# Estimating Interleaved Comparison Outcomes from Historical Click Data

Katja Hofmann
k.hofmann@uva.nl

Shimon Whiteson
s.a.whiteson@uva.nl

Maarten de Rijke
derijke@uva.nl

ISLA, University of Amsterdam

## ABSTRACT

Interleaved comparison methods, which compare rankers using click data, are a promising alternative to traditional information retrieval evaluation methods that require expensive explicit judgments. A major limitation of these methods is that they assume access to live data, meaning that new data must be collected for every pair of rankers compared. We investigate the use of previously collected click data (i.e., historical data) for interleaved comparisons. We start by analyzing to what degree existing interleaved comparison methods can be applied and find that a recent probabilistic method allows such data reuse, even though it is biased when applied to historical data. We then propose an interleaved comparison method that is based on the probabilistic approach but uses importance sampling to compensate for bias. We experimentally confirm that probabilistic methods make the use of historical data for interleaved comparisons possible and effective.

## Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

## Keywords

Information retrieval, Interleaved comparisons, A/B testing, Implicit feedback, Evaluation, Reusability

## 1. INTRODUCTION

*Interleaved comparison* methods [8, 17, 18], which compare rankers using naturally occuring user interactions such as clicks, are quickly gaining interest as a complement to traditional evaluations for information retrieval (IR). Compared to evaluations based on manual judgments from expert annotators, interleaved comparison methods rely only on data that can be collected cheaply and unobtrusively. Since this data is based on the behavior of real users, it more accurately reflects how well their information needs are met [18].

Existing interleaved comparison methods suffer from a drawback: they assume access to *live data*, i.e., data gathered during the evaluation itself. Comparing two rankers requires presenting users with interleaved result lists based on those rankers and observing how they interact with them. Since *historical data*, collected while comparing two different rankers, is unlikely to contain the same result lists, it is unclear how it can be exploited by interleaved comparison methods. No existing research has addressed this question.

Without the ability to estimate comparison outcomes using historical data, the practical utility of interleaved comparison methods is limited. If all comparisons are done with live data, then applications such as learning to rank, which perform many comparisons, need prohibitive amounts of data. Since interleaving result lists may affect the user's experience of a search engine, the collection of live data is complicated by the need to first control the quality of the compared rankers using alternative evaluation setups.

Using previously collected data is already possible for evaluation methods that use explicit assessments [1]. Thus, although obtaining explicit judgments is initially expensive, this cost can be amortized over the whole set of evaluations that reuse the same judgments. Interleaved comparisons are less expensive initially but cannot be amortized, as live data is needed for each new evaluation.

We remedy this shortcoming by investigating the use of historical data by interleaved comparison methods. First, we analyze to what degree existing approaches can exploit historical data. We show that, while the most widely known approach, called team-draft [17, 18], cannot do so effectively, a recently developed probabilistic method [8] can. This method makes efficient use of sample data by marginalizing over the ways in which observed result lists may have been constructed. In the live data setting, the probabilistic method is unbiased, i.e., its expected value equals the expected outcome. However, it is biased when applied to historical data. Second, we introduce an interleaved comparison method that is based on the probabilistic approach but uses *importance sampling* to correct for bias. Third, we present an empirical analysis of the use of historical data by interleaved comparison methods. Because the estimated values are ultimately used only to make a binary decision, the original, biased approach is surprisingly robust. When lots of historical data is available, the unbiased approach performs better. The reuse of historical data enables the application of interleaved comparisons to learning to rank and large-scale evaluation feasible.

## 2. RELATED WORK

In this section, we discuss IR literature that is related to the use of clicks for IR evaluation as well as approaches to *off-policy evaluation* using historical data in reinforcement learning.

**Click-based evaluation in IR.** Click data is a promising source of information for IR systems as it can be collected at low cost, is abundant in frequently-used search applications, and reflects user behavior and preferences. There are ongoing efforts to incorporate click data in retrieval algorithms, e.g., for pseudo-relevance feedback [12], and in learning to rank [10, 11].


