Improving Realism of Topic Tracking Evaluation

Anton Leuski and James Allan
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
{leuski,allan}@cs.umass.edu

ABSTRACT

Topic tracking and information filtering are models of interactive tasks, but their evaluations are generally done in a way that does not reflect likely usage. The models either force frequent judgments or disallow any at all, assume the user is always available to make a judgment, and do not allow for user fatigue. In this study we extend the evaluation framework for topic tracking to incorporate those more realistic issues. We demonstrate that tracking can be done in a realistic interactive setting with minimal impact on tracking cost and with substantial reduction in required interaction.

Categories and Subject Descriptors

H.3.3 [**Information storage and Retrieval**]: Information Search and Retrieval—*information filtering, relevance feedback*

Keywords

Filtering, interactive tracking, topic detection and tracking

1. INTRODUCTION

The tasks of information filtering and topic tracking require monitoring a continuous stream of information to select only those items that are of interest to the user. There is an underlying message in discussion of the tasks that as they mature in capability, they can be deployed in an operational system with little change. However, in the laboratory evaluations that are used to measure system quality, both tasks make necessary simplifying assumptions that move them away from a realistic model of users.

Information filtering, as evaluated in the TREC workshops, is the task of monitoring a stream of documents to find those that are on a particular topic of interest (e.g., documents about the effects of osteoporosis). The system is given a brief description of the topic (that has ranged in size from several sentences to a few words) and then proceeds to look for stories on that topic. The system is permitted to adapt its representation of the topic by asking an oracle whether or not a particular document is relevant (it can, of course, adapt without asking the oracle, too). The oracle simulates a user who would judge whether the document is relevant.

Topic tracking, part of the Topic Detection and Tracking evaluation program [2], is similar in that it requires a system to monitor a stream of news stories for additional stories on the same topic. Here, all stories are news articles.[1] A topic is derived out of the events that are discussed in news, and generally incorporates all related events that grow out of some seminal happening (e.g., an earthquake might trigger a topic which would include reporting on the earthquake, rescue efforts, and so on).[2] For the tracking task, a system is presented with a small number of stories (one to four) that are known to be on the same topic. At that point, active "user" interaction ends. The system is required to determine the topic from those sample stories and then to select stories from the rest of the stream without user interaction. If the system finds a story that might be a match, it cannot ask the user for confirmation—i.e., it must make a *yes* or *no* decision about every story on its own.

In that way, TDT's tracking evaluation is less user-oriented than TREC's filtering evaluation. Filtering simulates interacting with the user to supervise the process, whereas tracking operates as if the user were not there. In this study, we are interested in bringing simulated user interaction from filtering to tracking, and then moving beyond that to make the interaction even more realistic, giving a better sense of the effectiveness of these technologies in a real-world setting. We will model users interacting with the system during the tracking process itself to explore what is plausible and what is effective in that area. There are obviously other ways a user could be involved (e.g., highlighting specific relevant text, adding outside news stories for clarification, editing the internal representation of a topic, etc.). We will defer exploration of those interactions for later research.

In the next section we outline some related work in the area of interactive tracking. In Section 3 we describe a model of our interactive tracking system and show how TREC fil-

[1] TREC filtering evaluations have generally used news articles, too, though there is not a strong emphasis on that point.

[2] Note the difference between a tracking topic and a filtering topic. The latter is subject-based whereas the former is event-based. All event-based topics are also subject-based, but the reverse is not true. It would be reasonable to filter for stories about the effects of osteoporosis, but there is no corresponding event.

SIGIR'02, August 11-15, 2002, Tampere, Finland.

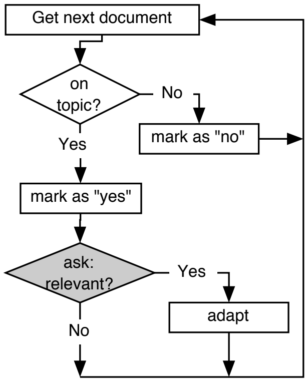

Figure 1: Shows the flow of control in TREC filtering task.

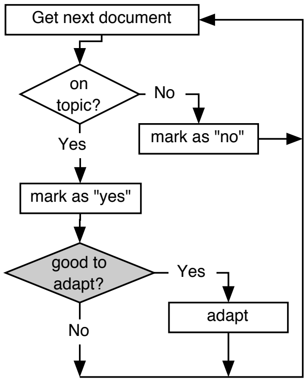

Figure 2: Shows the flow of control in TDT tracking task.

tering and TDT tracking fit into that model. We also present several more realistic variations of the task. Section 4 describes how we adjust the TDT tracking evaluation to incorporate interaction and Section 5 describes our experimental setup, including the system and evaluation corpora used. We present our results in Section 6, showing that TREC-style filtering interaction requires large numbers of user judgments and that nearly equivalent effectiveness can be realized with a more realistic usage model and substantially less user interaction. We conclude in Section 7 and describe the direction we are taking this work.

