# H5 at TREC 2008 Legal Interactive: User Modeling, Assessment & Measurement

Christopher Hogan
chogan@h5.com

Dan Brassil
dbrassil@h5.com

Shana M. Rugani, Esq.
srugani@h5.com

Jennifer Reinhart
jreinhart@h5.com

Misti Gerber
mgerber@h5.com

Teresa Jade
tjade@h5.com

H5
San Francisco, CA

## ABSTRACT

Treating the information retrieval task as one of classification has been shown to be the most effective way to achieve high performance on a particular task. In this paper, we describe a hybrid human-computer system that addresses the problem of achieving high performance on IR tasks by systematically and replicably creating large numbers of document assessments. We demonstrate how User Modeling, Document Assessment and Measurement combine to provide a shared understanding of relevance, a means for representing that understanding to an automated system, and a mechanism for iterating and correcting such a system so as to converge on a desired result.

## 1. INTRODUCTION

The extraordinary effectiveness of the Relevance Feedback (RF) paradigm is well established. Recent work [19] treating the information retrieval task as a form of classification has demonstrated that the most effective way to achieve high performance on a particular task is to acquire a large number of document assessments. How these assessments are acquired, however, is often left unspecified: within evaluations, such as the TREC series of conferences, assessments performed for a particular task one year are reused for Relevance Feedback the next. In real world, time-synchronous tasks, we cannot wait for assessments before addressing the task: such assessments, if they are to be used, must be created while addressing the task. In this paper, we describe a hybrid human-computer system that addresses the problem of achieving high performance on IR tasks by systematically and replicably creating large numbers of document assessments.

The impact of large number of document assessments has been indirectly tested in previous TREC tasks, including those within the Legal Track [18]. In several cases, TREC tasks have been created to test the capabilities of Relevance Feedback systems. Testing such systems, however, imposes a fundamental challenge to the organizers of such a task: (non-pseudo) relevance feedback presumes the existence of feedback judgments by a user who is knowledgeable about the topic. Generating such assessments, however, is a potentially expensive proposition, and acquiring a sufficient quantity of assessments to test the asymptotic properties of the tested systems is even more so. A simple accommodation is therefore applied, wherein assessments produced for a topic during evaluation of the *ad-hoc* task in previous years are reused to stand in for actual relevance assessments within the RF task in subsequent years. Approaching the development of training data in this manner has the effect of easily affording the creation of large amounts of relevance data for the RF task.

The reuse of evaluation assessments in the RF task also enables us to perform a kind of *gedankenexperiment* to assess the effect of various sources of information in the IR task. In both the original *ad-hoc* task, conducted the first year, and the relevance feedback task, conducted in subsequent years, the topic is the same, allowing comparison of results. In some cases, results have improved substantially between the original run of the topic and subsequent runs. We must therefore examine what has changed between the two runs in order to afford improved results. It is possible that additional understanding of the topic by the experimenters enabled better system design, but the general focus on general designs suggests that this is not the case. It is also possible that new or improved algorithms became available in the intervening period and that these algorithms produced better results. That the RF results were produced using algorithms that have been known for some time, such as SVM, also suggests that algorithmic improvements are not responsible for the improvement. After eliminating other possibilities, it is clear that the obvious difference between the runs is also that most responsible for the exhibited improvements: namely the additional information available in the form of document assessments.

The performance of information retrieval systems is therefore seen to be a function not only of the inherent properties of the system, such as the algorithms used, but also of the information available as input to the system. Indeed, the nature of the input information, including specifically the quality and quantity of such information is a critical determinant of performance. That additional information can bring improved results has been recognized within the evaluation community for some time, as expressed through the existence of evaluation tasks such as the Interactive [10] and

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| **NOV 2008** | | **00-00-2008 to 00-00-2008** |

| 4. TITLE AND SUBTITLE | 5a. CONTRACT NUMBER |
|---|---|
| **H5 at TREC 2008 Legal Interactive: User Modeling, Assessment & Measurement** | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **H5,71 Stevenson Street,San Francisco,CA,94105** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES
**Seventeenth Text REtrieval Conference (TREC 2008) held in Gaithersburg, Maryland, November 18-21, 2008. The conference was co-sponsored bythe National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).**

