Convergence Analysis of Ranking Measures for Web Search *

Di He HEDI@CIS.PKU.EDU.CN
Peking University
Beijing, P.R. China

Liwei Wang WANGLW@CIS.PKU.EDU.CN
Peking University
Beijing, P.R. China

Wei Chen CHENWEI@AMSS.AC.CN
Chinese Academy of Sciences
Beijing, P.R. China

Tie-Yan Liu TYLIU@MICROSOFT.COM
Microsoft Research Asia
Beijing, P.R. China

Abstract

This paper theoretically verifies whether a ranking measure is appropriate for Web search, from the convergence point of view. Many ranking measures have been proposed in the literature, such as Precision@k, WTA and NDCG, most of which contain a position discount function to reflect users' attention on top-ranked documents. A common practice to use these measures to evaluate ranking models is to check the values of these measures on a benchmark dataset, which usually contains a relatively small number of labeled documents per query. However, as we know, in real Web search, one usually needs to deal with an extremely large number of documents for many queries. Therefore, it is unclear whether the evaluation result in terms of a given measure on the benchmark dataset can consistently reflect the performance of the model in real Web search. If not, we think the ranking measure and the corresponding evaluation results cannot be used to select ranking models in a reliable manner. In this regard, we argue that the convergence of a ranking measure with the increasing number of documents is a dispensable property from both theory and application points of view. We then perform formal study on the convergence of ranking measures. Our theoretical analysis indicates that (i) when the discount function in a ranking measure decreases sharply with respect to positions (e.g., truncated at top k positions), the ranking measure will not converge; (ii) when the decrease is slow (e.g., with a logarithm discount), the ranking measure will converge to a model-independent constant; (ii) only when the decrease rate is in a certain range, the ranking measure can converge to a model-dependent value and be feasible for model selection. These findings can not only help us judge whether a ranking measure is good, but also provide a way of improving it. We have conducted experiments on both toy and real data. The corresponding experimental results well verified the above theoretical findings.

. *This work was conducted when the first and third authors visited Microsoft Research Asia.

1. Introduction

Web search has played a more and more important role in our daily life, especially in the era of information explosion. How to build a powerful search engine has gained increasing attention from both industry and academia. Researchers in various areas, such as information retrieval and machine learning have been working on different aspects of Web search, such as ranking model construction Beaulieu et al. (1995); Herbrich et al. (1999); Ponte and Croft (1998), and search result evaluation Basilico and Hofmann (2004); Breese et al. (1998); Järvelin and Kekäläinen (2000, 2002); Le et al. (2009).

Due to the large amount of information contained in the Web, for many queries there will be millions of related documents or even more. However, according to a user study, 62% of search engine users only click on the results within the first search result page. In other words, people tend to pay more attention to web documents ranked on the top of the search result, which are supposed to be more relevant to the query. Accordingly, when testing the effectiveness of a search engine, the evaluation measures should also emphasize the top-ranked documents. Widely-used evaluation measures like NDCG Järvelin and Kekäläinen (2000, 2002), Precision@k, WTA, and NERU Basilico and Hofmann (2004); Breese et al. (1998) all include certain position discount functions in their definitions. With these measures, a common practice for evaluation is to construct a benchmark dataset, which contains a number of queries and a set of labeled (e.g., relevant or irrelevant) documents associated with each query. For example, in the benchmark datasets used in many TREC tracks Clarke et al. (no date.), a query is associated with tens or hundreds of human-labeled documents. Then given a ranking model, one tests its performance on the benchmark dataset according to certain ranking measures and then uses the evaluation results to predict the effectiveness of the model (and/or compare different models) in real Web search scenarios. This is usually referred to as the Cranfield paradigm Sparck-Jones (1981) in the literature of information retrieval.

While the above evaluation paradigm has been widely used, we would like to point out an important problem with it. As can be seen above, the size of the benchmark dataset is much smaller than that of real Web search data, in the sense that there are only tens or hundreds of documents per query in the dataset but we may need to rank millions of or even more documents in Web search. Then the question is whether the evaluation results obtained from such a small benchmark dataset can be reliable enough, and can reflect the true performance of a ranking model when it handles a very large number of Web documents.

One may find that an issue related to the above question has been studied in the literature of machine learning. Specifically, in classification, one also cares about whether the classification error rate (which can be regarded as an evaluation measure) observed on the finite seen data can be closed to the error rate on the infinite unseen data. The conclusion regarding this is quite trivial: the classification performance of a model on unseen data is just the expectation of the performance on seen data, and the gap between them converges to zero by the Law of Large Numbers. Then a natural question is whether it is equally trivial to obtain conclusions in our case of Web search ranking. Unfortunately, as will be shown later, the answer is no and it turns out to be a new challenge when we are dealing with Web search. This is mainly because the ranking measures for Web search have much more complex forms than the classification error rate. Specifically, most ranking measures used in Web search are weighted sum of rank statistics and no longer sum-i.i.d. In this case, as far as we know, there is no existing conclusion on whether the ranking measures can converge with respect to the increasing sample size. Furthermore, even if they can converge, it is still unclear whether the convergence value will be meaningful and can be used to distinguish different models. For example, in practice, we may encounter one of the following situations.

(1) For some ranking measures, when the number of documents increases, the ranking performance of a model does not converge. That is, the model may be better than other models when evaluated on a number of documents, but it may become worse when the number of documents increases. Due to this vibrating performance, we can hardly make reliable prediction on its performance in real Web search based on the observations on the benchmark datasets.

(2) For some ranking measures, when the number of documents increases to infinity, the performance of a model converges to a model-independent constant. The convergence ensures that a model will have consistent performance even if the number of documents increases, which is good. However, since the convergence value is model-independent, all models will perform similarly when the number of documents is large, and a model with outstanding performance on the benchmark dataset may not make a difference in real Web search settings.

(3) For some other measures, when the document number approaches infinity, the performance of a model converges to a model-dependent value. In this case, when the number of documents increases, a model will have consistent performance and different models can be effectively distinguished. Only in this case, we can trust the evaluation result on benchmark datasets to certain extent.

