umd.edu/~jimmylin/ivory`
[2] `http://terrier.org`
[3] We could have used the TREC 2009 Web track, however with only 50 adhoc queries, 16 of which are single term queries, we perceived this as insufficient to provide both training and testing queries.

In our experiments, we apply both MRF and DFR-based proximity approaches, and study both sequential dependence and full dependence variants. In addition, we test two different window sizes, namely $k = 2$ and $k = 8$. In terms of global statistics, our experiments cover the use of the *target* corpus for global statistics (i.e., .GOV2 or CW09B), as well as using global statistics derived from a larger *external* corpus, namely a Web search engine, through the use of the Microsoft Web N-gram Service API [27]. In contrast to these settings, we also test proximity models that do not use global statistics. In particular, for MRF, we test using the default of $F = \frac{N}{50}$ used by Ivory. For DFR proximity, we compare PL2 and BiL2 – these models are similar (modulo different factorial approximations) except that PL2 uses F, while BiL2, does not. During our analysis, we use significance testing to determine if there is any statistically significant differences between proximity models that do use global statistics and those that do not.

For each setting, we train the proximity models to give high performance on the training query set. In particular, highly performing values for the normalisation parameters (μ for MRF or c for DFR) and proximity weight ϕ are found using simulated annealing [12], by directly maximising the mean average precision (MAP) evaluation measure on the training query set. No query independent features are considered, i.e., $\omega = 0$ in Equation (1).

5. RESULTS

This section is structured as follows: Section 5.1 addresses the first research question, by reporting on the overall results, using only the target corpus (i.e., .GOV2 or CW09B) for global statistics; Section 5.2 focuses our investigation by analysing results based on query length; Later, in Section 5.3, we experiment with using an index of the Web for external corpus global statistics.

5.1 Target Corpus Global Statistics

Table 2 reports the results in terms of MAP and P@10 of our experiments on the 50 test queries for each setting (k, proximity approach, SD or FD, global statistics and corpus). In each setting, the best performance for each evaluation measure is highlighted. Moreover, if the other performance is statistically significant from the best performance (as per the Wilcoxon signed-rank test), it is denoted by * ($p < 0.05$) and ** ($p < 0.01$). For comparative baselines, the results of BM25 without applying any proximity weighting models are given in the first row.

Firstly, we note that, as expected, applying proximity improves the effectiveness of BM25 for both test collections in almost all settings. The only exception to this is for FD $k = 2$ for the CW09B corpus. We will return to this point later when we discuss full dependence in detail.

Next, we examine the effect of global statistics on the MRF proximity approaches. From the top half of Table 2, we note that from the results for both SD and FD for the MAP and P@10 measures, MRF using global statistics in addition to the local statistics shows no significant improvements over when local statistics are used alone. In fact, there are three cases where not using global statistics results in significantly higher performance ($k = 8$ for .GOV2, and both $k = 2$ and $k = 8$ for CW09B). For SD, MRF with and without global statistics performs similarly. However, for FD, while performances on .GOV2 are similar, for CW09B, MRF performs lower than the baseline without proximity, and significantly lower than MRF without global statistics. We believe this to be a form of overfitting. Consider that during training, if MRF using global statistics was harmful to retrieval effectiveness, then ϕ would receive a low weight. Instead, it appears that the usefulness of the global statistics differs between the training and test sets, and that

Corpus	Setting	TREC Numbers	Total	# of Queries by Length 1	2	3	4	> 4	Mean Length	# of Pairs SD	FD
GOV2	Train	701–800	100	1	29	45	22	3	2.97	197	326
GOV2	Test	801–850	50	1	14	22	13	0	2.94	97	158
CW09B	Train	20051–20210	100	13	39	37	8	3	2.49	149	228
CW09B	Test	20211–20290	50	9	20	17	4	0	2.32	66	95

Table 1: Details of the query sets used in our experiments. TREC query numbers, the number of queries broken down by length are shown, and the number of SD and FD pairs are shown.