*CIKM'12,* October 29–November 2, 2012, Maui, HI, USA.

Using click data to evaluate retrieval systems has long been a promising alternative to expensive explicit judgments ("editorial data"), but the reliability of click-based evaluation has been found to be problematic. While reliability appears to be high in professional search, where trained archivists search professionally maintained archives [7, 21], evaluation methods that interpret clicks as absolute relevance judgments in more broadly used settings were found to be unreliable, due to large differences in click behavior between users and search topics [13, 18].

One line of research that addresses the problem of high variance in user clicks has resulted in click behavior models that combine explicit judgments and click data *per query* [3, 4]. These models are trained to predict clicks and/or relevance of documents that have not been presented to users at a particular rank, or that have not been presented at all for the given query. Such models leverage click data to allow more accurate evaluations with relatively few explicit judgments [2, 15]. The models can be reused but, unlike our method, require access to editorial data.

An alternative to click-based evaluation that does not require editorial judgments and is robust to noise in user clicks is interleaved comparison [6, 8, 18]. Such methods infer relative preferences between a pair of rankers. They work by combining pairs of document rankings into interleaved document lists, which are then presented to the user, instead of the original lists. User clicks on the interleaved list are observed and projected back to the original lists to infer which list would be preferred by users. Repeating this interleaving over many queries leads to very reliable comparisons. The work in this paper is based on the team-draft [17, 18] and probabilistic interleave [8] methods, which are described in §3.[1]

**Off-policy evaluation.** The problem of estimating interleaved comparison outcomes using historical data is related to *off-policy evaluation* [20] in *reinforcement learning* (RL), a branch of machine learning in which agents learn from interaction with an environment by taking actions and receiving rewards. Solving RL problems requires being able to evaluate a *policy* that specifies what actions the agent should take. The challenge in off-policy evaluation is to use data gathered with one policy to evaluate another one even though the two policies may specify different actions for a given context.

Algorithms for off-policy evaluation have been developed for tasks similar to IR, namely news recommendation [14] and ad placement [19]. Many existing algorithms are not directly applicable to the IR setting because they assume reward can be directly observed (e.g., in the form of clicks on ads). Since clicks are too noisy to be treated as absolute reward in IR [13, 18], only relative feedback can be inferred. We consider how to reuse historical data for interleaved comparison methods that work with implicit, relative feedback.

One tool employed by existing off-policy methods that is applicable to our setting is *importance sampling* [16, 20]. Importance sampling can be used to estimate the expected value $E_T[f(X)]$ under a *target distribution* $P_T$ when data was collected under a different *source distribution* $P_S$. The importance sampling estimator is:

$$E_T[f(X)] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)\frac{P_T(x_i)}{P_S(x_i)}, \tag{1}$$

where $f$ is a function of $X$, and the $x_i$ are samples of $X$ collected under $P_S$. These are reweighted according to the ratio of their probability of occurring under $P_T$ and $P_S$. This estimator is unbiased (i.e., its expected value is equal to $E_T[f(X)]$) as long as the source distribution is non-zero at points where the target distribution is.

Importance sampling can be more or less efficient than using the target distribution directly, depending on how well the source distribution focuses on regions important for estimating the target value. In this work, we use importance sampling to derive an unbiased estimator of interleaved comparison outcomes using historical data.

# 3. METHODS AND ANALYSIS

We discuss three approaches for estimating interleaved comparison outcomes from historical data. First, we analyze to what degree current state-of-the-art interleaved comparison methods (the team-draft and probabilistic interleave methods) can be applied to historical data.[2] Then we propose an unbiased approach that is derived by applying importance sampling to probabilistic interleave.

## 3.1 Team-draft

*Team-draft* [17, 18] is based on the idea that team captains (rankers) select their most valued players (documents) in a friendly match (gaining clicks on result documents). Like other interleaved comparison methods, team-draft compares two ranked lists of documents, $l_1$ and $l_2$, given a query $q$, using two steps: (1) interleaving and (2) comparison. During the first step, an interleaved result list is generated from the original lists. For each pair of ranks to be filled with result documents, a coin flip determines which of $l_1$ or $l_2$ gets to select a document first. This list then contributes its highest-ranked document that is not yet part of the interleaved list. The algorithm also records which list contributed which document in an assignment vector $a$. Once the result list is constructed, it is presented to the user, whose clicks are then recorded. During comparison, clicks are attributed to the original lists that contributed the clicked documents. The list obtaining more clicks wins the comparison.

