Modeling Expected Utility of Multi-session Information Distillation

Yiming Yang and Abhimanyu Lad

Language Technologies Institute,
Carnegie Mellon University,
Pittsburgh, USA.
{yiming,alad}@cs.cmu.edu

Abstract. An open challenge in information distillation is the evaluation and optimization of the utility of ranked lists with respect to flexible user interactions over multiple sessions. Utility depends on both the relevance and novelty of documents, and the novelty in turn depends on the user interaction history. However, user behavior is non-deterministic. We propose a new probabilistic framework for stochastic modeling of user behavior when browsing multi-session ranked lists, and a novel approximation method for efficient computation of the expected utility over numerous user-interaction patterns. Using this framework, we present the first utility-based evaluation over multi-session search scenarios, using the TDT4 corpus of news stories and compare a state-of-the-art distillation system against a relevance-based retrieval engine. We demonstrate that the distillation system obtains a 44% utility enhancement over the retrieval engine due to multi-session adaptive filtering, accurate novelty detection, and utility-based adjustment of ranked list lengths.

Key words: Multi-session distillation, utility evaluation based both on novelty and relevance, stochastic modeling of user browsing behavior.

1 Introduction

Information distillation is an emerging area of research where the focus is to effectively combine ad-hoc retrieval (IR), novelty detection (ND) and adaptive filtering (AF) over temporally ordered documents for global utility optimization [12, 2, 11]. An information distillation system is typically designed for use over multiple sessions by a user or analyst. In each session, the system processes a new chunk of documents and presents a ranked list of passages[1] based on the utility of the passages to the user, where utility is measured in terms of *relevance* as well as *novelty*. The novelty of each passage in turn depends on the history of user interaction with the system, i.e., which passages were already seen by the user in the past. User behavior is typically non-deterministic, i.e., not every document in the system-produced ranked lists is necessarily read by the

[1] We use "passage" as a generic term for any retrieval unit, e.g., documents, paragraphs, sentences, etc.

user. They may skip passages, or abandon going further down a ranked list after reading the top few passages due to various reasons, e.g. satisfaction, frustration, and so on. The nondeterministic nature of user-browsing behavior has raised an important question – how should the expected utility of a distillation system be defined, estimated and maximized over all plausible patterns of user interactions in multi-session distillation? Current literature in IR, ND and AF has not offered a satisfactory solution for the whole problem, but only partial answers for sub-problems.

Recently, there has been increasing interest in evaluation metrics that are based on a model of user behavior. For example, Moffat et al. proposed *Rank-Biased Precision* (RBP) [8], which corresponds to the expected rate of gain (in terms of graded relevance) obtained by a user who reads a ranked list *top down*, and whose stopping point in a ranked list is assumed to follow a geometric distribution. Similarly, Robertson et al. re-interpreted *Average Precision* as the expected precision observed by a user who stops with uniform probability at one of the relevant documents in the ranked list returned by the system [9]. To evaluate retrieval systems in multi-session search scenarios, Järvelin et al. proposed an extension to the Discounted Cumulated Gain (DCG) metric, known as session-based DCG (sDCG) [7] that discounts relevant results from later retrieval sessions[2], to favor early retrieval of relevant information in multi-session search scenarios, based on the assumption that examining retrieved results and reformulating the query involves an effort on the part of the user.

However, all these metrics are designed for measuring utility purely in terms of *relevance* – binary or graded. In many retrieval settings, especially scenarios involving multiple search sessions, *novelty* of information plays a crucial role in determining the overall utility of the system. Adding novelty to the definition of traditional IR metrics is not straight-forward, mainly due to its dynamic nature. Unlike *relevance*, which can be "pre-defined" for each document-query pair, novelty is an ever-changing function of which passages were read or skipped by the user in his or her interactions with the system up to the current point. Therefore, we cannot measure novelty without accounting for the dynamic and non-deterministic nature of user interaction.