In this paper, we have performed theoretical study on the convergence of ranking measures. Specifically, we look at the position discount function in a ranking measure, and investigate the conditions for the ranking measure to fall into one of the above cases. In the literature Kanoulas and Aslam (2009), different choices of the discount functions in ranking measures have also been discussed, however, mainly from the efficiency and stability point of view. Our discussions from the convergence perspective can be a good complement to these previous works, since this new perspective provides a mathematical (but not an empirical) justification of a given discount function. To fulfill our study, we employ a condition-partition technique, and obtained the following conclusions, which is distribution free: (1) if the discount function $D(r) = o(r^{-1-\epsilon})$ for some $\epsilon > 0$, the ranking measure does not have convergence property; (2) if the discount function $D(r) = \Omega(r^{-\epsilon}), \forall\epsilon > 0$, the ranking measure converges to 1, independent of the ranking model; (3)if the discount function $D(r) = \Theta(r^{-\alpha})$ for some $\alpha \in (0.1]$, the ranking measure converges to a value dependent on the ranking model.

By applying these general conclusions to widely-used ranking measures, we can obtain the following results: (1) WTA, NDCG@k, Precision@k, and all other top-k ranking measures do not converge; NERU does not converge; (2) NDCG with logarithm discount function, i.e., $D(r) = \frac{1}{\log(r+1)}$, converges to 1 for all the ranking models; (3) NDCG with polynomial discount function, i.e., $D(r) = r^{-\alpha}$ where $\alpha \in (0, 1]$ converges to a value dependent on ranking models.

We have conducted experiments on both simulation data and real data. The experimental results have verified the correctness of our theoretical findings. According to these findings, many ranking measures commonly used today are not suitable to evaluate the ranking performance in Web search from the convergence point of view and we may need to make adjustments in order to take good use of them. For example, we had better use a certain polynomial discount function in NDCG. In this sense, our findings do not only have theoretical values but also have practical impact.

The rest of the paper is organized as follows. In Section 2, we introduce ranking measures. In Sections 3, we introduce convergence property of ranking measures, and we present our main theoretical results in Section 4. The simulation and real experiment results are provided in Section 5. Conclusions and future work are discussed in the last section.

2. Ranking Measures

In web search, when a query is submitted to the search engine, the search engine will first retrieve a set of related web documents, apply a ranking model to assign each document a score indicating its relevance to the query, and rank the documents in the descending order of their scores. To judge the performance of a ranking model, several ranking measures have been proposed, such as NDCG, Precision@k, WTA, and NERU. For ease of discussion, we use M to denote a ranking measure. According to Le et al. (2009), most of the ranking measures, especially the ones mentioned above, can be written in the following form:

$$M(\pi, \mathcal{Y}) = \frac{1}{N_n} \sum_{r=1}^{n} G(y_{\pi(r)}) D(r),$$

where π is a permutation representing the ranked list; $\mathcal{Y} = \{y_1, \ldots, y_n\}$ where $y_i \in \{0, \ldots, L-1\}$ is the relevance judgments for each web document; $G(y)$ is called the gain function, which is an increasing function of label y and equals zero when the document is irrelevant; $D(r)$ is called the position discount function, which is a decreasing function of position r; and N_n is a normalization term. Suppose a web document is represented by a feature vector $x \in \mathcal{R}^d$. Given a set of such documents, $\mathcal{X} = \{x_1, x_2, ..., x_n\}$, their labels $\mathcal{Y}$, and a ranking model f, the ranking measure M can be reformulated as below.

$$M(f, S_n) = \frac{1}{N_n} \sum_{r=1}^{n} G(y_{\pi_f(r)})) D(r),$$

where $\pi_f(r)$ is the index of the document ranking at position r by f, and $S_n = \{\mathcal{X}, \mathcal{Y}\}$.

We list the discount and gain functions of some commonly-used ranking measures as follows.

(1) **NDCG** (Normalized Discounted Cumulative Gain) usually uses an exponential gain function and a logarithm discount function:

$$D(r) = \frac{1}{\log(r+1)}, \; G(y) = 2^y - 1.$$

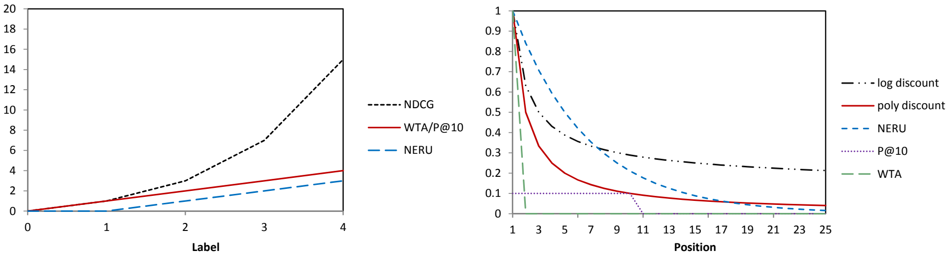
Sometimes, other discount functions such as polynomial functions are also used, such as $D(r) = r^{-1}$ Järvelin and Kekäläinen (2000).

Considering that users may only browse the first page of the search result, top-k version of NDCG (denoted as NDCG@k) is also widely used in practice, which truncates the discount function at position k.

(2) **WTA** (Winner-Takes-All) only cares about the first position in the ranking result. If the document at the first position is relevant,[1] the WTA score is 1; otherwise it is 0. Correspondingly, the discount and gain functions for WTA are as follows:

$$D(r) = \begin{cases} 1 & r = 1 \\ 0 & \text{otherwise} \end{cases}, \; G(l) = l.$$

[1] When the labels are given in terms of K-level ratings ($K > 2$), one can fix a level k^*, and regard all the objects whose ratings are higher than k^* as relevant and the rest as irrelevant Qin et al. (2010).



(a) Gain functions

(b) Discount functions

Figure 1: Gain and discount functions of ranking measures

(3) **Precision@k** measures the proportion of relevant documents in the top k positions. Its discount and gain functions are listed as follows.

$$D(r) \quad = \quad \begin{cases} \frac{1}{k} & r \leq k \\ 0 & \text{otherwise} \end{cases}, \; G(l) = l.$$

(4) **NERU** (Normalized Expected Rank Utility) is often used in the scenario of collaborative filtering. Its discount and gain functions are defined as follows.

$$D(r) \quad = \quad 2^{\frac{1-r}{\beta-1}}, \; G(l) = \max(l-d, 0).$$

where d is a neutral vote and α is the so-called viewing halflife Le et al. (2009).