A naive way of applying the team-draft method to historical data is to use only observed lists that could have been constructed under the target rankers for the given query. The effectiveness of this approach depends on the similarity of the document lists under the original and target rankers. The more these differ, the less likely it is to find matching interleaved result lists and, typically, the amount of overlap is expected to be very low. Furthermore, even in cases where the original and target pairs are very similar pairs, relying on historical data is problematic, as it can lead to bias. For example, the interleaved result lists collected using the original ranker pair may overlap only with target document lists for specific queries (e.g., only "easy" queries with a clearly identifiable relevant document) that poorly reflect the relative performance of the target rankers. This can lead to incorrect decisions about the preferred target ranker.

There is no apparent way to derive an unbiased estimator based on team-draft that reuses historical data. Such an estimator can be derived using the probabilistic interleave method, described below.

## 3.2 Probabilistic Interleave

The probabilistic interleave method [8] is a recent interleaved comparison method based on a probabilistic formulation of the team-draft method. Using this formulation, the method can make more effective use of observed click data than previous methods.

Like team-draft, probabilistic interleave consists of an interleaving step and a comparison step. During the first step, the interleaved result list is generated as follows. For each rank of the interleaved

[1] Like all interleaved comparison methods, our approach may be affected by bias in click behavior. However, they have been demonstrated to work well in practical settings [17].

[2] The balanced interleave [18] and document constraint [6] methods are not analyzed in detail. Our arguments regarding the data efficiency and bias of the team-draft method when applied to historical data apply to these methods as well.

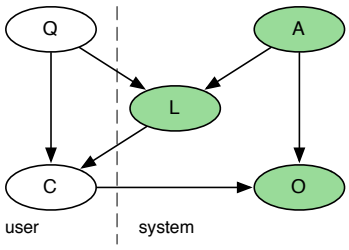


**Figure 1: Probabilistic model for comparing rankers, following [8]. Shaded nodes indicate that the probability distributions for these variables are known.**

list, it randomly selects an original list from $(l_1, l_2)$ to contribute a document (recorded in an assignment vector $a$, as in team-draft). Then, selecting a document from this list is formulated as random sampling without replacement from a probability distribution over documents. The distribution over documents is instantiated such that the highest-ranked document is the most likely to be drawn and probabilities quickly decrease with lower ranks. Consequently, all possible permutations of result documents are possible.

The second step, comparison, could be implemented exactly as in the team-draft method. However, a more effective comparison method can be derived using the observation that the interleaving process can be represented by a probabilistic model (cf., Fig. 1). The model describes the relationship between queries $q \in Q$, assignments $a \in A$, interleaved lists $l \in L$, clicks $c \in C$, and outcomes $o \in O$. $Q$ and $C$ are unknown but samples of both are observed. We also treat the outcome $O$ as a random variable. $O$ has three possible values and is deterministic given $a$ and $c$. If $l_1$ obtained more clicks than $l_2$, then $O = -1$. If both lists obtained the same number of clicks, then $O = 0$. If $l_2$ obtained more clicks, then $O = 1$.

With this model definition, we can derive an estimator of the expected comparison outcome, given $n$ samples of the form $(c_i, l_i, q_i)$. To allow for efficient use of available sample data, the observed, noisy, assignments are ignored. The expected value of the comparison outcome $E[O]$ is estimated by marginalizing across all possible assignments that could have led to an observed list $l$. This results in graded click assignments that depend on the similarity between the compared rankers (i.e., small differences are inferred if a document is moved up by one rank, and bigger differences are inferred for more dramatic changes). The method also reduces noise resulting from randomized assignments, making it more effective than methods that directly use observed assignments.

