Enhancing Relevance Scoring with Chronological Term Rank

Adam D. Troy Guo-Qiang Zhang
Department of Electrical Engineering and Computer Science
Case Western Reserve University
10900 Euclid Avenue
Cleveland, Ohio USA
adam.troy@case.edu gqz@eecs.case.edu

ABSTRACT

We introduce a new relevance scoring technique that enhances existing relevance scoring schemes with term position information. This technique uses *chronological term rank* (CTR) which captures the positions of terms as they occur in the sequence of words in a document. CTR is both conceptually and computationally simple when compared to other approaches that use document structure information, such as term proximity, term order and document features. CTR works well when paired with Okapi BM25. We evaluate the performance of various combinations of CTR with Okapi BM25 in order to identify the most effective formula. We then compare the performance of the selected approach against the performance of existing methods such as Okapi BM25, pivoted length normalization and language models. Significant improvements are seen consistently across a variety of TREC data and topic sets, measured by the major retrieval performance metrics. This seems to be the first use of this statistic for relevance scoring. There is likely to be greater retrieval improvements possible using chronological term rank enhanced methods in future work.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models,Search process*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation*

General Terms

Algorithms, Experimentation, Performance

Keywords

Relevance ranking, term position, similarity scoring, document structure, chronological term rank, term weighting

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.

1. INTRODUCTION

In the beginning, there was term frequency. As many have pointed out [4, 7], the notion of utilizing term frequencies in order to estimate term significance was asserted by Luhn as early as the late 1950s [11]:

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

Following the first element of Luhn's proposal, term frequency based methods have since become the benchmarks by which new work in relevance scoring is judged, including the work presented here. Two term frequency statistics are most commonly used. The first, referred to simply as *term frequency*, is the number of occurrences of a term in a particular document, denoted here as tf. The second, the number of documents in the collection containing the term of interest is generally referred to as *document frequency* and denoted as df. Document frequency is most commonly used in term weights as inverse document frequency (idf), as proposed by Jones in 1972 [9].

Two ($tf \cdot idf$)-based relevance estimation techniques have become particularly dominant: Okapi BM25 [15] and pivoted length normalization [16]. A popular formulation of the Okapi BM25 model is shown in Eq. 1.

$$\sum_{t \in d \cap q} ln \frac{N - df + 0.5}{df + 0.5} \cdot \frac{tf}{0.5 + 1.5 \cdot \frac{dl}{avdl} + tf} \quad (1)$$

In addition to tf and df as defined above, t is a term in the query q and document d, N is the total number of documents in the collection, dl is the length of d and $avdl$ is the average length of all documents in a collection. One version of the pivoted normalization scheme is shown in Eq. 2, where qtf is the number of occurrences of t in q.

$$\sum_{t \in d \cap q} \frac{1 + ln(1 + ln(tf))}{0.8 + 0.2 \frac{dl}{avdl}} \cdot qtf \cdot ln \frac{N + 1}{df} \quad (2)$$

Three common features of these techniques are worth noting. First, term frequency is used as the core indicator of

document relevance. Second, the importance of a term is inversely related to its commonality: rare terms are more useful indicators of relevance. Finally, document length is utilized to correct for the greater likelihood of retrieving longer documents simply because they contain a greater number of words.

While term frequency is certainly one of the most useful indicators of relevance, relying solely on term frequency amounts to viewing a document simply as a *bag-of-words* and ignores much of its structure. With this in mind, researchers have developed techniques incorporating document structure information.

The second element of Luhn's proposal, that *term position* is also important, has historically seen only intermittent interest. With the growing difficulty of achieving further retrieval improvements using only term frequencies, there is an increasing interest in information derived from document structure.

Term order is an example of structural information that some recent approaches utilize. Such approaches assign higher relevance scores to documents in which query terms appear in the same order as they appear in the query.

Document-feature based approaches take into account the occurrence of query terms in locations of varying prominence, such as title, header text, or body text. The most prominent document-feature based approach is BM25F, built upon Okapi BM25 [14, 19]. Document feature techniques are most commonly used in web search as features can be easily identified due to the presence of HTML tags.

Term proximity, which is most closely related to the work presented here, is an idea which has seen recent interest. In this context, term proximity refers to the lexical distance [7] between query terms, calculated as the number of words separating query terms in a document. Keen carried out some of the earliest investigations of using term proximity in the 90s [10]. Hawking and Thistlewaite provided a framework for formal discussion of proximity based methods [7]. Rasolofo and Savoy combined proximity information with Okapi and found improved retrieval performance, particularly among the top scoring documents [13]. Beigbeder and Mercier achieved improved precision through the use of fuzzy proximity statistics [4]. Büttcher, Clarke and Lushman also added proximity information to Okapi BM25, with positive results [5]. Many research groups now use proximity enhanced approaches as part of their TREC submissions [18]. Modern experimental retrieval systems such Indri support proximity query operators [17].