14. ABSTRACT
**see report**

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | |
| **unclassified** | **unclassified** | **unclassified** | **Same as Report (SAR)** | **9** | |

HARD [1] tracks. Such evaluations have sought to bring additional information to the information retrieval task in a controlled manner, limiting both the degree and the manner of information transfer. To some extent, such limitations were driven by the need to pose a controlled experimental paradigm wherein observed improvements could be reasonably attributed to the effect of the additional information. Some of the limitations, must, however, be attributed to the difficulty of making available the resources necessary to perform information transfer experiments at large scale. The results of such experiments, while showing that additional information does indeed help, are limited by the fact that the nature and amount of information was limited by the experimental conditions.

For the first time, the Legal Interactive task admits the possibility of experimenting with large amounts of information as input to the IR task. As stated in the guidelines, the purpose of the interactive task is "... to enable the task to model more completely and accurately the conditions and objectives of e-discovery in the real world"[3]. One such property being modeled is that of the Lead Attorney as user: although document review is typically delegated to more junior attorneys or out-sourced, it is ultimately the Lead Attorney whose notion of relevance must be considered. Although the amount of time (10 hours) allocated within the Interactive Task for consultation with the Topic Authority (TA) is less than that typically experienced in document reviews of this magnitude, it is sufficient to explore more sophisticated approaches to interactive IR than have been explored in TREC in the past. In particular, because the interaction is not limited to a single exchange, iterative exploration of the topic becomes possible, as explained in our analysis, below.

At the same time that the Interactive Task guidelines provide for the possibility for incorporating larger amounts of input information into the Information Retrieval task, they also impose a much more stringent notion of relevance than has been required in the past. While relevance was in prior years based on the consensus of the reviewers, and therefore not completely defined until after the task had been completed, this year's Interative guidelines require that the TA come to a fairly complete understanding of what relevance means for a particular topic prior to providing guidance to the individual teams. Of course this ideal is not always met: the TA may change his or her mind regarding relevance, and cannot help but be influenced by discussions of relevance and exposure to particular documents. Interactive systems, therefore, must take into account not only the possibility that relevance is being defined external to a particular representation, but that the very notion of what is relevant may be changing over time. In exchange for this added complication, however, systems are provided with a single target of relevance, and are not limited by the amount of agreement that can be achieved by uninformed assessors.

The key questions to be answered are therefore these: How can we most effectively harness the knowledge that the user makes available to the system in order to improve performance? Given limitations on the user's time and attention, what is the best way to structure the conversation with the user so as to acquire the most information with the least effort? Given a certain amount of information, how best to go about the task of representing it in a way that is consumable by an automatic system? And finally, how can such a system deal with the real world exigencies posed by operating in such an environment, including a fallible user whose interpretation is subject to change? These are the questions with which this paper is concerned.

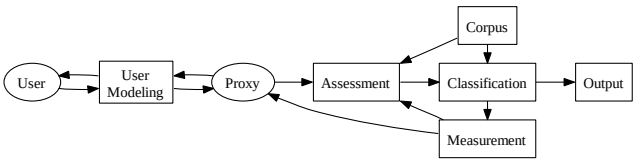


**Figure 1: System Architecture**

## 2. SYSTEM OVERVIEW

Our system comprises one main agent, the proxy, and four separate, yet interconnected, processes: User Modeling, Assessment, Classification and Measurement. A diagram is shown in Figure 1.

### 2.1 Proxy

The proxy is an internal agent who co-constructs a theory of relevance with the user via User Modeling. The proxy provides guidance to document assessors and resolves intra- and inter-assessor discrepancies to ensure that errors are resolved in favor of the proper interpretation of relevance.

### 2.2 User Modeling

User Modeling is the process by which the proxy co-determines a theory of relevance with the user (in this case the TA), iterating the process to increase the likelihood of relevance within the system's output.