	Window Size k	Global Stats.	Sequential Dependence (SD) .GOV2 MAP	.GOV2 P@10	CW09B MAP	CW09B P@10	Full Dependence (FD) .GOV2 MAP	.GOV2 P@10	CW09B MAP	CW09B P@10
BM25	-	-	0.2743	0.5383	0.1935	0.1830	0.2743	0.5383	0.1935	0.1830
MRF	2	✔	**0.2964**	0.5760	0.2271	**0.2200**	0.2763	0.5640	0.1505**	0.1540**
MRF	2	✘	0.2945	**0.5880**	**0.2292**	**0.2200**	**0.2819**	**0.5840**	**0.2276**	**0.2200**
MRF	8	✔	0.3012*	0.5780	**0.2302**	**0.2280**	0.3063	**0.6120**	0.2032**	0.1880**
MRF	8	✘	**0.3037**	**0.5960**	0.2293	0.2220	**0.3089**	0.6080	**0.2324**	**0.2260**
PL2	2	✔	**0.2998**	0.5840	**0.2299**	**0.2180**	0.2769	**0.5740**	0.1603**	0.1620**
BiL2	2	✘	0.2888*	**0.5920**	0.2178	0.2120	**0.2862**	0.5640	**0.1990**	**0.2080**
PL2	8	✔	**0.3024**	**0.5860**	**0.2304**	**0.2280**	**0.3048**	**0.6060**	0.2050	0.1900
BiL2	8	✘	0.2858**	0.5500*	0.2289	0.2200	0.2897	0.5740	**0.2105**	**0.2040**

Table 2: Results for the various proximity models, without global statistics and when using global statistics from the target collection. The best result in each setting is highlighted, with statistically significant different results from the best (as per the Wilcoxon signed-rank test) denoted by * ($p < 0.05$) and ** ($p < 0.01$).

FD is more sensitive to this than SD. Given the lack of benefit in retrieval performance brought by the global statistics, we suggest that it is safer to use models which use local statistics alone and do not consider F.

For the DFR proximity approaches, we compare BiL2, which does not consider global statistics, with PL2. For SD, BiL2 almost always performs worse than PL2, which does consider global statistics (a single exception is $k = 2$ for .GOV2 P@10 measure). However, only for MAP on the .GOV2 corpus are these differences statistically significant. This suggests that while the global statistics are slightly benefiting retrieval performance, any impact is minimal, and usually not significant. For FD, similar to MRF, global statistics do not provide any significant improvements for .GOV2. For CW09B, the PL2 results are inferior to not applying proximity, and also worse than BiL2 (significantly so for $k = 2$). Similar to MRF, we believe that overfitting is again occurring, because the retrieval performance is negatively impacted.

Comparing between the .GOV2 and CW09B test collections, we observe similar results for sequential dependence. For full dependence, where more pairs of query terms are considered (see Table 1), the 100 training queries for CW09B do not appear to provide a good indication of the quality of global statistics on the 50 test queries. Finally, comparing the window sizes $k = 2$ and $k = 8$, we note slightly higher overall performance for $k = 8$, in line with the results reported in [19].

Overall, we conclude that using local statistics alone is sufficient for effective retrieval, and that the presence of global statistics has little impact on the effectiveness of both MRF and DFR proximity approaches. In particular, only two small significant degradations in retrieval performance are observed when not using global statistics. Moreover, when global statistics are used, the resulting models are more likely to be overfitted and less robust, particularly for full dependence, where there are more pairs of query terms. These results are promising, as they indicate that effective, robust proximity weighting models can be implemented without need of provisions for global statistics.

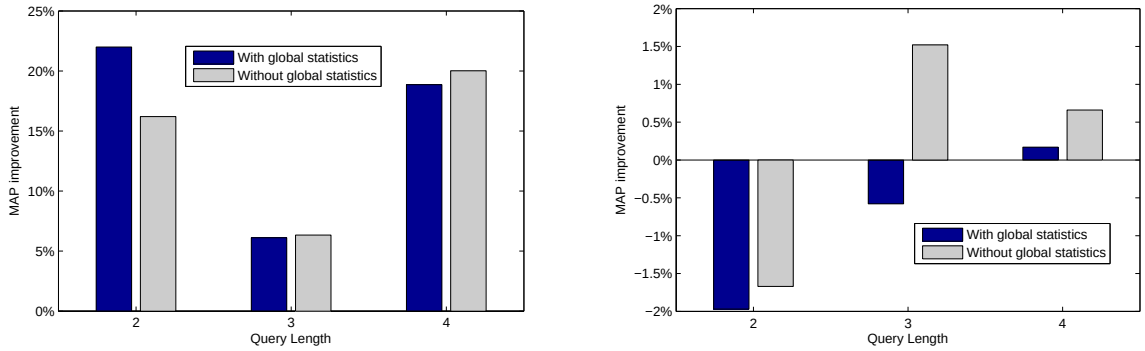
5.2 Query Length Analysis

In this section, we investigate our second research question, namely whether queries of different lengths are impacted differently by the presence of global statistics. In particular, for queries with more than one pair (i.e., queries with more than two query terms), the global statistics should assist in discriminating between occurrences of different pairs, as per the pair discrimination constraint. Moreover, the global statistics also play a role in measuring the likely usefulness of proximity weighting in conjunction to term weighting (see the corollary in Section 3).