In this work, we examine the use of a relatively neglected term position statistic: chronological term rank (CTR). We were not able to locate any similar work in the literature. Chronological term rank is the rank obtained from the position of the term in the sequence of words in the original document. Throughout this paper the phrase "term rank" refers to chronological term rank unless otherwise indicated. Intuition indicates that this statistic may be useful; terms most strongly related to the main content of a document are likely to occur near the beginning. We augment Okapi BM25 with CTR, resulting in a hybrid scoring method that improves retrieval performance across a variety of TREC data and topic sets, measured by major retrieval performance metrics.

In the next section we introduce chronological term rank, our chronological term rank model and discuss some functional considerations for its use. In Section 3 we experimentally evaluate these considerations in order to identify an optimal approach. In Section 4 we compare the optimized approach with several current benchmark relevance estimation methods on a variety of data and topic sets. We end with some concluding remarks.

2. CHRONOLOGICAL TERM RANK

Just as document-structure based methods move beyond the traditional *bag-of-words* approaches to relevance estimation, so does our method, utilizing chronological term rank.

The chronological rank of a term (CTR) within a document is defined as the rank of the term in the sequence of words in the original document. We refer to this statistic as "chronological" to emphasize its correspondence to the sequential occurrence of the terms within the document from the beginning to end. We also use this name to differentiate it from the many other uses of "rank" in the literature, such as Anh and Moffat's frequency-based term rank [2].

Intuitively, CTR is the sequence of words as they are encountered as the document is read from start to finish. This intuition is very simple; perhaps even more clear and succint than that behind term proximity. Authors tend to state the main ideas and topics of a piece as early as possible to quickly convey core ideas and results and grasp attention. This practice is explicitly reflected in titles, the use of abstracts in scientific papers and the inverted-pyramid style of writing favored by journalists [6]. Journalism students are taught to place the most important and relevant content at the beginning of an article while progressively including less important and smaller details as the article continues. As such, important terms may be placed at the beginning of a document and may not be explicitly mentioned again, potentially spoiling the correlation between term frequency and the importance of the term. Terms mentioned many times, but only at the end of a document, may not be as important. Chronological term rank systematically and elegantly incorporates these ideas by utilizing one of the simplest possible structural observations.

The CTR values of each term in the query are independent of one another. This differs from other structural approaches such as term proximity or term order, making CTR easier to incorporate into existing systems. Indexing systems need only to store one additional integer along with each term frequency if document-term impact scores are not computed at indexing time. Fully determining order or proximity between all occurrences of query terms within a document or using document features can be expensive in both storage space and computation time. Existing term frequency based systems would likely need to undergo significant modifications to utilize structural information other than CTR.

2.1 Chronological Term Rank Model

The CTR model is defined as follows: let $D = (t_1, \ldots, t_n)$ be a document where t_i are terms (words) ordered according to their sequence in the original document. Let $tr_t := i$, where the chronological rank tr of term t is assigned as the subscript i of the *earliest occurrence* of t in D.

2.2 Functional Considerations

How should CRT be utilized? In this subsection we describe some of the approaches we studied. This is by no means exhaustive. Many other uses of CTR may be pos-

sible and researchers may develop more effective schemes incorporating CTR in the future. As with other successful structural approaches, we use Okapi BM25 as the base for our ranking functions, adding an additional term which incorporates CTR. We use BM25 as formulated in Eq. 1. In our preliminary testing, the pivoted length normalization formula did not integrate with term rank successfully, at least for the approaches we attempted. Thus we will focus on the integration of CTR with BM25, with the following considerations.

- **Multiplicative vs. Additive** — Term rank can either be incorporated into BM25 as a scaler (multiplicative) term or as an additional weighted term (additive). This is related to the relationship which term rank plays with the *tf* term.
- **Absolute vs. Percentage** — Term rank can be used as an inverse of the absolute rank, such as $\frac{C}{tr}$ or as rank percentage, $\frac{tr}{dl}$, for example.
- **Document Length** — If document lengths are used, such as in a rank percentage, it may be beneficial to use the maximum document length of the collection in order to improve the chances of short documents being retrieved.
- **Log Values** — Term rank may be more effective as a logarithm, for example if differences between smaller ranks are more interesting.
- **Range Variance** — Retrieval performance may be improved by limiting the range of the total impact of the score that the CTR affects.
- **Stopwords** — Term rank may be more effective if stopwords are ignored in the CTR calculation.