### 2.3 Assessment

The assessment process is designed to (i) generate a large amount of training data (ii) of the appropriate kind (iii) with minimal error. The assessment process consists of an initial assessment of all documents of interest and subsequent error correction procedures.

### 2.4 Classification

Document-assessment pairs generated during assessment are used as training data for a supervised classification system. The classifier is trained over available assessments and the resulting model used to perform a binary classification of all documents.

### 2.5 Measurement

The performance of the classification system is regularly evaluated in order to test its efficacy. The classification system is run over all documents in the corpus. Following classification, a random sample is drawn and reviewed by document assessors. Data generated by the evaluation process are used to tune the system and may result in the proxy and user modifying the theory of relevance.

## 3. USER MODELING

### 3.1 Introduction

The effectiveness of an IR system is measured on how well it retrieves relevant text from a corpus.[1] Relevance is a derived property that entails a user and an information need: a text is deemed relevant by a user if it satisfies that user's information need (*cf.* [16]). Thus, at some level of an IR system, there must exist a representation of a user and his information need (User Modeling). Moreover, User Modeling (UM) serves as a powerful source of input by providing a mechanism by which external knowledge can be formalized into the system via query development, vocabularies, *etc.* Indeed, this year's Interactive Task appears—in part—predicated on this aspect of IR by incorporating into its design a Topic Authority to serve as a knowledgeable yet "needful" user.

UM is understood as a two-fold endeavor: (i) constructing a definition of relevance and (ii) iteratively interacting with a user to increase the likelihood of relevance in the output. We follow [17] in positing that mediated interaction, that is interaction of a user, a human intermediary and an IR system, is the most effective form of UM in IR. Within such a model, an intermediary is an "intelligent agent constructing, implementing and modifying user models in all their complexity with considerable feedback"[17].[2]

## 3.2 UM as co-construction

There are two central tenets of our approach to UM: (i) a user is seeking to resolve an "anomalous state of knowledge" and (ii) the user is unable to precisely specify what information is needed to resolve the anomalous knowledge-state [4]. These tenets underlie our own endeavors as intermediaries: we are seeking to resolve an anomalous state of knowledge as it pertains to satisfying the user's information need and we are unable to precisely define what information will satisfy the user's information need. Moreover, we recognize that users and intermediaries have access to external knowledge sources (personal knowledge, reference guides, the target corpus, *etc.*) that can be leveraged to inform and refine the model. Thus, the act of UM is a co-construction of information needs and mutual knowledge[3] in a shared representation.

We assume a model, based on [6] and depicted in Figure 2, in which the representation serves as the common ground through which external knowledge is shared, mediated, negotiated and synthesized. It is this aspect of our approach to UM that allows the intermediary to become a *proxy* for the user thereby permitting the proxy to arbitrate whether information is assessed as relevant or not relevant (which allowed H5 assessors—at the direction and guidance of the proxy—to generate nearly 8000 assessments for training data; see §4 for further discussion). Alternative approaches to UM could require the user to make all 8000 assessments to serve as training data. However, the time constraints of TREC's Interactive Task make such an approach infeasible if not impossible.

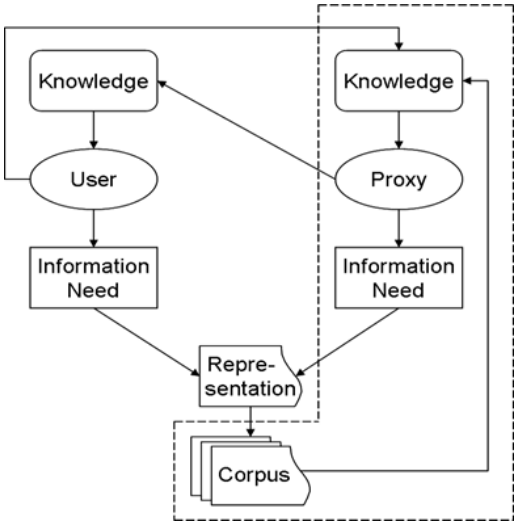


**Figure 2: Representation of User Modeling. The portion within the dashed lines is internal to the system**

For the Interactive Task, UM comprised four component areas: (i) use case, (ii) scope (iii) nuance and (iv) linguistic variability. The resultant representation is a description of subject matter, that, if found in a document, would make that document relevant (henceforth Subject Matter Model).