The gain and discount functions of the aforementioned ranking measures are plotted in Figure 1 (for NERU, the variables β and d are chose as 5 and 1 in the figure). From the figure, we can see that the discount functions in NERU and top-k ranking measures (including WTA, Precision@k and NDCG@k) decrease the fastest with respect to position r, followed by the polynomial discount function used in NDCG, while the logarithm discount function used in NDCG decreases the slowest.

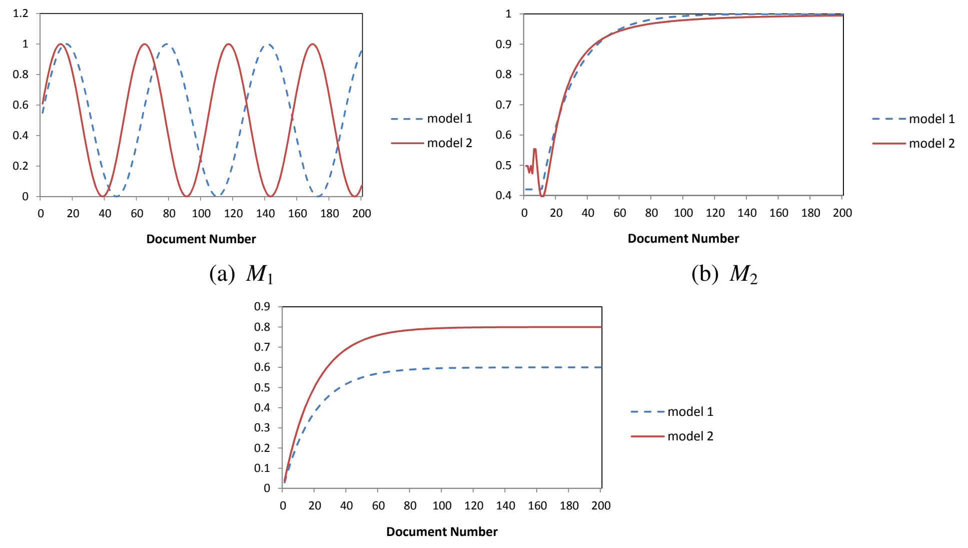
3. Convergence of Ranking Measures

In Web search, the search result is induced by a ranking model f, which maps the feature space onto an interval of R. For easy reading, the interval is supposed to be $[0, 1]$. For a given data set, the performance of f can be evaluated by the ranking measures mentioned in the previous section. As a common practice, people usually use some benchmark datasets to compare different ranking models. However, as mentioned in the introduction, the number of documents per query in benchmark datasets is usually much smaller than that in real Web search. As a result, it is not clear whether a ranking model f can work well in real Web search even if it has a good performance on the benchmark datasets.

In our opinion, this question is highly related to the *convergence of ranking measures*, which is defined as follows.

Definition 1 *Assume S_n contains n websites and their labels that are i.i.d. sampled. Given a ranking measure M and a ranking model f, if there exsits $C \in R$, s.t. $\forall \epsilon > 0$,*

$$\lim_{n\to\infty} P\{|M(f, S_n) - C| > \epsilon\} = 0, \; (i.e. \; M(f, S_n) \xrightarrow[n\to\infty]{p} C),$$



(a) M_1

(b) M_2

(c) M_3

Figure 2: Performance curves of f_1 and f_2 under three ranking measures

then we say f converges w.r.t. M, and denote C as $M(f)$. If $\forall f$ converges w.r.t. M, we say that M has the convergence property.

As for the convergence property, given a ranking measure M, we may encounter one of the following situations:

1. M doesn't have the convergence property, i.e., there exists a ranking model that doesn't converge w.r.t. M;

2. M converge and $M(f) \equiv const, \forall f$. We say that such ranking measures have trivial convergence property;

3. M converge and $M(f)$ is dependent on f. In this case, we say the ranking measure has non-trivial convergence property.

Here we use an example to further illustrate the aforementioned three situations. Suppose we have three ranking measures M_1, M_2, M_3 and two randomly selected ranking models f_1 and f_2. The performances of the ranking models w.r.t. different ranking measures with the increasing number of documents are plotted in Figure 2.

From the figure we can see that for ranking measure M_1, the performances of the two ranking models vibrate. Sometimes f_1 is better than f_2 and sometimes f_2 is better than f_1. As a consequence, if we observe one model is better than another on the benchmark dataset, we cannot make prediction whether this model will also outperform the other one in real Web search (where the number of documents increases). Therefore, it is not appropriate to use ranking measures with no convergence property in the evaluation of Web search.

For ranking measure M_2, the two ranking models perform similarly when the number of documents is large. However, if we only look at a small number of documents, their performances may be quite different. For example, if only the first 20 documents are contained in the benchmark dataset, f_1 will be regarded as better than f_2; if the first 40 documents are in the benchmark dataset, we would say that f_2 is better than f_1. However, the fact is that their performances in real Web search are very similar. In this sense, it is not appropriate either to use ranking measures with trivial convergence property for Web search evaluation.

For ranking measure M_3, when the number of documents increases, the performances of the two ranking models tend to be stable and the corresponding value for f_1 is better than that for f_2. As a result, as long as the benchmark dataset has a relatively large size, we can reliably predict which ranking model is better in real Web search based on their performances on the benchmark dataset. In this regard, we say that ranking measures with non-trivial convergence property can be used for the evaluation of Web search.

In summary, convergence w.r.t. the increasing number of documents is an essential property for ranking measures. It is therefore meaningful and also important to analyze the converge properties of existing ranking measures, and/or design new ranking measures with non-trivial convergence properties. This task is, however, not easy, mainly because ranking measures are usually complex in their mathematical forms. Comparatively speaking, the same convergence problem is straightforward for classification. As we know, in classification, the classification error rate of any classifier always converges and its convergence value is just the expected 0-1 loss, according to the Law of Large Numbers. However, because ranking measures are weighted sum of rank statistics and are no longer sum-i.i.d., the Law of Large Numbers cannot be simply applied. To tackle this challenge, we employee a condition-partition technique. The corresponding results will be presented in the next section. To our best knowledge, this is the first work that touches this issue and provides a sound theoretical result.

4. Main Results

In this section, we perform formal study on the convergence of ranking measures with respect to the increasing number of documents. First, we show our theoretical findings for binary relevance judgment in Section 4.1; and then we step forward to the case regarding multi-level relevance judgment in Section 4.2. Further discussions on commonly used ranking measures are provided in Section 4.3.