3. EXPERIMENTS

In this section we experimentally evaluate the functional considerations discussed in the previous section. Our goal is to construct the most effective relevance ranking formula using chronological term rank. In addition we will be able to identify the most effective range of parameters for the constructed function. Again, these experiments are by no means exhaustive. We are simply exploring some readily available avenues towards the effective use of chronological term rank.

Evaluation of each of the above considerations cannot be done completely independently. A complete term rank enhanced relevance function cannot be constructed without making some choices for each consideration. With that said, we attempt to keep the experiments as independent as possible.

In each of the following experiments the dataset used is *WSJ2*, which consists of the Wall Street Journal (1990–1992) portion of disk two of the TREC corpus. The topic set is the title fields from TREC topics 51–200. This data and topic set should be fairly good as an indicator of potential effectiveness for other data and topic sets; others have used it for similar purposes [2]. We focus on mean average precision (MAP), generally viewed as the most useful single measure of retrieval effectiveness.

We used our own retrieval system for this work, following standard guidelines in implementation [3]. As benchmarks, our implementations of BM25 and pivoted length normalization achieve MAP of 0.1478 and 0.1506 respectively for the selected data and topic sets. These results are consistent with other published results.

Formulae evaluated in the following subsections are compiled in Table 1 for easy reference. Specifically, a complete formula is constructed by selecting either base formula A or B and one of the CTR formulae, $\mathcal{R}$, a–j. For each formula to be evaluated, we measure MAP across the effective range of the parameter associated with the particular feature being examined.

Multiplicative vs. Additive

Here we evaluate whether chronological term rank is more effective as a multiplicative or additive supplement to BM25. This will determine whether the term rank statistic should be a scaler on the *tf* term or whether it is more effective as an independent term. Multiplicative use is shown in Eq. 3 where $\mathcal{R}$ is the CTR term (formula A in Table 1). Likewise, additive use is shown in Eq. 4 (formula B in Table 1).

$$\sum_{t \in d \cap q} ln \frac{N - df + 0.5}{df + 0.5} \cdot \frac{tf}{0.5 + 1.5 \cdot \frac{dl}{avdl} + tf} \cdot \mathcal{R} \quad (3)$$

$$\sum_{t \in d \cap q} ln \frac{N - df + 0.5}{df + 0.5} \cdot \left(\frac{tf}{0.5 + 1.5 \cdot \frac{dl}{avdl} + tf} + \mathcal{R} \right) \quad (4)$$

We evaluated a total of three functions. An additive function, specifically Eq. 4 where $\mathcal{R}$ is equal to $C \cdot (1 - \frac{tr-1}{dl})$ (formula [B,a]), essentially a rank percentage scaled by a constant, C. Two multiplicative functions were evaluated. Both are built upon Eq. 3 using different $\mathcal{R}$ terms. The first, referred to as `Multiplicative1` below, where $\mathcal{R} = 1 + (C \cdot (1 - \frac{tr-1}{dl}))$ scales the score for each term by a value between 1 and $1+C$, according to the percent rank (formula [A,b]). The second, referred to as `Multiplicative2`, where $\mathcal{R} = (1 - C) + (C \cdot (1 - \frac{tr-1}{dl}))$, scales a portion of the term score according to the term rank percentage (formula [A,c]), for a final value between $1 - C$ and 1.

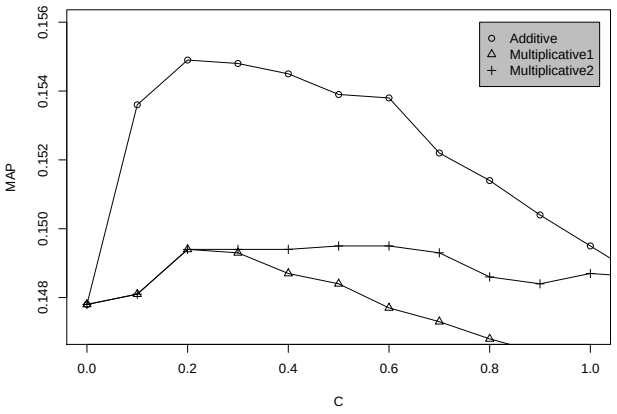

Figure 1: Comparison of MAP for additive and multiplicative term rank methods ([B,a],[A,b],[A,c]).

The MAP of these functions over varying C values is shown in Fig. 1. It is clear that the additive formula, with peak MAP around $C = 0.2$, is more effective than the multiplicative formulae. The multiplicative functions provide only a slight boost in MAP. Throughout the rest of these experiments, Eq. 4 is used as the base formula.