The resulting estimator computes the expected comparison outcome $E[O]$ given samples from an interleaving experiment conducted according to the graphical model in Fig. 1:

$$E[O] \approx \frac{1}{n} \sum_{i=1}^{n} \sum_{a \in A} \sum_{o \in O} oP(o|c_i, a)P(a|l_i, q_i). \qquad (2)$$

The probability of an assignment given observed lists and queries is computed using: $P(A|L, Q) = P(L|A, Q)P(A|Q)/P(L|Q)$ (Bayes' rule). Note that $P(A|Q) = P(A)$, because $A$ and $Q$ are independent, and $P(L|Q)$ is fully specified by the interleaving process. Finally, $P(L|A, Q)$ can be obtained using $P(L|A, Q) = \prod_{r=1}^{|L|} P(L[r] \mid A[r], L[1, r-1], Q)$. Here, $|L|$ is the length of the document list, $L[r]$ denotes the document placed at rank $r$ in the interleaved list $L$, $L[1, r-1]$ contains the documents added to the list before rank $r$, and $A[r]$ denotes the assignment at rank $r$, i.e., which list contributed the document at $r$.

Probabilistic interleave can be directly applied to historical data in the same way as team-draft. It is expected to be more effective, since it can infer information about two target rankers from clicks observed on any interleaved list. While it is unbiased in the live data setting, probabilistic interleave suffers from bias in the historical data setting in cases where the original distribution of interleaved result lists differs from the distribution under the target rankers. For example, suppose we compare two target rankers given historical data collected under two original rankers, where one of the target rankers reverses the result list of one of the original rankers. In the observed data, documents ranked low under the target rankers are much more likely to be highly ranked than those ranked highly under the target ranker. Hence, documents that are highly ranked under the target ranker had a lower probability of being inspected and ultimately clicked by the user when the data was collected, than if the data was collected under the target rankers. Thus, outcomes based on the historical data are biased against this target ranker.

## 3.3 Importance Sampling

Our analysis showed that existing interleaved comparison methods can be applied to historical data to various degrees, with team-draft expected to make very inefficient use of historical data, and probabilistic interleave expected to be able to utilize all historical data to some degree. A drawback in both cases is that outcomes inferred on the basis of historical data can be biased. Building on the properties of the probabilistic method we can apply statistical methods to derive an unbiased estimator.

Consider two pairs of rankers. Pair $S$ is the source ranker pair, which was compared in a live experiment using interleaved result lists from which the comparison outcome was computed using the resulting clicks. All data from this experiment have been recorded, and we want to compare a new ranker pair $T$ using this data. Observations for $S$ occur under the original distribution $P_S$, while observations for $T$ occur under the target distribution $P_T$. Our goal is to estimate the outcome of comparing $T$, given data from the past experiment of $S$, by compensating for the difference between $P_T$ and $P_S$. $P_T$ and $P_S$ can be seen as instantiations of the graphical model in Fig. 1. Both instantiations have the same event spaces and only the distribution over $L$ changes for different ranker pairs. The distributions over $A$ are the same by design of the interleaving process. Distributions over $C$ and $Q$ are the same for different ranker pairs because we assume that clicks and queries are drawn from the same static distribution, independently of the ranker pair used to generate the presented list.

A naive estimator of the expected outcome $E_T[O]$ from sample data observed under $P_S$ can be obtained from the definition of the importance sampling estimator in Eq. 1 with $f(C, A) = \sum_{o \in O} oP(o|C, A)$:

$$E_T[O] \approx \frac{1}{n} \sum_{i=1}^{n} \sum_{o \in O} oP(o|c_i, a_i) \frac{P_T(c_i, a_i)}{P_S(c_i, a_i)}. \qquad (3)$$

This estimator is unbiased but expected to perform poorly, as it merely reweights the original, noisy, estimates, which can increase the overall variance. To derive an efficient estimator, we need to marginalize over all possible assignments, as in the original probabilistic interleave method. This results in the following estimator, which is unbiased given samples from an interleaving experiment conducted according to the graphical model in Fig. 1 under $P_S$:

$$E_T[O] \approx \frac{1}{n} \sum_{i=1}^{n} \sum_{a \in A} \sum_{o \in O} oP(o|c_i, a)P(a|l_i, q_i) \frac{P_T(l_i|q_i)}{P_S(l_i|q_i)}. \qquad (4)$$

Note that $P(L|Q) = \sum_{a \in A} P(L|a, Q)P(a)$, which can be obtained using Bayes' rule and the fact that $P(a) = \frac{1}{|A|}$.