Nevertheless, most novelty detection approaches and benchmark evaluations conducted in NIST and TREC have shared a convention of producing novelty judgments in an *offline* manner – all the passages which are relevant to a query are listed in a pre-specified order, and a binary judgment about the novelty of each passage is made, based on how its content differs from previous passages in the list [1]. Such novelty judgments would be correct from a user's perspective only if *all* these passages were presented by the system to the user, and in the exact same order as they were laid out during the ground truth assignment. These conditions may not hold in realistic use of a distillation system, which could show both relevant and non-relevant passages to the user, ranked according to its own notion of "good" passages.

[2] A note on terminology – In this paper, a *distillation task* consists of multiple *search sessions*, each comprising a single query. In [7], a *session* consists of multiple *queries*.

In other words, conventional evaluation schemes for novelty detection are insufficient or inappropriate for evaluating the true novelty – and hence – the true utility (relevance plus novelty) of passages in realistic settings. Non-deterministic user interactions over multi-session ranked lists make the problem even harder. The novelty of each passage would depend not only on the the user history in the current session, but also the user history in all previous sessions. Since there are many possible ways for the user to interact with multi-session ranked lists, we must evaluate the *expected* utility of the system over all interaction patterns instead of assuming a fixed pattern of user interaction, e.g., "all users read the top 10 passages in each ranked list." A principled solution would be to create a stochastic model of user behavior, and define a probability distribution over user interaction patterns with respect to multiple ranked lists, and accordingly calculate the expected utility of the system.

The above challenge has been partially addressed by the NDCU (Normalized Discounted Cumulated Utility) scheme proposed by Yang et al. [12] for distillation evaluation. NDCU uses nugget-level relevance judgments to enable automated determination of relevance and novelty of each passage in a system-produced ranked list. The evaluation algorithm scans the ranked list from top to bottom, keeping a count of all nuggets seen in each passage, thus dynamically updating the novelty of each nugget as the evaluation proceeds. Despite these desirable properties, a major limitation of NDCU is that it is only well-defined for a single ranked list. In case of a K-session distillation process, when estimating the novelty of passages in the k^{th} ranked list, how should "user history" at that point be modeled? Should we assume that all the ranked lists in the past $k-1$ sessions were completely read by the user? This assumption is obviously unrealistic. Alternatively, if we assume that the previous ranked lists were only partially browsed, then we need a principled way to model all plausible user-interaction patterns, and to estimate the *expected* utility of the system as a function of the joint probabilistic distribution of user interaction patterns over multiple sessions.

A recent approach by Clarke et al. [4] is similar to NDCU in terms of counting both relevance and novelty in utility-based evaluation and with respect to exploiting nugget-level relevance judgments. However, it also shares the same limitation with NDCU, namely not modeling stochastic user interaction patterns in multi-session distillation. sDCG [7] accommodates multiple search sessions, but lacking a probabilistic model of user behavior, it cannot account for the non-determinism associated with which passages were read by the user in each ranked list. Therefore, one is forced to make deterministic assumptions about user behavior, e.g., "users read a fixed number of documents in each ranked list" (the authors truncate each retrieved list at rank 10). Therefore, sDCG does not accurately reflect the true utility perceived by a user who can flexibly interact with multiple ranked lists presented by the system.

Our focus in this paper is to address the limitations of current methods for utility-based evaluation and optimization of distillation systems. Specifically, (i) We propose a new framework for probabilistic modeling of user browsing patterns over multi-session ranked lists. Each pattern corresponds to a possible way for a user to browse through the ranked lists. By summing over all such patterns, we calculate the Expected Global Utility of the system. (ii) This model flexibility comes at the cost of increased computational complexity, which we address using an efficient approximation technique. (iii) Using this framework, we present the first utility-based evaluation of a state-of-the-art distillation system, which produces ranked lists based on relevance as well as novelty of passages, and a popular retrieval engine `Indri` [10], which produces ranked lists based on relevance only. By comparing different configurations of the two systems, we highlight the properties of our evaluation framework, and also demonstrate that the distillation system obtains a 44% utility enhancement over `Indri` due to optimal combination of adaptive filtering, novelty detection, and utility-based adjustment of ranked list lengths.