Absolute vs. Percentage

We next determined whether absolute rank or rank percentage is more effective. In particular, for the percentage method we used the same formula as the additive function in the previous subsection (formula [B,a]). For absolute rank functions we use Eq. 4 as the base with $\mathcal{R}$ equal to $C \cdot \frac{1}{tr}$ (formula [B,d]) and $C \cdot \frac{1}{log(tr)}$ (formula [B,e]).

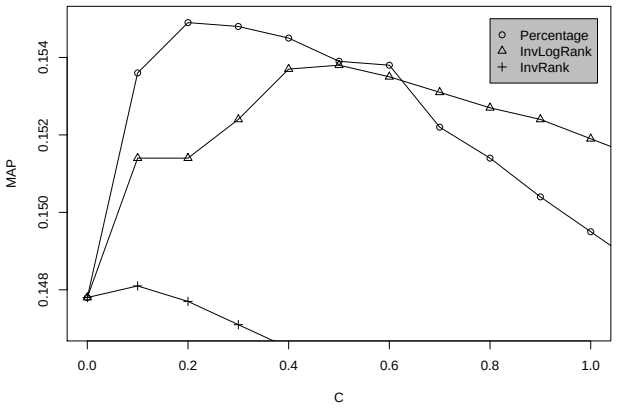

Figure 2: Comparison of MAP for percentage and absolute term rank methods ([B,a],[B,e],[B,d]).

MAP for each of these three formulae over varying C values is shown in Fig. 2. The percentage formula handily outperforms the two inverse absolute formulae. The non-log inverse rank formula provides virtually no improvement at all. The percentage formula also has the advantage of being less temperamental over different values of C.

Document Length

Perhaps a term occurring at the end of a short document is more indicative of relevance than a term occurring at the end of a long document. If this is the case, individual rank percentages may not be optimal. In order to evaluate this hypothesis, we again use formula [B,a] as the benchmark compared to base formula B with $\mathcal{R} = C \cdot (1 - \frac{tr-1}{maxdl})$ where $maxdl$ is the length of the longest document in the collection (formula [B,f]). This will give shorter documents somewhat of a boost in score.

The comparison of MAP for these two formulae over varying C values are shown in Fig. 3. The maximum document length formula outperforms the individual length formula with its maximum MAP at about $C = 0.3$, though for larger values of C the individual formula performs slightly better.

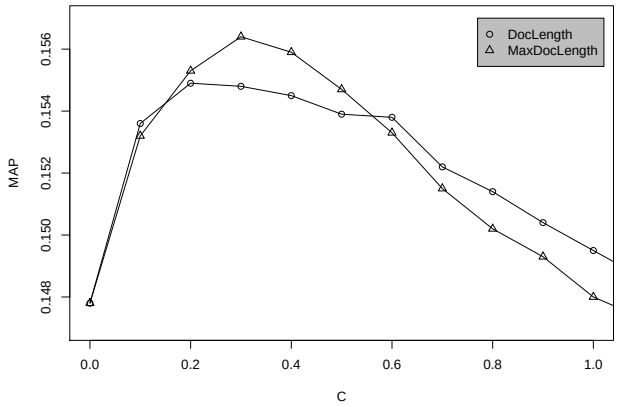

Figure 3: Comparison of MAP for term rank percentage methods using individual document length and max document length ([B,a],[B,f]).

Log Values

Logarithms of term rank statistics may improve effectiveness. For instance, differences in term ranks may be most important in the beginning portions of documents. In order to evaluate this, we compare formulae [B,g], [B,h] and [B,i] which use logarithms of term rank and document length statistics. Formula [B,g] uses the logarithm of both term rank and document length. Formula [B,h] uses the logarithm of scaled term rank and scaled document length, with the scaler being $\frac{1}{D}$. The scaler helps the rank statistics to better fit into the useful range of the logarithm. Formula [B,i] uses the maximum document length of the collection rather than individual document length. These log values do not change the magnitude of the rank term, rather as part of a percentage they change the shape of the function over the range of term ranks.

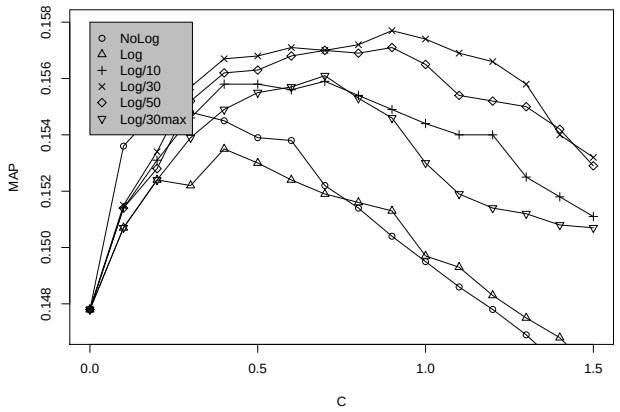

Figure 4: Comparison of MAP for term rank methods using log values ([B,a],[B,g],[B,h],[B,i]).