2. RELATED WORK

This work builds on the sequence of Topic Detection and Tracking evaluations that have run from 1998 through 2001 [2, 16]. The baseline system that we are using was adapted from the University of Massachusetts' TDT 2000 tracking system [3], a system that performed well in the evaluation.

This work is also strongly related to the TREC evaluations of information filtering [13]. We are considering interactive tracking, but there is no relationship between this work and the TREC interactive track [6] that focuses on user studies and the process of information retrieval.

One of the goals of the work discussed below is to reduce the amount of interaction that a user is required to perform, but to keep the effectiveness of the system unchanged. This is similar in spirit to earlier results that showed that TREC routing[3] results could be remain similar even if substantially fewer documents were judged [1]. Other approaches to reducing the user's workload include presenting summaries of single documents [11] or multiple documents [12]. Our focus is on extending the classic document at-a-time approach used in tracking and filtering, so we do not consider here how summarization would perform in these condition.

The "oracle" approach to evaluation, where a user's judgment is made by looking at a truth file, is common in information retrieval research. The adaptive tracking task uses such a technique [13] as has some work evaluating interaction strategies in the context of classic ranked retrieval [8, 9].

3. MODELING INTERACTIVE TRACKING

Figures 1 and 2 summarize the control flow in both TREC filtering and TDT tracking systems. Each system begins by extracting the next document from the stream of incoming documents. Then it decides whether the document is on-topic. The documents deemed off-topic are labeled as non-relevant and the system returns to the stream for the next document. If the document passes the test, it is marked as relevant and the system has to decide whether the document is good enough to adapt the topic representation. This last test is the main difference between the TREC filtering and TDT tracking approaches. The corresponding test block is highlighted in gray on both figures. A TREC filtering system asks the oracle (or the user in the real settings) to confirm that the document is relevant and adapts the topic representation upon receiving the confirmation. A TDT system makes its decision without any external input. A TREC system could also decide to adapt without user confirmation, but the evaluation model of TREC filtering does not encourage that: both cases are treated the same. Note that in this paper we consider only positive judgments while adapting the topic representation. A negative judgment made by either the user or the system may also be used to adapt the topic representation, however we defer this question for future study.

Both approaches make two unrealistic assumptions about the user's behavior. First, a traditional system either forces the user to judge *every* document it labels relevant as in TREC filtering or it avoids any dialog with the user as in TDT tracking. We are interested in an intermediate scenario where the system may request the user's judgments for some of the returned documents but it does not ask her to judge all of them. In her turn, the user may decide to ignore and not judge some of the documents requested by the system. For example, consider a situation when the system constantly keeps requesting judgments for documents and at some point in time the user grows tired and loses interest in the judging task. We expand the control flow for filtering and tracking to represent our model of interactive tracking (or filtering) and show it in Figure 3. The extended decision block just discussed is in gray. Here the system decides whether to involve the user and checks if the user is interested in making the judgment. If both conditions are satisfied it asks the user for the relevance judgment, otherwise it proceeds to decide

[3] TREC's routing task was a batch-oriented approximation of what eventually became filtering. It permitted massive amounts of relevance information for query adaption.

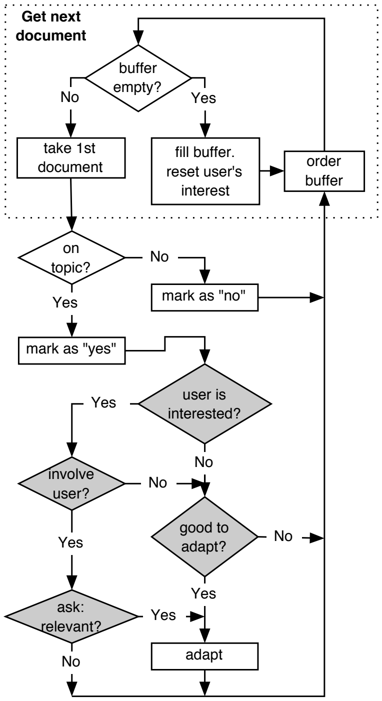

Figure 3: Shows the flow of control in the Interactive Tracking task.

whether the document can be used for adapting the topic representation on its own.

The second assumption, most apparent in the TREC filtering task, has two parts: (1) the system has to label each document before it can look at any other documents that follow and (2) the user is readily available at any time to make the judgments. An alternative to having the user at hand would be to suspend processing until she returns, but the delay may be inappropriate in many settings.

In contrast to this assumption, we consider a situation where the user interacts with the system at discrete time intervals. The system may defer its decisions until the next user's session. This can be viewed as the system collecting the incoming documents into an intermediate session buffer. At the beginning of each session the system lists the documents in the buffer for the user. If we can assume by analogy with the ranked list that the user will start at the top of the list and follow it down, examining one document after another, then the order of the documents in the session buffer is very important. For example, the system may elect to place the documents that it requests to be judged at the top of the list. After it receives the user's feedback it adapts the topic representation and possibly re-orders the rest of the list. Figure 3 shows this extended control block surrounded by a dotted line.