As mentioned before, many ranking measures contain a gain function and a position discount function. Our study shows that the gain function does not affect the convergence property while the discount function does. For ease of presenting our findings, we focus on the following three types of discount functions.

Definition 2 ***Three Types of Discount Functions***

1. *We call a discount function $D(r)$* ***Type 1*** *discount function, if $D(r) = o(r^{-1-\epsilon})$ for some $\epsilon > 0$.*
2. *We call a discount function $D(r)$* ***Type 2*** *discount function, if $D(r) = \Omega(r^{-\epsilon}), \forall \epsilon > 0$.*
3. *We call a discount function $D(r)$* ***Type 3*** *discount function, if $D(r) = \Theta(r^{-\alpha})$ for some $\alpha \in (0, 1]$.*[2]

These three types of discount functions can cover most commonly used ranking measures in Web search. For instance, the exponential discount function in NERU and the discount functions

[2] Please refer to Knuth (1998) for the definitions of Θ, Ω, o.

in top-k ranking measures (including WTA, Precision@k, and NDCG@k) belong to Type 1; the logarithm discount function used in NDCG belongs to Type2; and the polynomial discount function used in another NDCG implementation belongs to Type 3.

Before we present our main results, we would like to introduce two notations that are used throughout our analysis. One is $P_f(s) = P(f(X) > s)$, which denotes the distribution of ranking scores produced by ranking model f. The other is $g_f(s, y) = P(Y = y)p(f(X) = s|Y = y)$, where $p(f(X) = s|Y = y)$ is the conditional probability density function for scores $f(X)$ given the relevant judgment $Y = y$.

4.1 Results for Binary Relevance Judgment

In this subsection, we perform convergence analysis for ranking measures in the setting of binary relevance judgment. Binary judgment is a special data assumption, with which the documents only have two relevance levels, i.e., 1 (relevant) and 0 (irrelevant).The analysis regarding the three types of discount functions are presented in the following subsections..

4.1.1 Analysis for Type 1 Discount Functions

The convergence properties of ranking measures with Type 1 discount functions are stated in the following theorem.

Theorem 3 *If the discount function of a ranking measure belongs to Type 1, it does not have convergence property.*

Proof We just need a counter example in order to prove the theorem. Since $D(r) = o(r^{-1-\epsilon})$, for some $\epsilon > 0$, there exists $N \in \mathbb{N}$, s.t. $D(r) < r^{-1-\epsilon}, \forall r > N$. Thus, $\sum D(r) < \sum_{r=1}^{N} D(r) + \sum_{r=N+1}^{\infty} r^{-1-\epsilon} < \infty$. Assume that $\sum D(i)$ is bounded by K. For a distribution satisfying $g_f(s, 1) = g_f(s, 0) = 1/2, \forall s$, the probability of label 0 or 1 appearing at any position is the same (i.e., 0.5). Therefore, for any two lists in which only the label of the top1 position are different, their probabilities will be the same. However, the difference between their ranking measures is at least $1/K$. Then we can come to the conclusion that the convergence property does not exist. ■

Since top-k ranking measures and NERU have Type 1 discount functions, we have the following corollary.

Corollary 4 *Any Top-k ranking measure and NERU do not have convergence property.*

Remark: Usually, people have thought that the top-k ranking measures can elegantly reflect the user behavior in web search. However, according to our theoretical result, they are not good choices to measure the performance of a ranking model. For example, if we retrieve 1000 web documents for a given query and find a ranking model is better than the other in terms of NDCG@10, we have no idea whether we will have the same observation if we retrieve 2000 web documents. In this regard, we should doubt the wide use of top-k ranking measures and the reliability of all previous evaluation results on the benchmark datasets in terms of top-k ranking measures.

4.1.2 Analysis of Type 2 Discount Functions

The convergence property of ranking measures with Type 2 discount functions are given in the following theorem. The proof is in Appendix.

Theorem 5 *If the discount function of a ranking measure M belongs to Type 2, then $M(f, S_n) \xrightarrow[n\to\infty]{p} 1$, i.e., M has trivial convergence property.*

Based on this theorem, the following corollary is straightforward.

Corollary 6 *NDCG with the logarithm discount function $D(r) = 1/\log(r + 1)$ has trivial convergence property.*

Remark: As we know, NDCG with logarithm discount has been widely used in the literature. However, our theoretical results show that the convergence value of NDCG with such a discount function is the same for any ranking model. Even if a ranking model ranks all the relevant documents at the bottom, the value of NDCG will still converge to 1 as long as we have enough documents to rank. This result indicates that even if a ranking model has outstanding performance in terms of NDCG on the benchmark dataset, when we use it in real Web search, its performance will be just similar to other ordinary ranking models. In this regard, the evaluation results in terms of NDCG with logarithm discount on benchmark datasets are not as reliable as expected.

4.1.3 Analysis of Type 3 Discount Functions

The convergence property regarding this kind of discount function is given in the following theorem.

Theorem 7 *If the discount function of a ranking measure belongs to Type 3, then we have:*

$$M(f, S_n) \xrightarrow[n\to\infty]{p} \begin{cases} \frac{(1-\alpha)\int_0^1 g_f(s,1)(P_f(s))^{-\alpha} ds}{(P(Y=1))^{1-\alpha}} & \textit{if } \alpha \in (0,1) \\ P(Y = 1|f(X) = 1) & \textit{if } \alpha = 1 \end{cases}$$

To prove this theorem, we need to introduce two lemmas whose proofs are in Appendix.

Lemma 8 *If the discount function $D(r) = \Theta(r^{-\alpha})$ where $\alpha \in (0, 1]$, then*

$$\mathbf{E}[M(f, S_n)] \xrightarrow[n\to\infty]{} \begin{cases} \frac{(1-\alpha)\int_0^1 g_f(s,1)(P_f(s))^{-\alpha} ds}{(P(Y=1))^{1-\alpha}} & \textit{if } \alpha \in (0,1) \\ P(Y = 1|f(X) = 1) & \textit{if } \alpha = 1 \end{cases}$$

Lemma 9 *If the discount function $D(r) = \Theta(r^{-\alpha})$ where $\alpha \in (0, 1]$, then*

$$Var(M(f, S_n)) \xrightarrow[n\to\infty]{} 0$$

Proof of Theorem 7 With the above lemmas, we can get

$$\mathbf{E}[M(f, S_n)] \xrightarrow[n\to\infty]{} \begin{cases} \frac{(1-\alpha)\int_0^1 g_f(s,1)(P_f(s))^{-\alpha} ds}{(P(Y=1))^{1-\alpha}} & \text{if } \alpha \in (0,1) \\ P(Y = 1|f(X) = 1) & \text{if } \alpha = 1 \end{cases}$$

and $Var(M(f, S_n)) \xrightarrow[n\to\infty]{} 0$. Using the Chebyshev Inequality, we can obtain

$$M(f, S_n) \xrightarrow[n\to\infty]{p} \begin{cases} \frac{(1-\alpha)\int_0^1 p_f(s,1)(P_f(s))^{-\alpha} ds}{(P(Y=1))^{1-\alpha}} & \text{if } \alpha \in (0,1) \\ P(Y = 1|f(X) = 1) & \text{if } \alpha = 1 \end{cases}.$$

This exactly proves Theorem 3.