Equation 4 completes our importance sampling estimator for comparing rankers using historical data that was collected using the probabilistic interleave process defined in Fig. 1.[3]

## 4. EXPERIMENTAL SETUP

Interleaved comparison methods have not been applied to historical data in previous work, so we cannot follow accepted experimental practice. Here, we propose to follow a simulation-based approach that allows us to assess the effectiveness of interleaved comparison methods for historical data under different conditions. The setup is based on the simulation framework used for evaluation in [8] and combines learning to rank data sets and click models to simulate users' interactions with a retrieval system. Here, we extend this framework to allow for the simulation of historical interactions. Below, we give an overview of our data, procedure, user model, and experimental runs.

Our experiments are run on the $18,919$ queries of the training set of fold 1 of the MSLR-WEB30k Microsoft learning to rank data set [8].[4] This data set encodes relations between queries and candidate documents in 136 precomputed features, and provides (manual) relevance judgments on a 5-point scale (from 0 – "non-relevant" to 4 – "highly relevant"). We generate rankers from the features provided with the data set. This means that our experiments simulate the task of comparing the effectiveness of individual features for retrieval using varying amounts of historical data.

Our experiment investigates the bias-variance tradeoff inherent to the probabilistic methods by studying the convergence of comparison methods for individual queries. It assumes that a large amount of historical data has been collected for a set of queries and ranker pairs (e.g., for frequently occurring queries in web search). To simulate historical data, we use the following procedure. First, we randomly sample a query $q$ and an original ranker pair from the available data, record both, and then generate $n = 500{,}000$ interleaved result lists and clicks for $q$ using the original ranker pair. The interleaved result lists are generated according to the interleaved comparison method being evaluated (i.e., either using TD, or probabilistically, depending on the method).

User clicks are generated using the following model of user interactions with a result list. A user interaction consists of submitting a query to the system, examining the top 10 documents of the returned result list, and clicking links to promising documents. Users inspect and click documents following the Dependent Click Model [5]. They start with the top-ranked document and proceed down the list, clicking on promising documents and, after viewing a document, decide whether to stop or examine more documents. Click and stop probabilities are instantiated using the graded relevance assessments provided with the data set. It is assumed that users are more likely to click on more relevant documents, based on the attractiveness of e.g., the document title and snippet. Our experiments use a *perfect click model* [8], with a stopping probability of zero (i.e., the user inspects all top 10 results of each interleaved list), and a click probability that linearly increases with document relevance (with click probability $P(c) = 0.0$ for non-relevant documents and $P(c) = 1.0$ for highly relevant documents).

[3]Note that the resulting estimator does not depend on the assignments observed in the original data and could be applied to data collected in an arbitrary way, as long as the distribution over result lists is known and non-zero for all lists that are possible under the target distribution. This opens up directions for future work on more effective sampling algorithms for interleaved comparisons.

[4]http://research.microsoft.com/en-us/projects/mslr/default.aspx

After generating a set of historical data, we randomly sample a target ranker pair and use the historical data to infer a comparison outcome for this target pair, using either our baseline (TD) or our experimental methods. We measure the performance of the interleaved comparison methods in terms of accuracy compared to the Normalized Discounted Cumulative Gain (NDCG) [9]. We repeat the experiment 500 times, and report the accuracy of each method averaged over repetitions after observing $n$ historical data points.