The complete flow of control in this extended tracking scenario is shown on Figure 3. We call this model the *Interactive Tracking* task. In the following sections we discuss how we evaluate an Interactive Tracking system.

4. EVALUATING INTERACTIVE TRACKING

In the TDT evaluation program [5] tracking is viewed as a detection task and its performance is characterized in terms of the probability of miss and false alarm errors. These error probabilities are then combined into a single cost, K, by assigning costs to miss and false alarm errors:

$$K = K_{miss} \cdot P_{miss} \cdot P_{target} + K_{fa} \cdot P_{fa} \cdot (1 - P_{target})$$

where K_{miss} and K_{fa} are the costs of a miss and false alarm, respectively, P_{miss} is the conditional probability of a miss or the proportion of relevant documents that were labeled as non-relevant by the system, P_{fa} is the conditional probability of a false alarm or the proportion of the non-relevant documents that were labeled as relevant by the system, and P_{target} is the *a priori* target probability. In this paper we compute the normalized version of the cost measure $Cost = K/K_{min}$, where

$$K_{min} = min(K_{miss} \cdot P_{target}, K_{fa} \cdot (1 - P_{target}))$$

The values for normalized costs of a miss ($C_{miss} = K_{miss} \cdot \frac{P_{target}}{K_{min}}$) and false alarm ($C_{fa} = K_{fa} \cdot \frac{(1-P_{target})}{K_{min}}$) are set to 1 and 4.9 as per conditions of the TDT tracking experiments [5] ($K_{miss} = 1$, $K_{fa} = 10$, and $P_{target} = 0.02$).

This evaluation accounts for the simple model of the user-system interaction assumed in the TDT tracking task: the user views all documents labeled relevant by the system and she does so in the order the documents were received. This interaction is *passive* as the system receives no feedback from the user.

In TREC filtering the interaction is *active* – the user is required to judge all documents returned by the system and the system is not expected to proceed without receiving the feedback, since the judgment is needed to adapt the query.

In Interactive Tracking we want to balance the system's tracking errors against the active interaction with the user. Here we differentiate among the notions of documents *presented*, *examined*, and *judged*. The documents *presented* to the user are the document labeled as relevant by the tracking system. We also consider the documents the user reads through or the *examined* documents and the *judged* documents – the documents that user labeled as relevant or non-relevant for the system. In TREC filtering these are all the same document sets. In TDT tracking it is assumed that the user will read all documents returned by the system, so the set of presented documents is the same as the set of examined documents, while the set of judged documents is empty. The Interactive Tracking model does not expect the user to read every document, so the set of examined documents might be smaller than the set of presented ones. For example, consider a situation when the user is reading the documents in the session buffer and sees too much non-relevant material. The user grows irritated and decides to stop reading – "There is nothing interesting in the news today!". We measure the user's involvement in the tracking process, or the user's *activity*, as the proportion of documents that were judged by the user among all documents examined by the user. We want to minimize the user's activity without a significant drop in the tracking cost:

$$Activity = \frac{\#\ of\ judged}{\#\ of\ examined}$$

An Interactive Tracking system buffers the incoming documents and may defer its labeling decisions until the next user's session. When the user begins the session, the system orders the documents in the buffer and presents them to the user. We assume that the user's goal is to locate the relevant material in the news stream as quickly as possible. To achieve this goal the user would prefer to have the relevant documents at the top of the list. At the same time we want to get the most improvement from our interaction with the user. If we want to request the user's feedback about some documents, we want to do it as early as possible and therefore we want to place these documents close to the top of the list. To measure the quality of the document ordering we compute the average precision of the document buffer ranked in the order it is examined.

To summarize, our goal is to keep the TDT cost value as low as possible while simultaneously reducing the measure of activity. We will in Section 6 that this is possible.

5. EXPERIMENTAL SETUP

For our study we created an Interactive Tracking system and conducted a number of simulation experiments. We simulated a user interacting with the tracking system with the help of the topic and relevance judgment information available for the TDT evaluation. In this section we describe our tracking system and the experimental data set.

5.1 System design

Our Interactive Tracking system uses the vector-space approach where each document is represented by a vector of term weights V. The weight of the ith term in the vocabulary, v_i is computed using the Inquery weighting formula, which uses Okapi's *tf* score [14] and Inquery's normalized *idf* score:

$$v_i = \frac{tf}{tf + 0.5 + 1.5\frac{doclen}{avgdoclen}} \cdot \frac{\log(\frac{colsize+0.5}{docf})}{\log(colsize+1)}$$

where *tf* is the number of times the term occurs in the document, *docf* is the number of documents the term occurs in, *doclen* is the number of terms in the document, *avgdoclen* is the average number of terms per document in the collection, and *colsize* is the number of documents in the collection. The document frequency and the collection statistics were determined from a training data set. The similarity between a pair of documents is measured by the cosine of the angle between the corresponding vectors [15].