It is clear that the convergence value given in Theorem 3 will take different values for different ranking measures. In other words, it is model-dependent and can be used for model comparison and selection. In this regard, ranking measures with Type 3 discount functions are more suitable for Web search evaluation. One example of such ranking measures is NDCG with a polynomial discount function $D(r) = r^{-\alpha}, \alpha \in (0, 1]$.

4.2 Results for Multi-level Relevance Judgment

In many recent practices, the relevance judgments are not binary, but multi-level ratings instead. For example, the ratings can be {highly relevant, relevant, irrelevant}, or {perfect, excellent, good, fair, bad}. Some ranking measures like NDCG are specifically designed for such relevance judgment. But other ranking measures like Precision@k and WTA can also be computed, by simply assuming some of the ratings as relevant and the rest as irrelevant. For example, we can regard both "relevant" and "highly relevant" as relevant. In this section, we will analyze the convergence of ranking measures when multi-level relevance judgments are used. Our findings are summarized in the following theorem, which indicates that the convergence property only depends on the type of discount functions, regardless of the gain function. One can see that the findings regarding multi-level ratings are similar to those regarding binary relevance judgement. Actually the proof is also similar, and therefore we omit it here.

Theorem 10 *Suppose we have L-level rating labels, i.e.,* $y \in \{0, \ldots, L-1\}$. *For any increasing gain function* $G(y)$*, we have the following conclusions for ranking measure M:*

(1) if the discount function $D(r)$ *belongs to Type 1, then M has no convergence property;*

(2) if the discount function $D(r)$ *belongs to Type 2, then M has trivial convergence property;*

(3) if the discount function $D(r)$ *belongs to Type 3, then M has non-trivial convergence property.*

By applying the above general results to some specific ranking measures, we can obtain more specific conclusions, as shown in the following corollary.

Corollary 11 *Suppose we have L-level rating labels, i.e.,* $y \in \{0, \ldots, L-1\}$. *For any increasing gain function* $G(y)$*, we have the following conclusions for ranking measure M:*

(1) top-k ranking measure M does not converge;

(2) when $D(r) = \frac{1}{\log(r+1)}$, $M(f, S_n) \xrightarrow[n\to\infty]{p} 1$;

(3) when $D(r) = r^{-1}$,

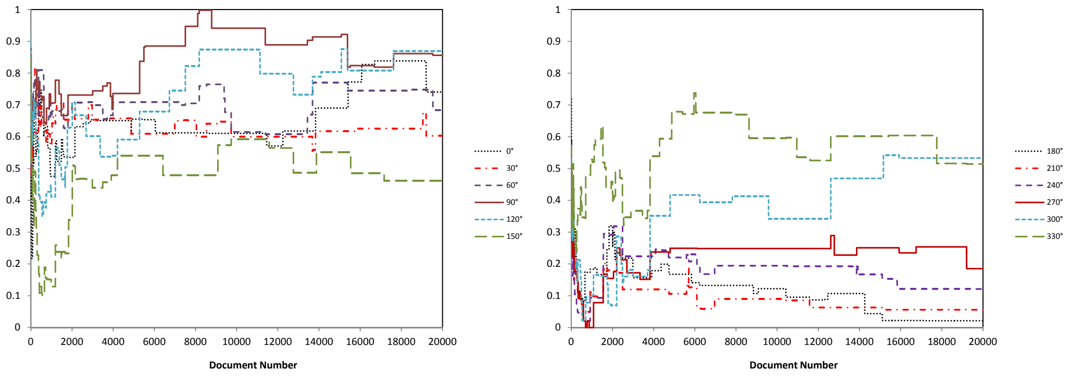
$$M(f, S_n) \xrightarrow[n\to\infty]{p} \frac{\sum_{i=1}^{L-1} G(i)P(Y = i|f(x) = 1)}{\max_{P(Y=l)\neq 0} G(l)},$$

(4) when $D(r) = r^{-\alpha}, \alpha \in (0, 1)$, *we have* $M(f, S_n) \xrightarrow[n\to\infty]{p}$

$$\frac{(1-\alpha)\int_0^1 \sum_{i=0}^{L-1} G(i) g_f(s, i)(P_f(s))^{-\alpha} ds}{\sum_{i=0}^{L} G(i)((\sum_{j=i}^{L-1} P(Y = j))^{1-\alpha} - (\sum_{j=i+1}^{L-1} P(Y = j))^{1-\alpha})},$$

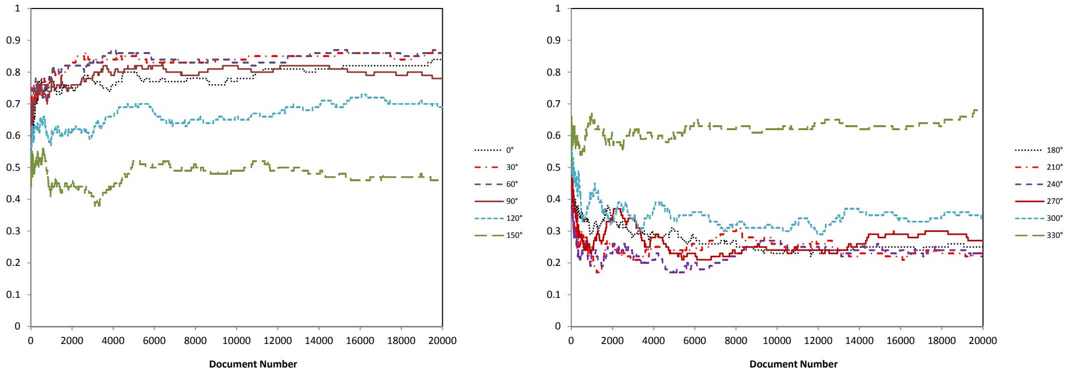
4.3 Summary

In this section, we have analyzed the convergence properties of different ranking measures in different settings. So far, the conclusions indicate that some widely-used ranking measures either do