### 3.2.1 Use Case

Use case discussions allowed us to take into account the user's objectives: to produce to opposing counsel a set of documents deemed responsive to the Request for Production (RFP) [primary objective] and to mitigate the risk of being accused of under-producing (*i.e.* intentionally withholding responsive documents) or over-producing (*i.e.* intentionally delivering non-responsive documents) [secondary objective]. The decision to prioritize one risk over the other has far-reaching design decisions: under-production > over-production implies a narrow, more exclusive conception of relevance whereas under-production < over-production implies a broad, more inclusive conception of relevance. During UM, we learned the user felt that the risk of under-production accusations outweighed the risk of over-production accusations. Thus, when entering into scope, nuance and linguistic variability discussions, we tested where and how the user's risk-mitigation considerations might manifest.

### 3.2.2 Scope

We define scope as the breadth of concepts considered relevant by the user. When engaging in scope discussions, we seek to define the boundaries of relevance for a given conceptual domain. For example, when engaging with the user for

[1]We follow [5] in using text to be an information-bearing object. Corpus is to be taken as any collection of texts, that is, any collection of information-bearing objects.

[2]The relationship of the intermediary to the user and the IR system is one of systems boundaries. Buckland and Plaunt [8] write that "systems boundaries define what is considered the 'system' and what is considered the 'environment' ". On this definition, whether or not the intermediary is within the system is determined by how integrated the intermediary is into design of the overall system.

[3]For more on co-construction of knowledge and mutual understanding, see [7] and [15].

RFP 103[4] we sought to understand how the user interpreted *retail marketing campaigns*. We analyzed the phrase, creating questions that tested the scope of each word: types of retail outlets, the activities that constitute marketing, and the characteristics of a campaign. Based on these questions, we provided the user examples to assess, discussing the ramifications and logical extensions of her responses. We iterated the process until a shared definition was agreed to.

#### 3.2.3 Nuance

Nuance refers to the degree of specificity required to be relevant. In the context of the TREC Interactive Task, discussions of nuance and specificity centered on the semantic relations hyponymy and hypernymy[5]. For instance, it was agreed to that a hyponym of *campaign*, such as *Marlboro Ranch* (a name of a specific marketing campaign) should be considered, in and of itself, a marker of relevance, whereas the non-specific hypernym *campaign* should not be considered, in and of itself, a marker of relevance.

#### 3.2.4 Linguistic Variability

Linguistic variability is related to, but distinct from, nuance. We define linguistic variability as the variety of ways a concept can be expressed, whether lexically or syntactically. During UM, linguistic variability was discussed in the context of cigarette brands, activities that constitute retail marketing, advertising slogans, *etc.* Two approaches were evaluated: defining each concept as a closed set or defining each concept in terms of pertinent characteristics. It was determined that the user's use case (see §3.2.1) favored the latter over the former and thus, definition-by-characteristic was built into the representation.

### 3.3 Modification

Belkin [4] notes that "a change in one's state of knowledge, by virtue of having engaged with text, will be reflected in some change in the anomalous state of knowledge". Because our approach to UM assumes anomalous states of knowledge on the part of both the user and proxy, we built into the UM process a "check-in" procedure to occur during week 7 of the task: we supplied the user with 16 documents, each chosen to test whether the proxy's interpretation of relevance aligned with the user's for various aspects of the Subject Matter Model (SMM). Of the 16 documents, the user's assessment matched the proxy's for 14 of the 16 (one was resolved as H5-internal assessor error; the other discrepancy triggered a modification to the SMM).

In subsequent discussions concerning this discrepancy, two documents were discussed: `dug65f00` and `ccq45f00`. The user suggested the documents differed only in degree of specificity as it pertained to the promotion of cigarette brands via media outlets. Guidance was provided to modify the SMM in order to allow for a broader interpretation of relevance for the portion of the SMM under review. We modified the SMM which necessitated a course correction for our system (see §6 for further discussion).