We start by briefly describing the `NDCU` evaluation metric in the next section, followed by detailed explanation of the new framework.

2 Normalized Discounted Cumulated Utility

The Normalized Discounted Cumulated Utility (`NDCU`) scheme [12] is an extension of the popular `NDCG` metric [6] to model utility as the difference between the gain and cost incurred by a user in going through a ranked list presented by the system. Specifically, the utility of each passage is defined as:

$$U\left(p_i|l_q\right) = G\left(p_i|l_q\right) - aC(p_i) \tag{1}$$

where q is a query, l_q is the ranked list retrieved for the query, p_i is i^{th} passage in l_q, $G\left(p_i|l_q\right)$ is the *gain* (benefit) for reading the passage, $C(p_i)$ is the cost for reading the passage, and a is a pre-specified constant for balancing the gain and the cost of user interaction with the passage. The cost for reading the passage is defined as the passage length in terms of the number of words. The gain from reading the passage is defined in terms of its relevance and novelty, as follows:

$$G\left(p_i|l_q\right) = \sum_{\delta \in p_i} w(\delta, q)\gamma^{n(\delta, l_q, i-1)} \tag{2}$$

where δ is a nugget (a unit for relevance judgment), $w(\delta, q)$ is the graded relevance of δ with respect to the query q, $n(\delta, l_q, i-1)$ is the number of times δ appears in the ranked list l_q up to rank $i-1$. γ is a pre-specified *dampening factor*, reflecting the user's tolerance for redundancy. If $\gamma = 1$, the user is assumed to be fully tolerant to redundancy, and the evaluation reduces to be relevance-based only. At the other extreme of $\gamma = 0$, reading a nugget after the first time is assumed to be totally useless for the user, and hence incurs only cost. The use of nuggets as retrieval units allows flexible evaluation over arbitrary system output, as well as fine-grained determination of novelty.

The Discounted Cumulated Utility (`DCU`) of a list l_q is calculated as:

$$DCU(l_q) = \sum_{i=1}^{|l_q|} P(R_i = 1)\left(G\left(p_i|l_q\right) - aC(p_i)\right) \tag{3}$$

where $|l_q|$ is the number of passages in the ranked list, and $P(R_i = 1)$ is the probability that the passage with rank i in the list is read by the user. Since $P(R_i = 1)$ is typically a decreasing function of the rank, it serves as a discounting factor, similar to the logarithmic discount used in DCG [6]. The DCU score of the system can be normalized by the DCU score of the ideal ranked list to obtain Normalized Discounted Cumulated Utility (NDCU).

Combining relevance and novelty into a utility-based evaluation metric, and utilizing nugget-level judgments to enable automated calculation of novelty for passages in any ranked list, were the main accomplishments of the NDCU scheme. However, NDCU is only defined for a single ranked list, not supporting utility-based evaluation over multi-session ranked lists. We now describe our new framework, which extends novelty-based evaluation to multi-session retrieval scenarios.

3 New Framework

The core of the new framework is a well-defined probability distribution over user behavior with respect to multiple ranked lists. We define a utility function conditioned on user behavior, and sum over all possible user interactions to obtain an *expectation* of the utility.

Let $l_1, l_2, ..., l_K$ be a sequence of K ranked lists of passages, with lengths given by $|l_1|, |l_2|, ..., |l_K|$, respectively. We define Ω as the space of all possible user browsing patterns – each element $\omega \in \Omega$ denotes a possible way for a user to browse the ranked lists, i.e., to read a specific subset of the passages that appear in the ranked lists. Let P denote a probability distribution over the space Ω, such that $P(w)$ corresponds to how likely it is for a user to read this set of passages. Intuitively, P should assign higher probability to subsets that include passages at top ranks, reflecting common user behavior. We leave the specific details of modeling user behavior to Section 3.1.