Figure 4 compares the MAP of theses formulae using various constants. `NoLog` is, again, formula [B,a]. `Log` is formula [B,g], `Log/10`, `Log/30`, `Log/50` are formula [B,h] with D equal to 10, 30 and 50 respectively. `Log/30max` is formula [B,i] with D equal to 30. `Log/30` achieves the highest MAP. Interestingly, `Log/30max`, which uses *maxdl*, did not perform as well as would be expected based on the performance of the `MaxDocLength` method in the previous subsection. Overall, the log statistics provided significant improvement in MAP. Unintentionally, the way the logarithms used here gives similar rank percentages higher scores as if they occur in longer documents. This also likely contributed to the performance. This indicates that CTR may be a more useful statistic for determining relevance of longer documents.

$\mathcal{R}$ Range Variance

Next, we determine whether retrieval can be improved by limiting the range of C over which CTR has an effect. The way the logarithms are used above, for example, limits the effective range of the term rank percentage. This may be partly the cause of the improved performance of those functions. Here we will examine that notion further. This is done by using formula [B,j] where D determines the percentage of C which is affected by the term rank percentage, for a final term value between $C \cdot (1 - D)$ and C where D is some value from 0–1.

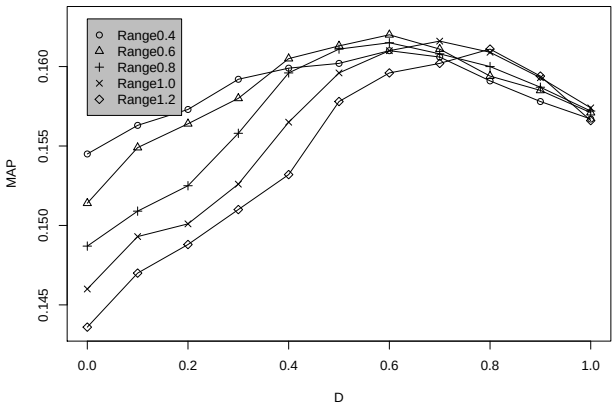

Figure 5: Comparison of MAP for term rank methods with limited C range variance ([B,j]).

The comparison of the MAP of this formula using several values of C and D is shown in Fig. 5. Overall, we see a dramatic improvement using this formula. In particular we see the maximum MAP at $C = 0.6$ and $D = 0.6$, though the peak MAP for each of the C values is not dramatically different.

Stopwords

Finally, we evaluate whether the inclusion or exclusion of stopwords in CTR calculation has any effect on retrieval performance. Specifically, when encountering a stopword while indexing, should the rank count be incremented or not? The stopword list we use throughout this work is `stoplist.orig` from Zobel[1] which contains 600 common words. Formula [B,a] was used with and without stopwords in CTR calculation.

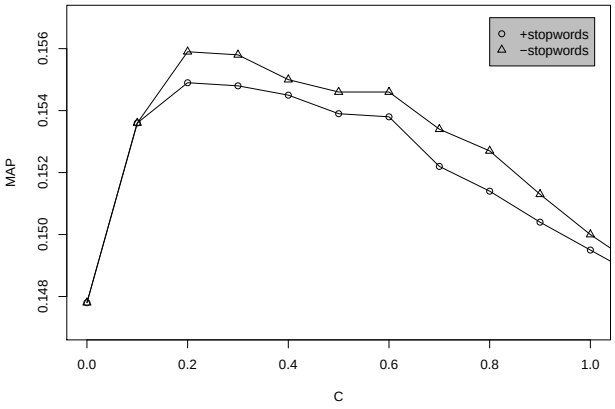

Figure 6: Comparison of MAP for term rank method with stopwords included and not included in rank calculation ([B,a]).

The comparison of MAP for formula [B,a] with and without stopwords as part of the CTR calculation is shown in Fig. 6 over varying C values. The exclusion of stopwords in rank calculations consistently outperforms rank that includes stopwords. The exclusion of stopwords likely produces a more truthful relative ordering of terms in the documents.