[4]RFP 103—"All documents which describe, refer to, report on, or mention any "in-store", "on-counter", "point of sale", or other retail marketing campaign for cigarettes."

[5]Hyponymy is the semantic relation in which the extension of a word is subsumed in the extension of another word (*e.g. dachshund* is a hyponym of *dog*). Hypernymy is the semantic relation in which the extension of a word subsumes the extension of another word (*e.g. dog* is the hypernym of *dachshund*).

## 4. DOCUMENT ASSESSMENT

The representation and quality of training data is a, if not the, primary determiner of the success of supervised learning [13]. The presence of much irrelevant or unreliable data can significantly reduce the ability of a learner to generalize or, at best, increase the amount of training data needed to generalize properly.[6] In this section, we describe the process we used to generate training data.

### 4.1 Goals

The motivation for the process described here is to (i) generate a large amount of training data (ii) of the appropriate kind (iii) with minimal error. Assessed documents and associated annotations form the primary input to classification. As descibed in the introduction, the amount of information contained in these artifacts is determinative of a high quality result.

It is generally accepted at this time that increased amounts of training data result in improved classification accuracy [9]. During participation in the task, we assessed over 8000 documents. While this represents a large number of assessments, it represents less than 1% of the population, and by itself cannot ensure proper representability of the topic. Additional mechanisms are therefore employed to actively determine likely sources of additional relevant documents with distinct language.

As is usual, we distinguish two sources of error: random error and systematic error. Random errors are the less serious of the two types of error, and the most easily handled by ordinary error checking mechanisms, such as double-assessment. Random errors also have less serious conseqences for the classification task, and can be dealt with by increasing the number of assessments.[2] Systematic errors, on the other hand, pose a much more serious challenge, particularly for a task with a very well-defined target, such as the interactive task. If systematic error is allowed to infiltrate the assessments, the resulting system could become very highly targeted on a topic other than that co-defined with the user. Although simple consensus mechanisms cannot combat systematic error, we discuss additional properties of our error correction procedures that are deployed to minimize systematic error. Our approach to minimizing error is critically dependent on user modeling.

In order to ensure consistency between assessors, a fraction of assessed documents are independently assessed a second time by another assessor. The resulting assessments are compared, and disagreements are resolved. While this is a fairly standard operating procedure in linguistic annotation tasks, the additional constraints imposed by the inter-

[6]The problem of training data quality has also been investigated within the framework of computational learning theory, where it has been shown that while it is possible to learn in the presence of random noise [11], learning is not in general possible with malicious errors [12]. In any case, the amount of training data required to learn in the presence of noise is increased [2].

active task mean that non-standard mechanisms must be employed to address the disagreements. If the requirement of the assessment task were merely to ensure that consensus had been achieved among assessors, then it would suffice to resolve disagreements at the level of the assessors themselves, perhaps by majority vote, or by asking assessors to resolve their differences in order to come to an agreement. Such methods, however, while they are able to address random error such as might occur through an oversight on the part of an assessor who then might be persuaded to overturn his or her mistake, cannot ensure that systematic errors do not overwhelm the true intent of the topic. Bringing mismatches to the attention of the proxy, who was instrumental in the co-constuction of the theory of relevance, ensures that systematic errors are resolved in the favor of proper topic interpretation.

## 4.2 Assessment Guide

The work of assessors is informed by the theory of relevance that the proxy has co-determined with the user. In order to communicate this intent, and to give added guidance to assessors in specific cases, assessment guidelines are drawn up by the proxy and communicated to and among the assessors. It has been shown, by *e.g.* [14], that annotator agreement can be enhanced by increasing amounts of detail in an annotation guide. The purpose of the assessment guide, then, is to provide detailed direction to assessors beyond that shared between the user and the proxy. To be sure, the guidance provided by the proxy is grounded in his or her understanding of the theory as shared with the user. The assessment guide, however, provides additional direction to the assessors on how to handle known and anticipated specific instances of the topic. The assessment guide is also maintained as a continuous record of decisions made about particular cases and the reasoning behind those decisions.