Once we have a way of representing different user interaction patterns ω, we can define the utility as a function of ω, i.e. $\mathcal{U}(\omega)$. Note that $\mathcal{U}(\omega)$ is a random quantity, since ω is a random variable. Therefore, the obvious next step is to calculate the expected value of U with respect to the probability distribution defined over Ω. We call this quantity as Expected Global Utility:

$$\text{EGU} = \sum_{\omega \in \Omega} P(\omega)\mathcal{U}(\omega) \tag{4}$$

3.1 User Browsing Patterns

As mentioned earlier, a user can read any subset of the passages presented by the system. We will use Ω to denote the set of all subsets that the user can read. Naturally, the most flexible definition of Ω would be the power set of all passages in the K lists, and the size of such a state space would be $2^{\sum_{i=1}^{K} |l_i|}$. This is a very large state space, leading to difficulties in estimating a probability distribution as well as computing an expectation over the entire space. Another alternative is to restrict the space of possible browsing patterns by assuming that the user browses through each ranked list *top down* without skipping any passage, until he or she decides to stop. Thus, each possible user interaction is now denoted by a K-dimensional vector $\omega = \{s_1, s_2, ..., s_K\}$, such that $s_k \in \{1..|l_k|\}$ denotes the stopping position in the k^{th} ranked list. This leads to a state space of size $\prod_{i=1}^{K} |l_k|$, which is much smaller than the earlier *all-possible-subsets* alternative. We further make a reasonable assumption that the stopping positions in different ranked lists are independent of each other, i.e., $P(\omega) = P(s_1, s_2, ..., s_K) = P(s_1)P(s_2)...P(s_K)$.

The particular form of $P(s)$, i.e., the probability distribution of stopping positions in a ranked list, can be chosen appropriately based on the given domain, user interface, and user behavior. For the purposes of this discussion, we follow Moffat et al. [8] and restrict attention to the geometric distribution with an adjustable (or empirically estimated) parameter, p. However, the standard geometric distribution has an infinite domain, but each ranked list in a distillation system will have a finite length. Therefore, we use a *truncated geometric distribution with a tail mass*, i.e., for a ranked list of length l, the left-over probability mass beyond rank l is assigned to the stopping position l, to reflect the intuition that users who intended to stop before rank l will be oblivious to the limited length of the ranked list, but all users who intended to stop at a rank lower than l will be forced to stop at rank l due to the limited length of the ranked list. Formally, for the k^{th} ranked list, the stopping probability distribution can be expressed by the following recursive formula:

$$P(S_k = s) = \begin{cases} (1-p)^{s-1}p & s < |l_k| \\ 1 - P(S_k < |l_k|) & s \geq |l_k| \end{cases} \tag{5}$$

3.2 Utility of Multi-session Ranked Lists Conditioned on User Browsing Patterns

The utility of multi-session ranked lists $l_1, l_2, ..., l_K$ depends on how a user interacts with them. We now define $\mathcal{U}(\omega)$ as the utility of multiple ranked lists conditioned on a user interaction pattern. Recall that $\omega = (s_1, s_2, ..., s_K)$ specifies the stopping positions in each of the ranked lists, allowing us to construct the list of passages actually read by the user for any given ω. We denote this list as $\mathcal{L}(\omega) = \mathcal{L}(s_1, s_2, ..., s_K)$, obtained by concatenating the top $s_1, s_2, ..., s_K$ passages from ranked lists $l_1, l_2, ..., l_K$, respectively. The conditional utility $\mathcal{U}(\omega)$ is defined as:

$$\mathcal{U}(\omega) = \sum_{i=1}^{|\mathcal{L}(\omega)|} G(p_i|\mathcal{L}(\omega)) - aC(p_i) \tag{6}$$

Comparing this formula with Equation 3, which defines the Discounted Cumulated Utility (`DCU`) for a single ranked list, we see that utility calculations in the two cases are almost identical, except (i) the single ranked list in `DCU` is replaced by the synthetic $\mathcal{L}(\omega)$ from the multi-session lists, and (ii) the discounting factor

$P(R_i = 1)$ is removed here because each passage in $\mathcal{L}(\omega)$ is assumed to be read by the user.