The topic being tracked by the system is represented as a cluster of documents. At the beginning of the tracking process the cluster consists of the one single training document provided by the TDT experimental conditions (this is the condition called "$N_t = 1$"). While the system examines the incoming documents and receives judgments, it may adapt the topic representation by extending the cluster with new documents. The similarity between the cluster and an incoming document is the average similarity between the incoming document and every document in the cluster. This approach is based on the system used by the University of Massachusetts at Amherst at the latest TDT workshop [3].

We have conducted an extensive study comparing different clustering algorithms for building the topic representation. Specifically, we compared the centroid, complete-link, group average, single-link, weighted centroid, and Ward methods [7]. We considered an "ideal" tracking system to be one that discriminates between relevant and non-relevant documents using a fixed threshold on the similarity between the document and the topic representation. If a document-topic similarity is above that threshold the system labels the document as on-topic, otherwise it labels the document as off-topic. The topic representation was updated using only relevant documents in the incoming stream even if they were labeled as off-topic. The system built around the group average algorithm showed the smallest tracking cost numbers.

For each document the Interactive Tracking system selects one of the following four actions: (1) it silently labels the document as non-relevant, (2) it marks it as relevant, (3) it marks it as relevant and uses the document to adapt the topic representation, and (4) it questions the user to judge the document and if the document is relevant, the system adapts the topic representation. Note that in the last case the document is labeled as relevant for evaluation purposes. In the discussion that follows we use n, y, $y+$, and q to designate these actions, respectively.

The exact choice of the action is a function of the document-topic similarity $a(x)$. The goal of the system is to minimize the expected cost of all actions:

$$\begin{aligned} E[Cost] &= \sum_i (C_{a_i,R}P(a_i|R) + C_{a_i,N}P(a_i|N)) \\ &= \sum_i \sum_{x \in A_i} (C_{a_i,R}P(x|R) + C_{a_i,N}P(x|N)) \\ &= \sum_x (C_{a(x),R}P(x|R) + C_{a(x),N}P(x|N)) \end{aligned}$$

where $C_{a_i,R}$ and $C_{a_i,N}$ are the costs of taking the ith action a_i when the corresponding document is relevant and non-relevant, respectively, $P(a_i|R)$ and $P(a_i|N)$ are the probabilities of taking the action a_i conditioned on the document's relevance, x is the document-topic similarity as observed during the tracking process, $A_i = \{x|a_i = a(x)\}$ is the set of values for x for which we take the action a_i, and $P(x|R)$ and $P(x|N)$ are the probabilities of observing x conditioned on the document's relevance.

Note that for the case of two actions – labeling the documents as either relevant or non-relevant – the expected cost is the TDT tracking cost [5]:

$$\begin{aligned} E[Cost] &= C_{y,R}P(y|R) + C_{y,N}P(y|N) \\ &+ C_{n,R}P(n|R) + C_{n,N}P(n|N) \\ &= C_{fa}P(y|N) + C_{miss}P(n|R) \end{aligned}$$

where $C_{y,R} = C_{n,N} = 0$, $C_{y,N} = C_{fa}$, and $C_{n,R} = C_{miss}$.

Assuming that all costs $C_{a,r}$ are non-negative, to minimize the expected cost we want to minimize the individual components of the $E[Cost]$. This way for each observed document-topic similarity we select action a_{best} such that

$$a_{best} = \arg\min_a C_{a,R}P(x|R) + C_{a,N}P(x|N)$$

The costs $C_{a,r}$ serve as parameters for our system and they are defined in Table 1. Here C_1 is the cost of asking the user about a relevant document, C_2 is the cost of adapting the topic representation with a non-relevant document, C_3 is the cost of asking the user about a non-relevant document, and C_4 is the cost of not adapting the topic representation with a relevant document.

The values for the cost parameters C_i, $i = 1 \ldots 4$ were selected by optimizing the tracking system on a training data set.

Table 1: Tracking costs $C_{a,r}$.

relevance	action			
	n	y	$y+$	q
R	$C_{miss} + C_4$	C_4	0	C_1
N	0	C_{fa}	$C_{fa} + C_2$	$C_{fa} + C_3$

The distributions for the conditional probabilities $P(x|R)$ and $P(x|N)$ are represented as parametric Beta distributions with probability functions $p(x|\alpha_R, \beta_R)$ and $p(x|\alpha_N, \beta_N)$ defined as

$$p(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

The corresponding parameters α_R, β_R, α_N, and β_N were determined by running the "ideal" tracking system on a training data set.

5.2 Experimental data sets

We used two different corpora for our experiments. The corpora were provided by Linguistic Data Consortium (LDC) for TDT experiments [4]. They contain stories from a variety of news sources including AP, CNN, New York Times, Voice of America, Xinhua, and Zaobao newswire. Our first corpus consists of manually transcribed English news stories from the TDT2 LDC corpus (January-June 1998). LDC also provides definitions and relevance judgments for 100 topics on that corpus. Out of those 100 topics we selected 84 topics that had more then one relevant document on the English part of the corpus. In this paper we call this data set TDT2.