In the following, we examine the improvement brought by the sequential and full dependence variants of MRF over the baseline for both the .GOV2 and CW09B corpus (results for $k = 2$ and $k = 8$ on both corpora are similar). We report only MRF results as those from DFR models are similar. In particular, we split the test topics into three different query sets by the length of the queries (see Table 1).

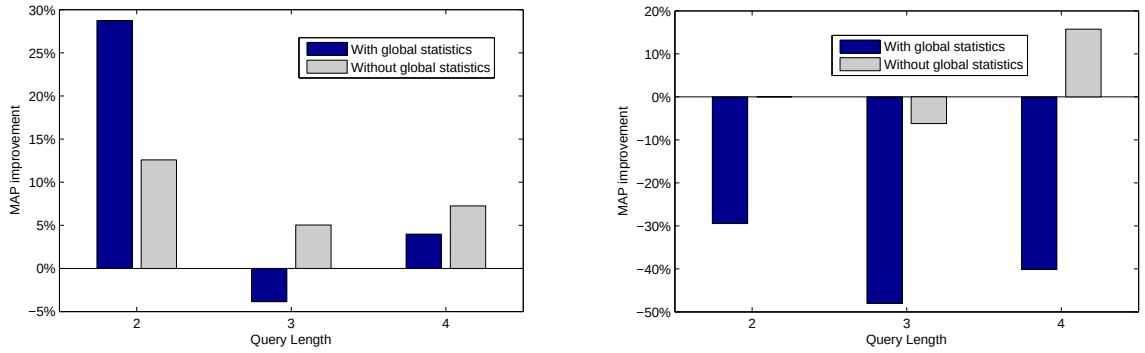
Figures 1 & 2 show the breakdown of relative improvement in MAP for queries of different lengths. From these figures, we observe that for .GOV2, global statistics are beneficial for two term queries. However, for queries of 3 or more terms, it is more effective not to use global statistics. For CW09B, the overfitting described in Section 5.1 ensures that using models without global statistics is always safer, particularly for longer queries. In general, the usefulness of the global statistics diminishes as the length of the query increases.

Indeed, the high performance of global statistics for queries of length 2 (i.e., a single pair) on .GOV2 illustrates that global statistics can play a role in balancing the proximity importance with that of the term weighting, as suggested by our corollary of the pair discrimination constraint. However, for longer queries with more pairs, finding a robust setting using global statistics is a problem, and reinforces our conclusion from Section 5.1 that it is safer to use models that do not consider global statistics. Hence, from these results, we conclude that the pair discrimination constraint does not appear to hold for proximity weighting models.



(a) .GOV2 (b) CW09B

Figure 1: Breakdown of performance for sequential dependence MRF, for $k = 2$.



(a) .GOV2 (b) CW09B

Figure 2: Breakdown of performance for full dependence MRF, for $k = 2$. MAP improvement for MRF without global statistics on CW09B, query length 2 is approx. 0%.

5.3 External Corpus Global Statistics

The Microsoft Web N-gram Service API [27] permits testing whether using a larger corpus can provide more refined global statistics for the proximity approaches. In particular, the Microsoft Web N-gram Service provides the (smoothed) probability of phrases (up to 5 terms in length) appearing on the Web, according to the index of the Bing search engine. Three different content types or fields are supported, namely estimates obtained from the statistics of the body of all documents, the titles of the documents, or the anchor text from all hyperlinks.

As the Microsoft Web N-gram Service is particularly new, its use in literature is somewhat sparse. We note the work of Huang et al. [11], which examined how the accuracy of some related IR tasks could be improved using the n-gram service, such as spelling correction and query segmentation. Indeed, while it is plausible query segmentation may be useful for proximity weighting, Huang et al. do not examine potential benefits to proximity weighting.