Substituting $G(.)$ and $C(.)$ in Equation 6 using their definitions from Section 2, we have:

$$\begin{aligned}\mathcal{U}(\omega) &= \sum_{i=1}^{|\mathcal{L}(\omega)|} G(p_i|\mathcal{L}(\omega)) - a \sum_{i=1}^{|\mathcal{L}(\omega)|} C(p_i) \\ &= \sum_{i=1}^{|\mathcal{L}(\omega)|} \sum_{j=1}^{|\Delta|} I(\delta_j, p_i) w(\delta_j, q) \gamma^{n(\delta, \mathcal{L}(\omega), i-1)} - a \sum_{i=1}^{|\mathcal{L}(\omega)|} \text{len}(p_i) \qquad (7)\end{aligned}$$

where Δ is the full set of nuggets in the data collection; $I(\delta_j, p_i) \in \{1, 0\}$ indicates whether or not nugget δ_j is contained in passage p_i, and $\text{len}(p_i)$ is the length of passage p_i.

The first term in Equation 7 is the cumulated gain (CG) from the synthetic list, which can be further calculated as:

$$\begin{aligned}CG(\omega) &= \sum_{j=1}^{|\Delta|} w(\delta_j, q) \left(\sum_{i=1}^{|\mathcal{L}(\omega)|} I(\delta_j, p_i) \gamma^{n(\delta_j, \mathcal{L}(\omega), i-1)} \right) \\ &= \sum_{j=1}^{|\Delta|} w(\delta_j, q) \left(1 + \gamma + \gamma^2 + ... + \gamma^{m(\delta_j, \mathcal{L}(\omega)) - 1} \right) \\ &= \sum_{j=1}^{|\Delta|} w(\delta_j, q) \frac{1 - \gamma^{m(\delta_j, \mathcal{L}(\omega))}}{1 - \gamma} \qquad (8)\end{aligned}$$

where $m(\delta_j, \mathcal{L}(\omega))$ is the count of passages that contain the nugget. An interesting insight we can obtain from Equation 8 is that the CG value depends on ω only through nugget counts $m(\delta_j, \mathcal{L}(\omega))$ for $j = 1, 2, ..., |\Delta|$. Thus, these nugget counts are the sufficient statistics for calculating CG.

The second term in Equation 7 is the cumulated cost (CC) weighted by a, which is dependent on $\mathcal{L}(\omega)$ only through the count of total word occurrences in the list. Thus the word count is a sufficient statistic for CC, and we denote it by $\text{len}(\mathcal{L}(\omega))$.

Rewriting utility $\mathcal{U}(\omega)$ as a function of the sufficient statistics, we have:

$$\begin{aligned}\mathcal{U}(\omega) &= g(m(\mathcal{L}(\omega)) - a\,\text{len}(\mathcal{L}(\omega)) \qquad (9) \\ &= \frac{1}{1 - \gamma} \sum_{j=1}^{|\Delta|} w(\delta_j, q) \left(1 - \gamma^{m(\delta_j, \mathcal{L}(\omega))} \right) - a\,\text{len}(\mathcal{L}(\omega)) \qquad (10)\end{aligned}$$

Expected Global Utility. Given the utility of multi-session ranked lists conditioned on each specific user browsing pattern, calculation of the expectation over all patterns is straightforward:

$$\begin{aligned}\mathbb{E}\left[\mathcal{U}(\omega)\right] &= \sum_{\omega\in\Omega} P(\omega)\mathcal{U}(\omega) \\ &= \sum_{s_1=1}^{|l_1|} \dots \sum_{s_K=1}^{|l_K|} \left(\prod_{k=1}^{K} P(s_k)\right) \mathcal{U}(\underbrace{s_1, \dots, s_K}_{\omega})\end{aligned} \tag{11}$$