The second corpus consisted of manually transcribed English and Mandarin news stories from the TDT3 LDC corpus (October-December 1998)[4]. There are two sets of 60 topics each defined on that corpus by LDC used in TDT'00 and TDT'01 evaluations [4]. In this paper we refer to this corpus and the corresponding topic sets as TDT3/1 and TDT3/2.

We report two sets of results. The first set was obtained on TDT3/1 by the tracking system trained on TDT2. The second set of numbers was produced on TDT3/2 by the system trained on TDT3/1. The latter is dangerously close to testing on the training data (the topics are different) but mirrors the official TDT'01 evaluation settings [16].

6. RESULTS

In this section we consider four different experimental questions. First we study four tracking systems that differ in the control flow in their interaction model. Then we investigate the effects of buffering and ordering of the document in the buffer on the overall system performance. Finally, we compare different users' strategies for interacting with the system.

6.1 Document at a time

The first experimental question in this paper is to compare four different tracking systems that process one document at a time and involve different amount of user interaction and adapting. The first system, *simple tracking* is the traditional TDT tracking system that discriminates stories into relevant and non-relevant categories and does not do any adapting of the topic representation (only actions n and y are allowed). The second approach, *tracking with adapting*, extends that system by allowing adaptation of the topic representation without involving the user (actions n, y, and $y+$ are allowed). The *filtering* system asks the user about every document it marks as relevant and adapts the topic representation if the user confirms the system's decision (only actions n and q are allowed). The *Interactive Tracking* system decides when to involve the user in the tracking process and when to adapt the topic on its own (i.e., it allows all four actions).

Table 2 shows the performance of these systems on both TDT3/1 and TDT3/2. The table shows the TDT tracking cost, conditional probabilities of a miss and false alarm, the number of documents that were labeled as non-relevant and are in fact non-relevant ("#n&non"), the number of documents labeled as non-relevant but are in truth relevant ("#n&rel"), and the same statistics for documents that the system marked as relevant but did not question the user about ("$yy+$") and the documents the system did question the user about ("q"). All the numbers shown are averaged across all topics in the corresponding data set.

The Interactive Tracking system ("n, y, q, $y+$") shows 48.5% and 45.0% improvement in the tracking cost over the simple tracking system ("n, y") on TDT3/1 and TDT3/2 respectively. It requires 25.9% and 25.6% fewer feedback requests to the user than the most successful filtering system ("n, q") in exchange for a small increase in the tracking cost: 6.9% and 3.4% on TDT3/1 and TDT3/2 respectively.

6.2 Ordering of documents

In the second experimental question we consider a scenario when the user interacts with the system at discrete time intervals. Specifically, she examines the documents twice a day at 9 AM and 4 PM each weekday. We simulated this setup by using the time stamps assigned to the documents in the corpus. The system buffers the documents that arrived between the sessions and opens the session by ordering and listing out the buffer. The documents labeled as non-relevant are excluded from the list (even if they are in truth relevant, since the system does not know that). It is assumed that the user will start at the top of the list and follow it down. We compare three different sorting criteria: the action label, document time, and document-topic similarity or the score assigned by the system.

There are three possible orderings based on the action: (1) the system may first request the user's feedback for some of the documents, then it lists the rest of the documents labeled relevant ("q, $y+$, y"); (2) the system first lists all the documents it does not require the feedback on and then questions the user about the rest ("$y+$, y, q"); and (3) the system starts by listing documents that it believes to be relevant and uses for adapting without asking for the user's judgment, then presents the documents that the system needs the user's feedback, followed by the rest of the documents ("$y+$, q, y"). Inside each group the documents are sorted either by time (in increasing order) or by score, i.e., in decreasing order of document-topic similarity. These document orderings are dynamic – user feedback may affect the topic adaptation, which in turn will adjust the document scores, potentially relabeling the documents. Sampling by uncertainty [10] is another ordering approach that might work but we have not yet investigated that method.

[4] The official evaluation corpus for TDT'01 prepends this corpus with news stories from the July-September period

Table 2: Average performance of four tracking systems that differ in the amount of user involvement and topic adaption. All systems process one document at a time.