Based on the results of these experiments, we discovered that the formula with the highest mean average precision uses base formula Eq. 4 with $\mathcal{R}$ equal to Eq. 5 below, with stopwords ignored in chronological term rank calculations.

$$\mathcal{R} = C - \left(C \cdot D \cdot \frac{log(\frac{tr-1}{30} + 10)}{log(\frac{dl}{30} + 10)} \right) \quad (5)$$

Intuition can provide clues as to why these functional choices result in top performance. The optimal performance of additive usage suggests that term rank is an independent relevance bearing statistic, rather than a modifier of term frequency. The superiority of rank percentage with individual document lengths indicates that relative rank is more useful than absolute rank. Log values put more emphasis on the differences of CTR in the early portion of the documents with those in the latter being more similar to one another, with a greater emphasis on ranks in longer documents. The limited range variance provides a small base value to the rank term. Computing rank without stopwords likely provides a more accurate chronological ordering of terms. Interestingly, Eq. 5 alone along with *idf* weights from BM25 produces MAP of 0.1199, a decent score, on the selected data and topic sets, which is suggestive of the utility of chronological term rank as a relevance indicator.

[1] `http://goanna.cs.rmit.edu.au/~jz/resources/stopping.zip`

Base Formulae

	Formula	Description
A	$\sum_{t \in d \cap q} ln\frac{N-df+0.5}{df+0.5} \cdot \frac{tf}{0.5+1.5\cdot\frac{dl}{avdl}+tf} \cdot \mathcal{R}$	Multiplicative
B	$\sum_{t \in d \cap q} ln\frac{N-df+0.5}{df+0.5} \cdot \left(\frac{tf}{0.5+1.5\cdot\frac{dl}{avdl}+tf} + \mathcal{R}\right)$	Additive

$\mathcal{R}$

	Formula	Description
a	$C \cdot (1 - \frac{tr-1}{dl})$	Rank Percentage
b	$1 + (C \cdot (1 - \frac{tr-1}{dl}))$	Rank Percentage Multiplicative Boost
c	$(1-C) + (C \cdot (1 - \frac{tr-1}{dl}))$	Rank Percentage Multiplicative Scale
d	$C \cdot \frac{1}{tr}$	Inverse Rank
e	$C \cdot \frac{1}{log(tr)}$	Inverse Log Rank
f	$C \cdot (1 - \frac{tr-1}{maxdl})$	Rank Maximum Percentage
g	$C \cdot (1 - \frac{log(tr+9)}{log(dl+10)})$	Rank Log Percentage
h	$C \cdot (1 - \frac{log((tr-1/D)+10)}{log((dl/D)+10)})$	Scaled Rank Log Percentage
i	$C \cdot (1 - \frac{log((tr-1/D)+10)}{log((maxdl/D)+10)})$	Scaled Rank Log Maximum Percentage
j	$C - (C \cdot D \cdot \frac{log((tr-1/30)+10)}{log((dl/30)+10)})$	Range Limited Scaled Rank Log Percentage

Table 1: Components for chronological term rank scoring formulae.

4. PERFORMANCE COMPARISON

In this section we compare the performance of our chronological rank method against established benchmark ranking formulae on larger data and topic sets. In particular, we compare against the performance of pivoted length normalization [16] and Okapi BM25 [15] as implemented in our system as well as the published performance of language model, title language model of Jin, Hauptmann, and Zhai [8], and the Global-By-Value and Local-By-Rank methods of Anh and Moffat [1, 2]. Variables such as stemming, stopwords and parsing were identical for all the methods implemented in our system. We used the Porter stemmer [12] and the stopword list indicated in Section 3. Performance of the BM25 and pivoted normalization formulae as implemented in our system is consistent, though not identically, to published performance of the same algorithms [2, 8]. The specific CTR formula used for this comparative study is Eq. 4 with $\mathcal{R}$ equal to Eq. 6, below. Stopwords were ignored in term rank calculations. This differs from Eq. 5 in the statistic scaler, which is 20 rather than 30. This difference makes the formula more effective when stopwords are ignored in term rank computation.

$$\mathcal{R} = C - \left(C \cdot D \cdot \frac{log(\frac{tr-1}{20}+10)}{log(\frac{dl}{20}+10)}\right) \quad (6)$$

Two datasets and four topic sets were used for the first set of comparisons. The *TREC12* dataset consists of disks one and two of the TREC corpus. *TREC45-CR* consists of disks four and five of the TREC collection excluding the congressional report documents. Title queries from TREC topics 51–200 were run against the *TREC12* dataset. Three topic sets were run against *TREC45-CR*: the title fields from topics 401–450 (TREC-8 ad hoc track), title fields from topics 351–450 (TREC-7 and TREC-8 ad hoc tracks), and the title fields from topics 301–450 and 601–700 (TREC 2004 robust track). These data and topic sets should be of sufficient size and contain sufficient variety of queries, along with other comparisons below, to provide a conclusive evaluation of our CTR approach.