4 Tractable Computation

Unfortunately, the exact utility calculation quickly becomes computationally intractable as the number and lengths of ranked lists grow. Therefore, we make an approximation. We first rewrite EGU in terms of expected gain and expected cost. Using Equation 9 we have:

$$\mathbb{E}\left[\mathcal{U}(\omega)\right] = \mathbb{E}\left[g(m(\mathcal{L}(\omega)))\right] - a\mathbb{E}\left[\text{len}(\mathcal{L}(\omega))\right] \tag{12}$$

We then approximate the gain[3](the first term above) as:

$$\mathbb{E}\left[g(m(\mathcal{L}(\omega)))\right] \approx g(\mathbb{E}\left[m(\mathcal{L}(\omega))\right]) \tag{13}$$

Thus, instead of calculating the *expected gain* with respect to different browsing patterns, we compute the gain for the *expected browsing patterns* $\mathbb{E}\left[(m(\mathcal{L}(\omega)))\right]$, i.e., the expected number of times each nugget will be read from all the ranked lists.[4]

Since the number of times each nugget will be read in a single ranked list only depends on the possible stopping positions in that list, and is independent of the stopping positions in other ranked lists, the computation can be decomposed into K terms as follows:

$$\begin{aligned}\mathbb{E}\left[m(\delta_j, \mathcal{L}(\omega))\right] &= \sum_{k=1}^{K} \mathbb{E}\left[m(\delta_j, l_k(s_k))\right] \\ &= \sum_{k=1}^{K} \sum_{s_k=1}^{|l_k|} P(s_k) m(\delta_j, l_k(s_k))\end{aligned} \tag{14}$$

where $m(\delta_j, l_k(s_k))$ denotes the number of times nugget δ_j is read in the k^{th} ranked list when the stopping position is s_k. Thus, the approximate computation

[3] Cost is easy to calculate due to its simple definition, and does not require any approximation.

[4] We can further approximate gain by moving the expectation operator further inside, i.e. $g(\mathbb{E}\left[m(\mathcal{L}(\omega))\right]) \approx g(m(\mathcal{L}(\mathbb{E}\left[\omega\right])))$, which is equivalent to calculating the gain based on the expected stopping position in each ranked list – in our case – $1/p$, i.e., the expected value of the geometric distribution with parameter p. This corresponds to the approximation used in [7] – a fixed stopping position in each ranked list. However, we do not pursue this extra approximation in the rest of this paper.

requires a sum over $O(|l_1| + |l_2| + ... + |l_K|)$ terms, instead of the $O(|l_1| \times |l_2| \times ... \times |l_K|)$ terms in the original definition, which must consider all combinations of stopping positions in the K ranked lists.

To verify the validity of the approximation, we compared the approximate calculation against the exact `EGU` calculation on randomly generated multi-session ranked lists. The approximate and exact EGU scores were found to be very close to each other.[5]

5 Utility Optimization

An important job of a distillation system is to determine how many passages to select for the user's attention, since reading the system's output requires user effort. However, relevance-based metrics like `MAP` and `NDCG` provide no incentive for the system to produce a limited-length ranked list. On the other hand, `EGU` takes into account the relevance, novelty, and cost of reading, and hence, provides an opportunity to tune the parameters of the distillation system for optimizing its utility.

We consider two ways of adjusting the lengths of the ranked lists: (i) `Fixed length ranked lists`: Using a held-off (validation) dataset, the optimal length of ranked lists (e.g., 5, 10, 20, or 50 passages) is determined, and then held fixed for the test phase, and (ii) `Variable length ranked lists`: Instead of fixing the absolute length of the ranked lists, the relevance and novelty thresholds of the system are tuned and then these thresholds are held fixed for the test phase. Only the passages whose relevance and novelty scores are both above the corresponding thresholds remain in the ranked lists. The second approach is more flexible since it allows the system to account for varying amounts of relevant and novel information in each retrieval session by adjusting the length of its ranked list accordingly.