System	Data set	Actions	Cost	P_{miss}	P_{fa}	#n&	#n&	#$yy+$&	#$yy+$&	#q&	#q&
						non	rel	non	rel	non	rel
simple tracking	TDT3/1	n, y	0.1948	10.16	1.90	34312.3	18.2	665.1	95.7	0.0	0.0
track. w/ adapting		n, y, $y+$	0.1728	8.63	1.76	34372.5	13.7	605.0	100.2	0.0	0.0
filtering		n, q	0.0935	2.08	1.48	34476.6	3.8	0.0	0.0	500.8	110.1
interactive track.		n, y, q, $y+$	0.1000	2.55	1.52	34460.3	4.1	144.5	30.0	372.7	79.8
simple tracking	TDT3/2	n, y	0.1475	10.65	0.84	31911.3	4.8	253.9	24.1	0.0	0.0
track. w/ adapting		n, y, $y+$	0.1269	8.57	0.84	31898.3	3.9	266.9	25.0	0.0	0.0
filtering		n, q	0.0785	3.82	0.82	31912.0	1.3	0.0	0.0	253.3	27.6
interactive track.		n, y, q, $y+$	0.0812	3.97	0.85	31900.1	1.3	75.1	8.5	190.0	19.1

Table 3 shows the performance of the Interactive Tracking system for these combinations of orderings. Additionally, the first two rows present the system performance for orderings that ignore the action assignments and sort the documents based on time or score only. Note that the former is equivalent to presenting the documents in the order they arrive into the system (i.e., the results for the Interactive Tracking system in Table 2).

We observe a small variation in the tracking cost due to the document ordering. The earlier the system requests the user feedback, the smaller the cost. The variations in the cost are small as all system requests are answered in any scenario. The user's activity is also constant. There are large differences in the buffer precision depending on the sorting order. It is not surprising that ordering the documents by their similarity to the topic increases the precision. However, some improvement in the tracking cost comes at the expense of the precision: the document ordering scenario that has the smallest cost – documents ordered by action with the document the system requests user's judgments going first and then by score ("q, $y+$, y; score") – shows a significant drop in precision (16.0% and 15.4% on TDT3/1 and TDT3/2 respectively) when compared to the second best scenario where the documents are ordered by score.

6.3 Buffering schedule

The third experimental question is how the frequency of the user-system interaction affects the Interactive Tracking performance. We considered four different schedules for the interaction: the user examines the documents (1) twice every weekday at 9 AM and 4 PM; (2) once every weekday at 9 AM; (3) once every Monday, Wednesday, and Friday at 9 AM; and (4) once every Monday at 9 AM. Countless other schedules are possible, but we will show that it does not matter much. The documents are ordered using the best-cost ordering scenario – documents ordered by action with the document the system requests user's judgments going first and then by score. It is assumed that the user answers all system requests for relevance judgments.

Table 4 shows that the buffering schedule has no effect on the tracking cost. The tracking cost does not degrade if the user is not present to judge the documents all the time as required in TREC filtering. However, it still requires the user to answer all system requests for relevance judgments – at the end of Section 6.4 we show that the buffering schedule has a more pronounced effect on the system's performance if this condition is not met. The user's activity goes up slightly as the number of interactions decreases – there is a higher proportion of the system requests for judgments among all documents examined by the user. The average precision of the document buffer decreases significantly. As the user works less often with the system, she has to wade through more non-relevant material before finding relevant information at each session.

Table 5: Performance of the Interactive Tracking system with the user limiting the number of her answers to the system's requests to a fixed number N. The results were obtained on TDT3/1 data set.

N	Cost	P_{miss}	P_{fa}	Activity	Precision	number
				(%)	(%)	examined
all	0.0992	2.48	1.52	72.29	88.28	625.9
10	0.1027	2.52	1.58	49.80	87.36	651.2
5	0.1107	3.14	1.62	34.88	86.70	666.0
3	0.1160	3.43	1.67	24.49	85.54	678.5
2	0.1168	3.66	1.64	18.22	84.49	665.8
1	0.1391	5.56	1.71	9.80	81.49	686.1

6.4 Interaction strategies

Our previous experiments assumed that the user is willing to examine all the documents labeled as relevant by the system and she is also willing to answer all system requests for relevance judgments. In this section we investigate the question of the user limiting her activity to more realistic scenarios. We consider a user that interacts with the system twice every weekday at 9 AM and 4 PM. The system lists the documents in the buffer in the order of decreasing document-topic similarity. Recall from Section 6.2 that such an ordering results in the best average precision with the second best tracking cost.

The first strategy we consider is when the user limits her active interaction with the system (or the number of judged documents is less than the number of examined documents). Table 5 shows how the Interactive Tracking system performance changes when the user restricts her answers on the system requests for relevance judgments to a fixed number per session. For example, from the fourth row in the table we see that if the user makes at most 3 relevance judgments per session, the tracking cost increases by 16.9%, while the activity goes down by 66.1%, the precision decreases by 3.1%, and the number of examined documents grows by 8.4%.

This strategy assumes that the user, while limiting her active interaction, still examines all documents labeled as relevant by the system (the number of examined documents equals to the number of presented documents). That re-

Table 3: The effect of document ordering in the session buffer for the Interactive Tracking system.