(a) (b)

Figure 3: Performance curves for NDCG@10 on simulation data



(a) (b)

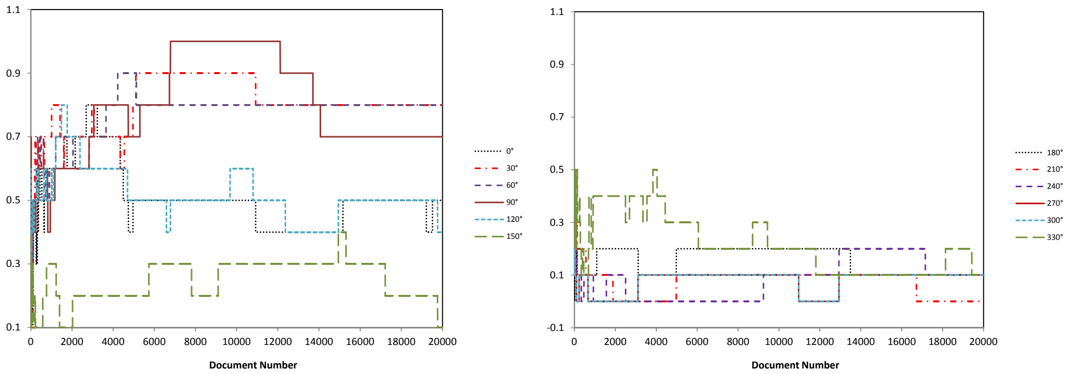
Figure 4: Performance curves for Precision@10 on simulation data

not converge or converge to some trivial constant, when the number of documents increases. This is quite negative, and implies that the evaluation results on benchmark datasets in terms of these ranking measures reported in the literature are not reliable.

Our theoretical studies also suggest that we can improve the situation by changing the widely-used logarithm discount in NDCG to a polynomial discount. Then the corresponding ranking measure will have non-trivial convergence property, and can be used for model comparison and selection in Web search. We think this finding has its value for both researchers and practitioners.

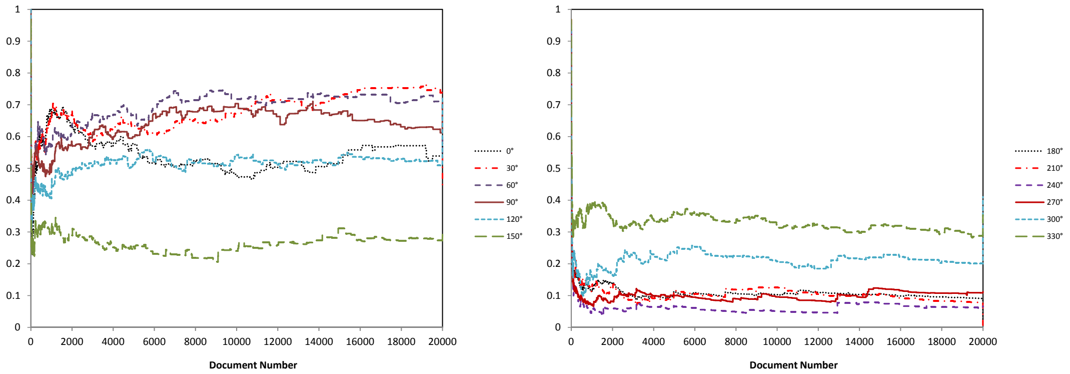
5. Experiments

In this section, we report our experimental results on the convergence of six ranking measures, i.e., NDCG and NDCG@k with logarithm discount function, NDCG with polynomial discount function $D(r) = r^{-1/2}$, Precision@k, WTA, and NERU. We have used both simulation and real data in our experiments.



(a) (b)

Figure 5: Performance curves for WTA on simulation data



(a) (b)

Figure 6: Performance curves for NERU on simulation data

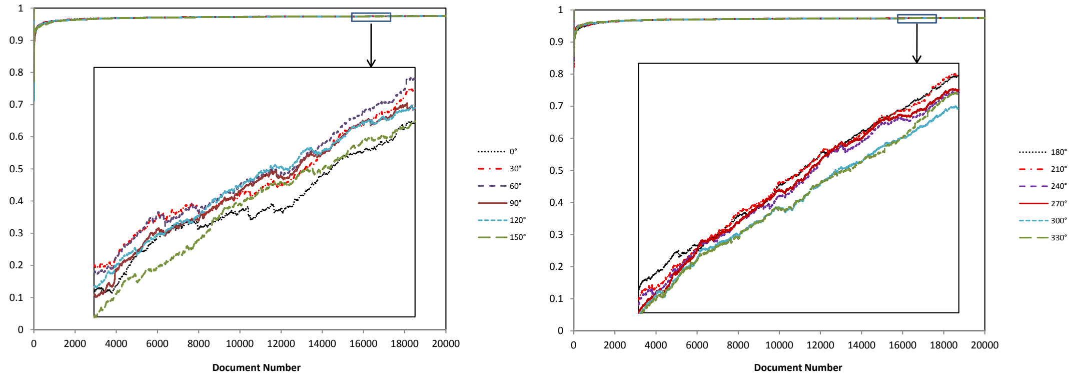
5.1 Experiments on Simulation Data

In this section, we report our experimental results on simulation data. As we mentioned previously, widely-used benchmark datasets such as MSLR Qin et al. (2010), Yahoo Learning to Rank Challenge dataset do not contain sufficient number of labeled documents per query. Therefore, they cannot be used for the experiments on convergence, which basically investigates the situation when the number of documents is large. In order to fully verify our theoretical results, we choose to conduct experiments on simulation data.