6 Experiments

To demonstrate the effectiveness of the proposed framework for evaluating and optimizing the utility of distillation systems, we conducted controlled experiments with two representative systems on a benchmark corpus.

Dataset. `TDT4` was a benchmark corpus used in Topic Detection and Tracking (`TDT2002` and `TDT2003`) evaluations. It consists of over 90,000 articles from various news sources published between October 2000 and January 2001. This corpus was extended for distillation evaluations by identifying 12 actionable events and defining information distillation tasks on them, as described in [5, 12]. Following [12], we divided the 4-month span of the corpus into 10 chunks, each comprising 12 consecutive days. A distillation system is expected to produce a ranked list of documents at the end of each chunk, receive feedback from the user, and then produce a new ranked list for the next chunk, and so on. We

[5] See detailed results at `http://nyc.lti.cs.cmu.edu/papers/utility/`

split the data into a validation set and a test set, each consisting of 6 events corresponding to 59 and 45 queries, respectively. We use the validation set to tune the lengths of ranked lists in the two ways mentioned in Section 5, and evaluate the performance of the system on the test set.

Metrics. We measure the system performance using the proposed metric, EGU. We also include MAP (Mean Average Precision) [3], which is a popular metric in traditional IR evaluations where relevance is the sole criterion. Comparing the behavior of these two metrics[6] allows us to highlight the properties and limitations of the metrics explicitly. As for the pre-specified parameters in EGU, we use $a = 0.01$ and $\gamma = 0.1$.

Systems. We conducted experiments with various configurations of two systems: (i) Indri [10], which is a state-of-the-art retrieval engine. It performs standard relevance-based retrieval without adaptive filtering and novelty detection, and (ii) CAFÉ [12], which is a state-of-the-art distillation system that combines adaptive filtering, ranked retrieval, and novelty detection. We try two settings in CAFÉ: adaptive filtering only (CAFÉ.AF), and adaptive filtering with novelty detection (CAFÉ.AF+ND).

Ranked list length optimization. For both systems, we also try two variants of controlling the lengths of ranked lists – Fixed and Variable, as described in Section 5.

6.1 Main Results

Table 1 summarizes the results of our experiments where the performance of each system with various settings is measured using two evaluation metrics.

Table 1: Performance scores for various configurations of two systems

System	Ranked list Length	EGU	MAP
Indri	Fixed	0.3235^{*}	0.3798
Indri	Variable	$0.3273^{*\ddagger}$	0.3798
CAFÉ.AF	Fixed	0.3001	**0.5019**
CAFÉ.AF+ND	Fixed	$0.3014^{\dagger}$	0.4737
CAFÉ.AF+ND	Variable	$\mathbf{0.4701}^{\dagger\ddagger}$	0.4737

Paired t-tests with n=45 queries:
* Statistically insignificant difference ($p > 0.05$).
† Statistically significant difference ($p = 0.01$).
‡ Moderately significant difference ($p = 0.04$).

[6] Due to space limitations, we avoid detailed comparisons with other metrics like NDCG and NDCU (which are unsuitable for multi-session utility evaluation, see Section 1), and instead, limit attention to MAP as a representative relevance-based metric.

Let us first focus on the performance in terms of EGU. With Fixed-length ranked lists, the EGU scores of Indri and CAFÉ are comparable, with Indri performing slightly better. However, with Variable-length ranked lists, the score of CAFÉ.AF+ND improved from 0.3014 to 0.4701 (+56%). On the other hand, the relative improvement by varying ranked-list length in Indri is much smaller (1%). Comparing the best result of CAFÉ with that of Indri (both of which occur with Variable length ranked lists), the former outperforms the latter by 44% (0.4701 vs. 0.3273) in EGU. These results suggest two necessary conditions for good utility-based performance: (i) the system's ability to properly assess the utility scores of passages in terms of both relevance and novelty, and (ii) the ability to control the length of the ranked lists to maximize their utility. With AF and ND, the CAFÉ system satisfies the first condition; with Variable length ranked lists the system satisfies the second necessary condition. On the other hand, Indri-Variable satisfies the second condition but not the first one. Since the relevance-based scores by Indri do not accurately reflect the true utility of passages due to the lack of novelty assessment, threshold tuning on such scores did not yield the same level of utility enhancement for Indri as it did for CAFÉ.