## 5. RESULTS

In this section we report on the result obtained in our experiments, as described in §4. The experiments are designed to compare the accuracy of different interleaved comparison methods when using large amounts of historical data. We compare the following interleaved comparison methods for historical data:

- **TD**: the team-draft method following [18], with naive reuse of historical data as described in §3.1.
- **PI+MA**: probabilistic interleaving that marginalizes over assignments but without bias correction, as defined in Eq. 2 (cf. §3.2, [8]).
- **PI+IS**: probabilistic interleaving with naive importance sampling, using the estimator defined in Eq. 3 (cf. §3.3).
- **PI+IS+MA**: probabilistic interleaving that marginalizes over assignments and corrects for bias using importance sampling, as defined in Eq. 4 (cf. §3.3).

Fig. 2 shows the accuracy of interleaved comparison methods over the number of historical samples. TD performs poorly, as it reuses only a small portion of samples for which the original and target pairs are similar enough to have overlap. We observe a strong effect of this limitation, with a low, almost constant accuracy of $7.4\%$.

The second-lowest performance is from PI+IS. Performance suffers from high variance caused by importance sampling for pairs where the distributions over interleaved lists under the target ranker pair is different from the original distribution. In these cases, a large portion of the observed samples has a small probability of occurring under the target distribution, and therefore receives a very low weight. Rarely, a result list occurs that is very unlikely under the original distribution, but very likely under the target distribution. If such a result list is observed, it receives a very high weight (in extreme cases importance weights on the order of $10^{10}$). Individual samples with extreme weights strongly affect comparison outcomes. For example, a large number of low-weight samples could be observed that indicate a preference for the wrong target ranker. The estimate of the outcome is then suddenly corrected when one sample with a high weight is added. This type of behavior results in the up-and-down pattern of the accuracy under the PI+IS method.

For PI+MA and PI+IS+MA, we observe a tradeoff between bias and variance. The accuracy of PI+MA converges comparatively quickly, with only small changes in comparison outcomes after 1,000 observed samples. Final accuracy is only $65.4\%$, $5\%$ lower than that of PI+IS+MA after processing the same number of samples. PI+IS+MA exhibits a pattern of noise similar to that of PI+IS, though to a much lesser degree. Compared to PI+IS, this method can use the available sample data much more effectively, resulting in lower variance of the importance weights and weighted outcomes. While PI+MA converges early to suboptimal performance, the performance of PI+IS+MA continues to improve with more sample data. In the limit, it is expected to converge to the level of performance achieved with live data.

This experiment confirms the hypothesized tradeoff between bias and variance for probabilistic interleaved comparison methods. Comparison outcomes under the original biased probabilistic

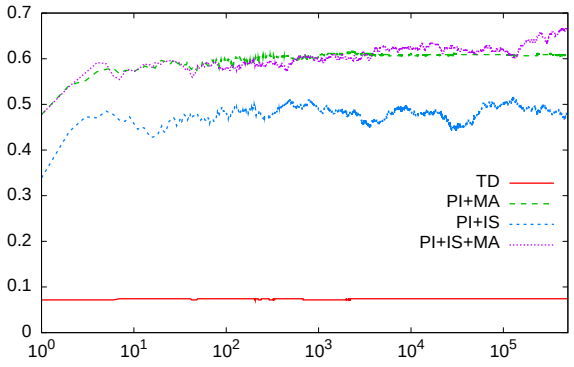


**Figure 2: Results. Accuracy over the number of historical impressions for interleaved comparisons methods applied to individual queries and ranker pairs.**

method exhibit low variance and estimates converge quickly. This can lead to relatively high performance when little sample data is available but in expectation does not produce the same outcomes as comparisons using live data. Methods that use importance sampling are unbiased, but high variance in importance weights can lead to slow convergence. Our importance sampling estimator that marginalizes over assignments (PI+IS+MA) achieves relatively low levels of noise by using the observed data more effectively than the naive importance sampling estimator.