5.1.1 Experimental Design

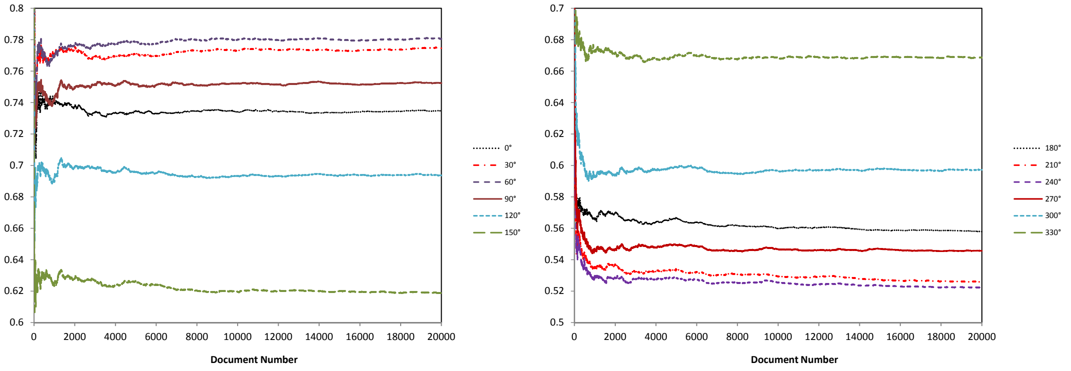
We generate the simulation data as follows. For simplicity, we assume that the document space is $\mathcal{R}^2$ and label y takes value from$\{0, 1, 2\}$. The probability of the three labels are $a_0 = 0.4, a_1 = 0.3, a_2 = 0.3$ respectively.[3] Given a label, the conditional probability of documents is a 2-dimensional normal distribution. We denote the conditional probabilities for the three labels as $N(\mu_0, \Sigma)$, $N(\mu_1, \Sigma)$,

[3] The probability does not affect our theoretical results by much, since our results are distribution free.



(a) (b)

Figure 7: Performance curves for NDCG with logarithm discount function on simulation data



(a) (b)

Figure 8: Performance curves for NDCG with polynomial discount function $D(r) = r^{-1/2}$ on simulation data

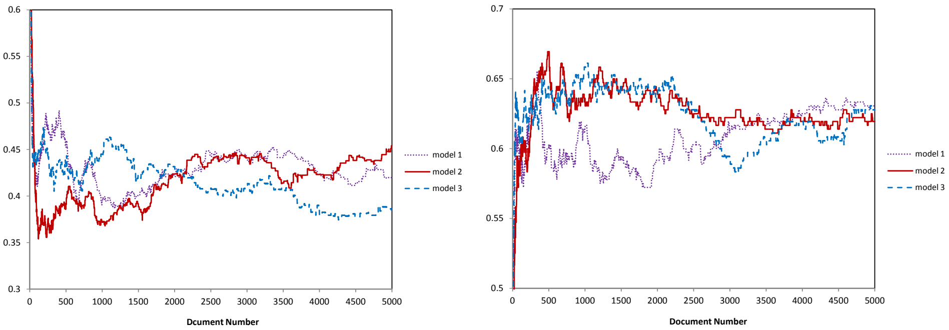
$N(\mu_2, \Sigma)$. The means and variance matrix are set as $\mu_0 = (0, 0), \mu_1 = (0.3, 0.3), \mu_2 = (0.5, 0.6)$, and $\Sigma = I_2$, where I_2 denotes the identity matrix of size 2.

Given the distributions, we sampled 20000 documents, and repeated the sampling for ten times. All of the experimental results reported in the following sections are the average results over the ten trials.

To comprehensively show the convergence properties of the ranking measures, we constructed twelve different linear rankers, whose angles with respect to the x axis are 0, 30, 60,..., 330, respectively.

5.1.2 Experimental Results

We evaluated the twelve ranking models under the six ranking measures and obtained their performance curves with respect to the increasing number of documents. For clarify, we plot the performance curves of first six models in a sub figure and the rest curves in another sub figure.



(a) NDCG@10

(b) Precision@10

Figure 9: Performance curves for NDCG@10/Precision@10 on real data

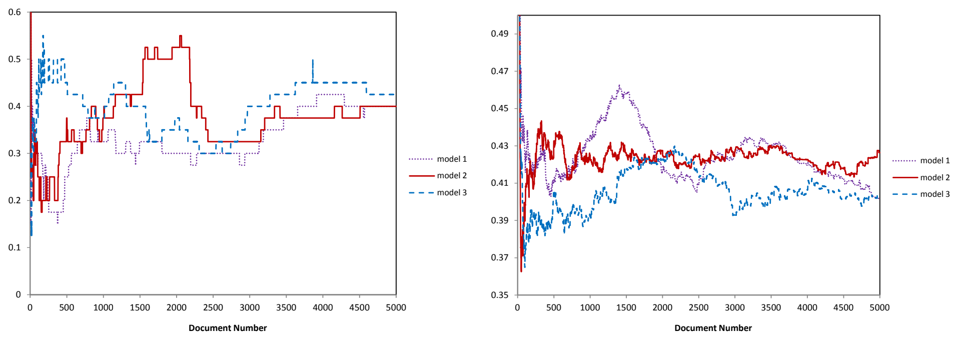
Performance curves of the models under ranking measures with Type 1 discount functions are plotted in Figure 3 to 6. From these figures, we can see that the performances of the models vibrate with the increasing number of documents, and do not converge even if the number of documents is very large. Furthermore, sometimes one model is better than the other and sometimes it is the other way around. Take Precision@10 in Figure 4 as an example. The performance curves of the model with angle 0 and the model with angle 90 cross each other for three times and the last time happens when the document number is about 15000. As a result, one can hardly conduct reliable model selection based on a given number of documents. All these experimental results are consistent with our theoretical results, providing empirical evidence that ranking measures with Type 1 discount functions do not have convergence property.

Performance curves of the models under NDCG with logarithm discount (which belongs to the Type 2 discount functions) are plotted in Figure 7. In this figure, as the document number increases, the performance curves of all the ranking models converge to 1. By zooming to the area where the document number changes from 15000 to 17000 , we found that the curves of different models interleave even if the document number is very large. These results are consistent with our theoretical results with respect to Type 2 discount functions, indicating that ranking measures with Type 2 discount functions are not suitable to evaluate and compare ranking models.

Performance curves of the models under NDCG with polynomial discount (which belongs to the Type 3 discount functions) are plotted in Figure 8. From the figure, we can see that the performance curves of all the twelve ranking models converge and the convergence values are different for different models. Moreover, when the number of documents is large (e.g., larger than several hundreds), the relative goodness of different models become quite stable and the curves seldom cross each other. In this case the empirical results on model selection obtained from a relatively small dataset have already been very reliable. These results are consistent with our theoretical findings on Type 3 discount functions.