Now let us focus on the MAP scores. Adding novelty detection (ND) to CAFÉ.AF decreases the MAP score. This does not indicate that novelty detection is useless; instead, it indicates a limitation of the MAP metric. That is, when a relevance-based ranked list is re-ranked based on both relevance and novelty, the resulting list is doomed to have decreased MAP score because the metric is relevance-based only. Another problem with MAP is its inability in accurately measuring the cost for the user to read non-informative passages. For instance, assume that there is only one relevant passage in a given chunk, and two systems, A and B, both place the relevant passage on the top of their respective ranked lists. However, system A's list contains the relevant passage only, while system B retrieves many irrelevant passages in addition. The MAP scores of both lists will be identical, i.e., 100%. However, system A is obviously better than system B in this case since users are likely to browse the system B's list beyond the top-ranking passage, and waste their effort on non-informative passages. Thus, comparing distillation systems based on MAP scores would be misleading when novelty and cost are important concerns in assessing the true utility of systems. This is evident in Table 1 – the optimal MAP scores for fixed as well as variable ranked lists is the same since MAP can never improve by truncating a ranked list.

7 Concluding Remarks

In this paper, we have proposed the first theoretical framework where non-deterministic user interactions over multi-session ranked lists are taken into account for evaluating and optimizing the utility of distillation systems. This model flexibility comes at the cost of increased computational complexity, which we address using an efficient approximation technique. We conducted the first utility-based evaluation over multiple-session search scenarios, using the TDT4 corpus of news stories, and show that a distillation system (CAFÉ) achieves 44%

better utility score than a purely relevance-based retrieval engine (`Indri`), due to multi-session adaptive filtering, novelty detection and utility-based length adjustment of ranked lists. Our framework can naturally accommodate more sophisticated probabilistic models of user behavior that go beyond the geometric distribution over stopping positions, which would be an interesting line of future research in the area of probabilistic evaluation frameworks.

Acknowledgements. This work is supported by the National Science Foundation (NSF) under grant IIS_0704689. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

1. J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 314–321, 2003.
2. O. Babko-Malaya. Annotation of Nuggets and Relevance in GALE Distillation Evaluation. *Proceedings LREC 2008*, 2008.
3. C. Buckley and EM Voorhees. Retrieval system evaluation. *TREC: Experiment and Evaluation in Information Retrieval*, pages 53–75, 2005.
4. C.L.A. Clarke, M. Kolla, G.V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666, 2008.
5. D. He, P. Brusilovsky, J. Ahn, J. Grady, R. Farzan, Y. Peng, Y. Yang, and M. Rogati. An evaluation of adaptive filtering in the context of realistic task-based information exploration. *Information Processing and Management*, 2008.
6. K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, (4):422–446, 2002.
7. K. Järvelin, S. Price, L. Delcambre, and M.L. Nielsen. Discounted cumulated gain based evaluation of multiple-query IR sessions. In *Proceedings of the 30th European Conference on Information Retrieval (ECIR 08).*
8. A. Moffat and J. Zobel. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems*, pages 1–27, 2008.
9. S. Robertson. A new interpretation of average precision. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 689–690, 2008.
10. T. Strohman, D. Metzler, H. Turtle, and W.B. Croft. Indri: A language model-based serach engine for complex queries. *Proceedings of the International Conference on Intelligence Analysis*, 2004.
11. JV White, D. Hunter, and JD Goldstein. Statistical Evaluation of Information Distillation Systems. *Proceedings of the Sixth International Language Resources and Evaluation LREC*, 8.
12. Y. Yang, A. Lad, N. Lao, A. Harpale, B. Kisiel, and M. Rogati. Utility-based information distillation over temporally sequenced documents. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 31–38, 2007.