Data set	order of documents		Cost	P_{miss}	P_{fa}	Activity (%)	Precision (%)
TDT3/1		time	0.1000	2.55	1.52	72.18	58.24
		score	0.0992	2.48	1.52	72.29	88.28
	q, $y+$, y	time	0.0995	2.55	1.51	72.76	58.23
		score	0.0986	2.47	1.51	72.97	74.14
	$y+$, y, q	time	0.1014	2.60	1.54	71.09	64.24
		score	0.1008	2.54	1.54	71.26	71.72
	$y+$, q, y	time	0.1000	2.56	1.52	72.12	73.36
		score	0.0992	2.48	1.52	72.29	88.28
TDT3/2		time	0.0812	3.97	0.85	71.43	57.61
		score	0.0811	3.97	0.84	71.45	80.13
	q, $y+$, y	time	0.0802	3.87	0.85	71.73	57.37
		score	0.0802	3.87	0.85	71.82	67.76
	$y+$, y, q	time	0.0813	3.97	0.85	70.63	64.52
		score	0.0815	3.99	0.85	70.71	68.75
	$y+$, q, y	time	0.0812	3.97	0.85	71.34	71.00
		score	0.0811	3.97	0.84	71.45	80.13

Table 4: The effect of the buffering schedule on the Interactive Tracking performance.

Data set	schedule	number of batches	Cost	P_{miss}	P_{fa}	Activity (%)	Precision (%)
TDT3/1	9am, 4pm, every weekday	90.4	0.0986	2.47	1.51	72.97	74.14
	9am, every weekday	45.6	0.0986	2.47	1.51	73.17	68.75
	9am, Mon, Wed, Fri	18.6	0.0980	2.39	1.51	73.66	63.05
	9am, Mon	9.9	0.0978	2.38	1.51	74.36	58.18
TDT3/2	9am, 4pm, every weekday	83.0	0.0802	3.87	0.85	71.82	67.76
	9am, every weekday	41.9	0.0802	3.87	0.85	72.04	63.13
	9am, Mon, Wed, Fri	17.1	0.0808	3.94	0.84	72.83	55.74
	9am, Mon	9.1	0.0810	3.96	0.84	72.71	52.31

quires the user to wade through the whole document set returned by the system, which may contain a lot of non-relevant material. The process is long and very tedious. What if the user stops reading the stories and “gives up” on the system after seeing too many non-relevant documents? In this case the number of examined documents is lower than the number of presented documents. The documents presented but not examined are evaluated as marked non-relevant and may generate additional misses.

Table 6 shows the performance of the Interactive Tracking system if the user stops after seeing a fixed number of non-relevant documents in the session buffer and *declares* the rest of the document in the buffer to be non-relevant. We observe a small (3.6%) improvement in tracking cost if the user gives up after at most 30 non-relevant documents. There is no change in both the activity and precision.

As the last strategy in this paper we consider a scenario where the user gives up when she sees N times more non-relevant documents than the relevant ones. If there are no relevant documents at the top of the session buffer, she stops after 5 non-relevant documents. Table 7 shows 13.1% improvement in tracking cost and 31.2% reduction in the number of examined documents for $N = 20$ versus the case when the user examines all document labeled relevant by the system (Table 3, the second row).

Finally, we take a look on how the buffering schedule affects the effectiveness of the user’s strategy. Table 8 shows the same data as presented in Table 7 with the only difference being that the user interacts with the system once a week instead of twice a day. We observe a significant increase in cost and precision. If the user is answering every request for relevance judgments posed by the system and she is not reading through all the material presented by the system, she must be willing to adjust her interaction strategy based on her viewing schedule.

Table 6: Performance of the Interactive Tracking system when the user stops examining the documents after seeing a fixed number N of non-relevant stories.

N	Cost	P_{miss}	P_{fa}	Activity (%)	Precision (%)	number examined
all	0.0992	2.48	1.52	72.29	88.28	625.9
50	0.0973	2.63	1.45	72.62	88.30	606.7
40	0.0959	2.67	1.41	72.90	88.30	594.4
30	0.0957	2.88	1.37	72.94	88.37	577.9
20	0.0968	3.42	1.28	72.98	88.66	546.7
10	0.1009	5.03	1.03	73.52	89.81	460.7
3	0.1653	14.05	0.51	72.84	93.05	270.3
1	0.3162	30.59	0.21	66.88	100.00	146.9

7. CONCLUSION AND FUTURE WORK

We have compared the TREC filtering and TDT tracking models and highlighted the simplifications that were made for evaluation purposes. We have extended the model of interaction that they use to a more realistic one—viz., by allowing user interaction at intervals, by reducing the amount of input required from the user, and by allowing for user

Table 7: Performance of the Interactive Tracking system when the user stops examining the retrieved documents after she sees N times more non-relevant documents than the relevant ones. The user interacts with the system twice every weekday.