## 6. CONCLUSION

We addressed how to make interleaved comparison methods more efficient through the reuse of historical data. Doing so is critical to making the advantages of such methods available in practical settings. We first analyzed to what degree existing interleaved comparison methods can reuse historical data. We found that a recently developed probabilistic approach can do so effectively, although it is biased. We then proposed an unbiased estimator by combining the probabilistic method with importance sampling. The first, naive approach performs poorly due to variance introduced by importance sampling. We experimentally demonstrated that probabilistic interleaved comparison methods can effectively reuse historical data. Best results were obtained by our unbiased method when large amounts of historical data are available.

We expect our methods to be the most beneficial in settings where many ranker pairs have to be compared, such as large-scale evaluation and learning to rank. In learning to rank, it is typical that the best of a set of rankers has to be identified quickly and reusing data collected for previous comparisons could result in faster and more robust learning for the same number of live impressions. Our future work will focus on exploring this direction.

**Acknowledgements.** This research was partially supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements nr 258191 (PROMISE Network of Excellence) and 288024 (LiMoSINe project), the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.-004.802, 380-70-011, 727.011.005, HOR-11-10, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP and BILAND projects funded by the CLARIN-nl program, the Dutch national program COMMIT, and by the ESF Research Network Program ELIAS.

## 7. REFERENCES

[1] B. Carterette. Robust test collections for retrieval evaluation. In *SIGIR '07*, pages 55–62, 2007.

[2] B. Carterette and R. Jones. Evaluating search engines by modeling the relationship between relevance and clicks. In *NIPS '07*, volume 20, pages 217–224, 2008.

[3] G. Dupret and C. Liao. A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. In *WSDM '10*, pages 181–190, 2010.

[4] G. Dupret, V. Murdoch, and B. Piwowarski. Web search engine evaluation using click-through data and a user model. In *In Proceedings of the Workshop on Query Log Analysis*, 2007.

[5] F. Guo, C. Liu, and Y. M. Wang. Efficient multiple-click models in web search. In *WSDM '09*, pages 124–131, 2009.

[6] J. He, C. Zhai, and X. Li. Evaluation of methods for relative comparison of retrieval systems based on clickthroughs. In *CIKM '09*, pages 2029–2032, 2009.

[7] K. Hofmann, B. Huurnink, M. Bron, and M. de Rijke. Comparing click-through data to purchase decisions for retrieval evaluation. In *SIGIR '10*, pages 761–762, 2010.

[8] K. Hofmann, S. Whiteson, and M. de Rijke. A probabilistic method for inferring preferences from clicks. In *CIKM '11*, pages 249–258, 2011.

[9] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.

[10] S. Ji, K. Zhou, C. Liao, Z. Zheng, G.-R. Xue, O. Chapelle, G. Sun, and H. Zha. Global ranking by exploiting user clicks. In *SIGIR '09*, pages 35–42, 2009.

[11] T. Joachims. Optimizing search engines using clickthrough data. In *KDD '02*, pages 133–142, 2002.

[12] S. Jung, J. L. Herlocker, and J. Webster. Click data as implicit relevance feedback in web search. *Information Processing & Management*, 43(3):791 – 807, 2007.

[13] J. Kamps, M. Koolen, and A. Trotman. Comparative analysis of clicks and judgments for IR evaluation. In *WSCD'09*, pages 80–87, 2009.

[14] L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *WSDM '11*, pages 297–306, 2011.

[15] U. Ozertem, R. Jones, and B. Dumoulin. Evaluating new search engine configurations with pre-existing judgments and clicks. In *WWW'11*, pages 397–406, 2011.

[16] D. Precup, R. Sutton, and S. Singh. Eligibility traces for off-policy policy evaluation. In *ICML'00*, pages 759–766, 2000.

[17] F. Radlinski and N. Craswell. Comparing the sensitivity of information retrieval metrics. In *SIGIR '10*, pages 667–674, 2010.

[18] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *CIKM '08*, pages 43–52, 2008.

[19] A. M. Strehl, J. Langford, L. Li, and S. M. Kakade. Learning from logged implicit exploration data. In *NIPS 2010*, pages 2217–2225, 2010.

[20] R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1998.

[21] J. Zhang and J. Kamps. A search log-based approach to evaluation. In *ECDL'10*, pages 248–260, 2010.