5.2 Experiments on Real Data

In this section, we report our experimental results on real data. As mentioned in the introduction, in real scenario of Web search, many queries have a large number of related documents. Examples include IPHONE, ThinkPad, and Adobe. In order to test the convergence of ranking measures on such queries, we have used the click-through logs of a commercial search engine, and derived rele-



(a) WTA (b) NERU

Figure 10: Performance curves for WTA/NERU on real data

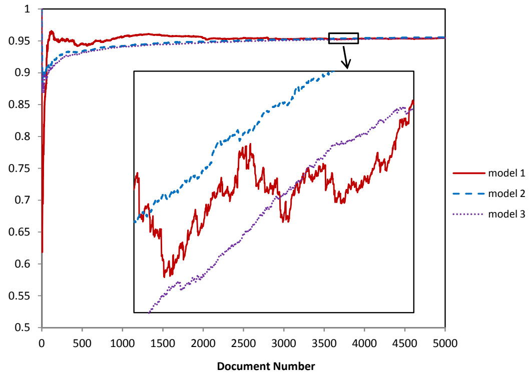


Figure 11: Performance curves for NDCG with logarithm discount function on real data

vance judgments from the click number on each document. The details about dataset construction is explained as below.

5.2.1 Experimental Design

We collected the clicked documents for 40 popular queries in total. Each query is associated with 5000 web documents with clicks. The label of each document is determined by its click numbers. In particular, we regard all document with more than 1000 clicks as having label 2, all documents with 100 to 1000 clicks as having label 1, and the remaining documents as having label 0. Overall, the ratio of documents with label 0, 1, and 2 is about 5:2:1. We extracted 40 features for each document representing its relevance to the query. The features include term frequency, inverted document frequency, BM25 Beaulieu et al. (1995), and language model for IR Ponte and Croft (1998). We randomly chose three linear ranking models, and evaluated them under the six ranking measures. The experimental results reported are averaged over the 40 queries.

5.2.2 Experimental Results

The performance curves of the models under ranking measures with Type 1 discount functions are shown in Figure 9 and 10. We can see from the figures that the performances do not converge for these measures. Take NDCG@10 as example, when the number of document per query is 1200,

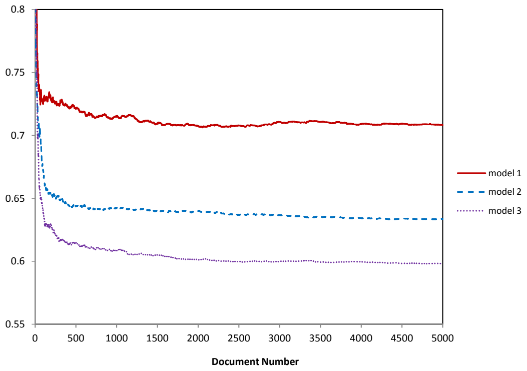


Figure 12: Performance curves for NDCG with polynomial discount function $D(r) = r^{-1/2}$ on real data

model 3 performs the best. However, when the number of documents increases to 3500, the best performer becomes model 1. When we have 5000 documents per query, model 2 outperforms the other two models. In this regard, it is not reliable to use NDCG@10 (and other measures with Type 1 discount functions) to evaluate the performance of a ranking model. This is consistent with our theoretical results.

Figure 11 shows the performance curves under NDCG with logarithm discount (which belongs to Type 2 discount functions). From the figure we can see that the NDCG values of all the three models converge to 1, which is the same with the simulation result. As a consequence, it is difficult to conduct model selection using NDCG since all the models have similar performances when the dataset becomes large. Moreover, in the enlarged subfigure, we can see that the curves of the three models are interwoven with each other heavily. Therefore, we can come to the conclusion that NDCG with logarithm discount cannot be used to reliably evaluate or select ranking models. This conclusion is consistent with our theoretical results.

From Figure 12, we can see under NDCG with polynomial discount function $D(r) = r^{-0.5}$, the performance of the three ranking models converge to different values as document size grows. And even if we draw a conclusion about model selection from a small number of documents per query, the conclusion is consistent with that drawn from a large number of documents. For example, model 1 always performs the best no manner how many documents we have. In this sense, we say that ranking measures with Type-3 discount functions are suitable for model evaluation and comparison. This is consistent with both the results on simulation data and with our theoretical findings.

To sum up, all the above experimental results well verify the correctness of our theoretical findings.

6. Conclusion and Future Work

In this paper, we have studied the convergence properties of ranking measures for Web search. Our theoretical analysis shows that NDCG with logarithm discount, Precision@k, WTA, and NERU are not good choices for model evaluation and comparison, although they have been widely used in the literature. In contrast, NDCG with polynomial discount seems to be a good ranking measure, since it will converge to a model-dependent value when the number of documents per query becomes large. We have verified these theoretical findings using both simulation and real data.

For future work, we plan to investigate the convergence properties of more ranking measures, such as Average precision and MRR. We will also investigate other theoretical aspects of ranking measures, such as consistency and uniform convergence.

References

J. Basilico and T. Hofmann. Unifying collaborative and content-based filtering. In *ICML*, page 9, 2004.

M. Beaulieu, M. Gatford, X. Huang, SE Robertson, S. Walker, and P. Williams. Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126, 1995.

J.S. Breese, D. Heckerman, C. Kadie, et al. Empirical analysis of predictive algorithms for collaborative filtering. In *UAI*, pages 43–52, 1998.

Charles L.A. Clarke, Nick Craswell, and Ian Soboroff. Overview of the trec 2009 web track. Technical report, no date.

R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. *NIPS*, pages 115–132, 1999.

K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *SIGIR*, pages 41–48, 2000.

K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

Evangelos Kanoulas and Javed A. Aslam. Empirical justification of the gain and discount function for ndcg. In *CIKM*, pages 611–620. ACM Press, 2009.

Donald E. Knuth. *The art of computer programming*, volume 3. Addison-Wesley Longman Publishing Co., Boston, MA, USA, 2nd edition, 1998.

Q.V. Le, A. Smola, O. Chapelle, and C.H. Teo. Optimization of ranking measures, 2009.

Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *SIGIR*, pages 275–281, 1998.

Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, 1:1386–4564, 2010.

K. Sparck-Jones. The cranfield tests. *Information retrieval experiment*, pages 256–284, 1981.