N	Cost	P_{miss}	P_{fa}	Activity (%)	Precision (%)	number examined
1	0.1463	11.32	0.67	74.60	91.53	327.7
2	0.1099	7.20	0.77	74.71	90.91	364.9
3	0.0987	5.81	0.83	74.51	90.57	382.9
5	0.0912	4.80	0.88	74.31	90.80	401.5
10	0.0874	4.15	0.94	73.88	90.63	420.7
20	0.0863	3.90	0.97	73.89	90.42	430.7

Table 8: Performance of the Interactive Tracking system when the user stops examining the retrieved documents after she sees N times more non-relevant documents than the relevant ones. The user interacts with the system once a week.

N	Cost	P_{miss}	P_{fa}	Activity (%)	Precision (%)	number examined
1	0.2613	24.87	0.26	71.59	86.08	169.8
2	0.2233	20.56	0.36	71.37	84.96	210.9
3	0.1812	15.97	0.44	71.74	82.30	236.5
5	0.1492	12.31	0.53	72.36	81.49	270.8
10	0.1386	10.70	0.65	71.49	80.88	308.3
20	0.1380	10.28	0.72	71.76	80.42	332.4

fatigue.

We showed that we can decrease the amount of necessary judgments from a user with only a small cost penalty. We next showed that arriving documents can be "batched up" with no significant impact on cost, although the precision of the batch improves when the batches are smaller (i.e., less time elapses). We also demonstrated that the documents in a batch are best presented to the user in order by their similarity to the topic (by score). In that case, one can optimize for cost by putting documents needing judgments first, or for precision by putting first those that the system is positive are relevant (i.e., no judgment requested) . Finally, we demonstrated that if the user gives up (stops providing necessary judgments) early, the cost of the final output is noticeably higher. However, if the user examines enough of the batch to be confident that the remainder is non-relevant, the cost stays low.

In sum, there are reasonable interactive strategies that do not grossly harm the cost of tracking. An interactive tracking system should be built with these results in mind.

The tracking system discussed in this paper uses fixed models of score distributions for relevant and non-relevant documents. These distributions are determined during the training phase of the experiments. We are currently investigating approaches that would allow to adjust the distributions on topic-by-topic basis using the user feedback.

In this paper we assumed that the documents are presented to the user in the form of the ranked list. There are alternative document set presentations that were proved to be more effective in directing the user towards relevant document such as different form of clustering [8]. We are interested in examining those approaches in the context of the Interactive Tracking system.

Acknowledgments

This material is based on work supported in part by the CIIR and in part by SPAWARSYSCEN-SD grant number N66001-99-8912. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

8. REFERENCES

[1] J. Allan. Relevance feedback with too much data. In *Proceedings of ACM SIGIR*, pages 337–343, 1995.

[2] J. Allan, editor. *Topic Detection and Tracking: Event-based Information Organization*. Kluwer Academic Publishers, Boston, 2002.

[3] J. Allan, V. Lavrenko, D. Frey, and V. Khandelwal. UMass at TDT 2000. Notebook publication for participants only, Nov. 2001.

[4] C. Cieri, S. Strassel, D. Graff, N. Martey, K. Rennert, and M. Liberman. Corpora for topic detection and tracking. In J. Allan, editor, *Topic Detection and Tracking: Event-based Information Organization*, pages 33–66. Kluwer Academic Publishers, 2002.

[5] J. G. Fiscus and G. R. Doddington. Topic detection and tracking evaluation overview. In J. Allan, editor, *Topic Detection and Tracking: Event-based Information Organization*, pages 17–31. Kluwer Academic Publishers, Boston, 2002.

[6] W. Hersh and P. Over. The TREC-9 interactive track report. In *Proceedings of the Text Retrieval Conference (TREC-9)*, pages 41–50, 2001.

[7] G. N. Lance and W. T. Williams. A general theory of classificatory sorting strategies: 1. hierarchical systems. *Computer Journal*, 9:373–380, 1967.

[8] A. Leuski. *Interactive Information Organization: Techniques and Evaluation*. PhD thesis, University of Massachusetts at Amherst, May 2001.

[9] A. Leuski and J. Allan. Evaluating a visual navigation system for a digital library. *International Journal on Digital Libraries*, 3(2):170–184, 2000.

[10] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of ACM SIGIR*, pages 385–404, 1994.

[11] I. Mani, T. Firmin, D. House, G. Klein, B. Sundheim, and L. Hirschman. The TIPSTER SUMMAC text summarization evaluation. In *Proc. of EACL'99*, 1999.

[12] K. McKeown and D. Radev. Generating summaries of multiple news articles. In I. Mani and M. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, Cambridge, Massachusetts, 1999.

[13] S. Robertson and D. A. Hull. The TREC-9 filtering track final report. In *Proceedings of the Text Retrieval Conference (TREC-9)*, pages 25–40, 2001.

[14] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In D. Harman and E. Voorhees, editors, *Third Text REtrieval Conference (TREC-3)*. NIST, 1995.

[15] G. Salton. *Automatic Text Processing*. Addison-Wesley, 1989.

[16] Proceedings of the TDT 2001 workshop. Notebook publication for participants only, 2001.