A summary of the performance of CTR, pivoted length normalization and BM25 on each of the data and topic sets above is shown in Table 2. Three metrics are shown: mean average precision (MAP), precision after 10 documents retrieved (P@10) and reciprocal rank (R. Rank). The column which is labeled CTR is the score of the chronological term rank method for the indicated metric. The next three columns contain the scores of pivoted length normalization, percent improvement of our method over pivoted length normalization and whether the difference is significant according to the Wilcoxon sign-rank test at 95% confidence. The next three columns are similar but refer to Okapi BM25. The amount of improvement of CTR over both pivoted length normalization and BM25 across the data and topic sets for the three retrieval metrics is striking. The differences are significant in all cases.

We next compared the performance of CTR with the published results of several other scoring schemes. First we compared against the recent performance of Anh and Moffat's Local-By-Rank and Global-By-Value scoring schemes [2, 1].

Collection	Query Set	Metric	CTR	Pivoted			BM25		
			Score	Score	%△	Sig.	Score	%△	Sig.
TREC12	051–200	MAP	0.2574	0.2339	+10.0	•	0.2325	+10.7	•
		Prec@10	0.5149	0.4564	+12.8	•	0.4752	+8.4	•
		R. Rank	0.7011	0.6288	+11.5	•	0.6255	+12.1	•
TREC45-CR	401–450	MAP	0.2632	0.2287	+14.5	•	0.2237	+17.7	•
		Prec@10	0.4740	0.4500	+5.3	•	0.4480	+5.8	•
		R. Rank	0.6455	0.6177	+4.5	•	0.5844	+10.5	•
TREC45-CR	351–450	MAP	0.2331	0.2036	+15.1	•	0.2013	+15.8	•
		Prec@10	0.4590	0.4180	+9.8	•	0.4310	+6.5	•
		R. Rank	0.7011	0.6184	+13.4	•	0.6142	+14.1	•
TREC45-CR	Robust04	MAP	0.2628	0.2318	+13.4	•	0.2327	+12.9	•
		Prec@10	0.4422	0.4060	+8.9	•	0.4181	+5.8	•
		R. Rank	0.7151	0.6420	+11.4	•	0.6433	+11.2	•

Table 2: Performance of CTR compared with pivoted length normalization and Okapi BM25 methods. Scores, percent improvement by CTR and significance are shown for multiple document and topic sets.

Datasets used for this comparison are *TREC12* and *TREC45-CR* as described previously. Title fields from topics 51–200 were run against *TREC12* while the title fields of 351-450 were run against *TREC45-CR*. Summary of results for these runs is shown in Table 3. Again, statistics include mean average precision (MAP), precision after 10 documents retrieved (P@10) and reciprocal rank (R. Rank). The score for each compared method is shown, as well as the percent of improvement made by CTR. Significance tests were not possible due to the lack of availability of individual query results. The improvements made here are again striking, though in some cases less so than the previous comparisons. MAP improvements are the most consistent. It should be kept in mind that these are not perfect comparisons; some performance differences may be due to unknown differences in parsing or stemming, for instance.

Finally, we compared our approach with the published performance of BM25, traditional language model (LM), and title language model (TLM) from Jin et al. [8]. For these comparisons four smaller TREC datasets were used: the Associated Press portions of disks two and three (*AP2,AP3*), San Jose Mercury News from disk three (*SJM*) and Wall Street Journal from disk two (*WSJ2*). The topic set for these comparisons was the description fields from TREC topics 201–250. This topic set is somewhat different than the ones used previously. In particular, the queries are much longer and consist of a complete sentence rather than a few words. The MAP of runs of the topic set against each dataset with each scoring method is shown in Table 4.

Here again we see dramatic improvement using CTR. The title language model performs most similarly to CTR, with CTR only performing slightly better on the *SJM* dataset, for instance. As with the previous comparisons, these are not perfect due to unknown retrieval system differences.

Overall, we can conclude from these comparisons that CTR provides significant improvements in retrieval performance compared to several traditional and more recent benchmark methods on a variety of established data and topic sets.

5. CONCLUSION

In this work we introduced enhanced relevance scoring with chronological term rank. Chronological term rank is the rank of a term obtained from the sequential ordering of terms in the original document. In particular, we enhanced Okapi BM25 with an additive chronological rank percentage term. The new chronological term rank method produces significant improvements in the major retrieval metrics when compared to the performance of existing relevance scoring formulae on a variety of TREC data and topic sets. Methods compared include Okapi BM25 [15], pivoted length normalization [16], traditional language model, title language model of Jin et al. [8], and Global-By-Value and Local-By-Rank models of Anh and Moffat [1, 2]. The CTR approach presented here has also been successfully implemented as part of our vertical digital library prototype [2], to appear in the future.