To test our third research question concerning the usefulness of global statistics obtained from a Web-scale corpus, we replace probability $\frac{F}{T}$ in Equation (3) for a pair of query terms with the probability as reported by the n-gram service for that pair. However, we note that our experiments thus far have used stemmed query terms, while the n-gram service only provides global statistics estimates for unstemmed phrases. To account for this, we replace the probability of the stemmed pair of query terms with that of the corresponding unstemmed pair of terms from the original query.

Moreover, as the n-gram service does not support wildcards, nor phrases of length 8, we restrict our experiments to $k = 2$. Furthermore, for full dependence, the ordering of the occurrence of a pair of query terms in each document is not considered. However, the n-gram service only provides ordered probabilities (i.e., $P(\text{“}a\,b\text{”}) \neq P(\text{“}b\,a\text{”})$). For this reason, we sum the probabilities of both orderings of each pair. Finally, we note that PL2 (Equation (9)) does not model directly $\frac{F}{T}$ (indeed, it models $\frac{F}{N}$), hence, the DFR proximity approaches are excluded from this experiment.

We experiment with all three content types (body, title and anchor text). We note that the document bodies are the largest overall content and hence may provide the most accurate global statistics. However, it also is plausible that particularly important phrases may be easier to identify using the title or anchor text statistics, because these content types typically consist of noun group phrases rather than large passages of text.

Table 3 presents the results of using the Web N-gram Service results from Bing for global statistics. Results are broken down by content type of the global statistics (body, title and anchor text). Results from Table 2 for BM25, MRF without global statistics (denoted None), and MRF using the global statistics of the target corpora (.GOV2 and CW09B) are also provided as baselines. From Table 3, we draw several observations. Firstly, it is apparent that when MRF uses the global statistics obtained from the Web, retrieval effectiveness is not enhanced compared to MRF that uses local statistics alone. Indeed, the MRF using Web global statistics

	Sequential Dependence				Full Dependence			
	.GOV2		CW09B		.GOV2		CW09B	
Global Stats.	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
(BM25)	0.2743	0.5383	0.1935	0.1830	0.2743	0.5383	0.1935	0.1830
None	0.2945	**0.5880**	**0.2292**	**0.2200**	**0.2819**	**0.5840**	**0.2276**	**0.2200**
Target	**0.2964**	0.5760	0.2271	**0.2200**	0.2763	0.5640	0.1505**	0.1540**
Bing Body	0.2697**	0.5420**	0.2126	0.1980	0.2680**	0.5480**	0.1807**	0.2040**
Bing Title	0.2654**	0.5280**	0.2182	0.2040	0.2647**	0.5380**	0.1790**	0.2060**
Bing Anchor Text	0.2675**	0.5320**	0.2174	0.1940	0.2672**	0.5460**	0.1753**	0.1940**

Table 3: MRF $k = 2$ when using various corpora for global statistics. Best performance in each column is highlighted. Statistically significant performances from the best in each column are denoted using * and ** as before.

performs significantly worse for all settings except CW09B for SD. Comparing to MRF using global statistics from the target corpora, we see that Web global statistics also perform poorer, except for FD on CW09B. We note that this setting exhibited poor retrieval performance in Table 2 above, and for this setting, the n-gram service statistics are more robust, yet still underperform compared to the BM25-only baseline.

Overall, we find that using only local statistics is still sufficient for effective retrieval, i.e. there is still no benefit in applying global statistics, even when they are obtained from the Web instead of the target corpus. The low performance of the global statistics obtained from the n-gram service might be due to the conversion of unstemmed statistics to a stemmed environment. We consider unstemmed proximity retrieval beyond the scope of this paper, and we leave it for future work.

6. CONCLUSIONS

The IDF component is an important aspect of all term weighting models. However, its benefit for proximity weighting models is unclear – i.e., if the global statistics of pairs of query terms is an important feature. We refer to the use of these global statistics as the pair discrimination constraint. In this paper, we examined the importance of global statistics for two statistically different and effective proximity approaches, namely Markov Random Fields language modelling and Divergence from Randomness-based proximity weighting models.

Through experiments on two large-scale TREC corpora, we compare proximity models with and without the use of global statistics. We found that proximity using only local document-level statistics was sufficient for effective retrieval. Indeed, while the global statistics are expensive to compute, they rarely led to significant improvements in retrieval effectiveness, while their usefulness decreased as queries becomes longer. Finally, using a Web-scale corpus to estimate the global statistics did not lead to improvements in retrieval effectiveness. Overall, the results in this paper suggest that the pair discrimination constraint is not a necessary feature for an effective proximity weighting model.