Chronological term rank goes beyond the prevailing *bag-of-words*, term frequency approaches to relevance scoring by incorporating fine-grained information regarding document structure in the relevance estimation process. The chronological term rank method has the advantage of being conceptually and computationally simple when compared to other document structure approaches such as those incorporating term proximity, term order or document features. Chronological scores of terms are independent of one another, and can be calculated by a single counter at indexing time. For these reasons, CTR can be more easily incorporated into many existing retrieval systems.

Further improvements in retrieval performance using CTR are likely possible. We experimentally evaluated many considerations in the use of CTR in order to identify an optimal strategy. These experiments provide a good foundation for future work in the development of other approaches incorporating chronological term rank.

6. REFERENCES

[1] V. N. Anh and A. Moffat. Impact transformation: effective and efficient web retrieval. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–10, 2002.

[2] V. N. Anh and A. Moffat. Simplified similarity scoring using term ranks. In *SIGIR '05: Proceedings of the*

[2] `http://memsworldonline.case.edu`

Collection	Query Set	Metric	CTR	Local-By-Rank		Global-By-Value	
			Score	Score	%△	Score	%△
TREC12	051–200	MAP	0.2574	0.2196	+17.2	0.2032	+26.7
		Prec@10	0.5149	0.4773	+7.9	0.4480	+14.9
		R. Rank	0.7011	0.6452	+8.7	0.6509	+7.7
TREC45-CR	351–450	MAP	0.2331	0.2180	+6.9	0.1995	+16.8
		Prec@10	0.4590	0.4470	+2.7	0.4260	+7.7
		R. Rank	0.7011	0.6873	+2.0	0.5412	+29.5

Table 3: Retrieval performance of CTR compared with the published performance of the Local-By-Rank and Global-By-Value methods of Anh and Moffat.

Collection	Query Set	CTR	BM25		LM		TLM	
		Score	Score	%△	Score	%△	Score	%△
AP2	201–250	0.2825	0.2463	+14.7	0.2238	+26.2	0.2667	+5.9
AP3	201–250	0.2856	0.2511	+13.7	0.2411	+18.5	0.2711	+5.3
SJM	201–250	0.2102	0.1727	+21.7	0.1845	+13.9	0.2081	+1.0
WSJ2	201–250	0.2176	0.1719	+26.6	0.1844	+18.0	0.1950	+11.6

Table 4: Mean average precision of CTR compared with the published performance of the BM25, language model (LM) and title language model (TLM) from Jin et al.

28th annual international ACM SIGIR conference on Research and development in information retrieval, pages 226–233, 2005.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[4] M. Beigbeder and A. Mercier. An information retrieval model using the fuzzy proximity degree of term occurences. In *SAC '05: Proceedings of the 2005 ACM symposium on Applied computing*, pages 1018–1022, 2005.

[5] S. Büttcher, C. L. A. Clarke, and B. Lushman. Term proximity scoring for ad-hoc retrieval on very large text collections. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 621–622, 2006.

[6] J. R. Dominick. *The Dynamics of Mass Communication*. McGraw-Hill Inc., 1990.

[7] D. Hawking and P. Thistlewaite. Relevance weighting using distance between term occurrences. Technical Report TR-CS-96-08, The Australian National University, August 1996.

[8] R. Jin, A. G. Hauptmann, and C. X. Zhai. Title language model for information retrieval. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–48, 2002.

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

[10] E. M. Keen. Term position ranking: some new test results. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 66–76, 1992.

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

[12] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[13] Y. Rasolofo and J. Savoy. Term proximity scoring for keyword-based retrieval systems. In *Proceedings of the 25th European Conference on IR Research (ECIR 2003)*, pages 207–218, April 2003.

[14] S. Robertson, H. Zaragoza, and M. Taylor. Simple bm25 extension to multiple weighted fields. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49, 2004.

[15] S. E. Robertson, S. Walker, and M. Beaulieu. Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7), NIST Special Publication 500-242*, pages 253–264, July 1999.

[16] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 21–29, 1996.

[17] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. Technical Report IR-416, University of Massachusetts Amherst, 2005.

[18] E. M. Voorhees and L. P. Buckland, editors. *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005), NIST Special Publication 500-266*. National Institute of Standards and Technology, November 15-18 2005.

[19] H. Zaragoza, N. Craswell, M. Taylor, S. Saria, and S. Robertson. Microsoft Cambridge at TREC 13: Web and hard tracks. In *Proceedings of the the Thirteenth Text REtrieval Conference (TREC 2004), NIST Special Publication 500-261*, 2004.