7. REFERENCES

[1] G. Amati. *Probabilistic Models for Information Retrieval based on Divergence from Randomness*. PhD thesis, University of Glasgow, 2003.

[2] G. Amati, C. Carpineto, and G. Romano. Italian monolingual information retrieval with PROSIT. In *Proceedings of CLEF 2002*, pages 257–264.

[3] J. Bai, Y. Chang, H. Cui, Z. Zheng, G. Sun, and X. Li. Investigation of partial query proximity in web search. In *Proceedings of WWW 2008*, pages 1183–1184.

[4] A. Z. Broder, D. Carmel, M. Herscovici, A. Soffer, and J. Zien. Efficient query evaluation using a two-level retrieval process. In *Proceedings of CIKM 2003*, pages 426–434.

[5] S. Büttcher, C. Clarke, and I. Soboroff. The TREC 2006 Terabyte track. In *Proceedings of TREC 2006*.

[6] S. Büttcher, C. Clarke, and B. Lushman. Term proximity scoring for ad-hoc retrieval on very large text collections. In *Proceedings of SIGIR 2006*, pages 621–622.

[7] B. Carterette, H. Fang, V. Pavlu, and E. Kanoulas. Million query track 2009 overview. In *Notebook of TREC 2009*.

[8] B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley, 2009.

[9] R. Cummins and C. O'Riordan. Learning in a pairwise term-term proximity framework for information retrieval. In *Proceedings of SIGIR 2009*, pages 251–258.

[10] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *Proceedings of SIGIR 2004*, pages 49–56.

[11] J. Huang, J. Gao, J. Miao, X. Li, K. Wang, F. Behr, and C. L. Giles. Exploring web scale language models for search query processing. In *Proceedings of WWW 2010*, pages 451–460.

[12] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[13] C. Lioma, C. Macdonald, V. Plachouras, J. Peng, B. He, and I. Ounis. University of Glasgow at TREC 2006: Experiments in Terabyte and Enterprise tracks with Terrier. In *Proceedings of TREC 2006*.

[14] D. Metzler. Personal communication, 2009.

[15] D. Metzler and W. B. Croft. A Markov random field model for term dependencies. In *Proceedings of SIGIR 2005*, pages 472–479.

[16] G. Mishne and M. de Rijke. Boosting web retrieval through query operations. In *Proceedings of ECIR 2005*, pages 502–516.

[17] M. Mitra, C. Buckley, A. Singhal, and C. Cardie. An analysis of statistical and syntactic phrases. In *Proceedings of RIAO 1997*, pages 200–214.

[18] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A high performance and scalable information retrieval platform. In *Proceedings of the OSIR Workshop 2006*, pages 18–25.

[19] J. Peng, C. Macdonald, B. He, V. Plachouras, and I. Ounis. Incorporating term dependency in the DFR framework. In *Proceedings of SIGIR 2007*, pages 843–844.

[20] V. Plachouras and I. Ounis. Multinomial randomness models for retrieval with document fields. In *Proceedings of ECIR 2007*, pages 28–39.

[21] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C: The Art of Scientific Computing, 2nd ed.* Cambridge University Press, 1992.

[22] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. Okapi at TREC. In *Proceedings of TREC 1*, pages 21–30, 1992.

[23] R. Schenkel, A. Broschart, S. Hwang, M. Theobald, and M. Gatford. Efficient text proximity search. In *Proceedings of SPIRE 2007*, pages 287–299.

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

[25] K. M. Svore, P. H. Kanani, and N. Khan. How good is a span of terms? Exploiting proximity to improve web retrieval. In *Proceedings of SIGIR 2010*.

[26] H. Turtle and J. Flood. Query evaluation: strategies and optimizations. *Inf. Process. Manage.*, 31(6):831–850, 1995.

[27] K. Wang, C. Thrasher, E. Viegas, X. Li, and B.-j. P. Hsu. An overview of Microsoft web n-gram corpus and applications. In *Proceedings of NAACL-HLT 2010 Demos*, pages 45–48.

[28] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR 2001*, pages 334–342.

[29] M. Zhu, S. Shi, M. Li, and J.-R. Wen. Effective top-k computation in retrieving structured documents with term-proximity support. In *Proceedings of CIKM 2007*, pages 771–780.

[30] M. Zhu, S. Shi, N. Yu, and J.-R. Wen. Can phrase indexing help to process non-phrase queries? In *Proceedings of CIKM 2008*, pages 679–688.