## 4.3 Assessment Process

The assessment process we use is designed to address the above goals while providing a straightforward and efficient workflow. The process consists of an initial assessment performed on all documents of interest, and subsequent error correction steps, performed on samples of the population with specific characteristics. Although shown as unitary, the process actually takes place over time, and provides for evolution of interpretation as new exemplars are sought and identified.

### *4.3.1 Initial Assessment*

Assessors review documents drawn randomly using internal sampling procedures. Documents are assessed for relevance (R) or non-relevance (NR).

### *4.3.2 Relevant Passage Identification*

Following initial assessment, a portion of the documents that have been asssessed as R undergo a second round of assessment to identify relevant passages in the document. Relevant passages form one of the inputs of the classifier, where they serve to narrow the focus to highly relevant portions of potentially very long documents.

To extract relevant passages, assessors re-read R-assessed documents, and attempt to identify portions of the text that serve as indicators of relevance.

In addition to generating additional training information, passage extraction serves the secondary purpose of validating the initial assessment of relevant documents. Documents for which no relevant passage can be found are flagged for review by the proxy. Upon review, the proxy may either indicate the relevant passage, leaving the document as R, or overturn the R assessment in light of the lack of passage evidence, changing the document assessment to NR.

Although logically related to assessment, passage extraction is performed separately by an assessor other than the one who provided the initial assessment. This is done to ensure that passage extraction fulfills its function as a part of quality control, insofar as a portion of the relevant documents are assessed independently by more than one assessor.

### *4.3.3 Cross Check*

Like R documents, documents with an initial assessment of NR must be quality checked via an independent second assessment. However, unlike R documents, no relevant passages can be expected in NR documents, and there is little marginal benefit to entertaining a distinct process. Therefore, a portion of NR documents are re-reviewed by a second asssessor. Disagreements between the initial and second review are identified and flagged for review by the proxy. Upon review, the proxy may choose to leave the document as NR, or may overturn the initial assessment and make the document R.

### *4.3.4 Other Quality Controls*

In addition to the Relevant Passage Identification and Cross Check procedures described above, which have been explicitly designed for quality control, improperly assessed documents are sometimes detected in other parts of the system. Although these *ad-hoc* controls individually contribute to only a small degree, taken together they form a third branch of quality control.

Because assessors differ in their capabilities, level of expertise and knowledge of the topic, additional quality control measures are employed on a per-assessor basis. The proxy therefore randomly selects documents that have been reviewed by each assessor for spot-checking until the proxy is confident of the assessor's abilities.

# 5. MEASUREMENT

Iterative approaches to information retrieval, such as relevance feedback, clearly offer benefits over a one-shot approach. Additional retrieval iterations provide the opportunity to uncover additional relevant documents or to refine judgments on previously identified documents, and can therefore potentially boost either Recall or Precision or both. However, in order to attain any advantage over a single-shot system, the iterative system must incorporate additional knowledge during the iteration process. Nevertheless, blindly incorporating additional information with no attention paid to the current state of the system or the likely effect of such knowledge, is a blunt instrument that neither offers insight into the progress of the retrieval process nor provides direction concerning those next steps which may be most effective.

The alternative paradigm, which we espouse, incorporates explicit measurement of the system at different stages of processing. While measurement entails a certain amount of effort, the benefits are great. Among the primary benefits of measurement is the insight it provides to establish the current state of the system and the degree to which it has attained desired outcomes. While the goal of systems in the TREC task is to establish the relative effectiveness of different approaches to Information Retrieval, in real-world applications, it is often possible to set minimum standards which will ensure that the information needs of the user are being met subject to other constraints. Measurement, therefore determines not only the current state of the system, but also determines how many iterations must be performed in order to achieve the desired outcome.

In addition to providing insight into an iterative process, measurement also informs decisions made during execution of the process and provides the direction that is necessary to make considered changes in the approach. Thus, for example, if precision is seen to be low, additional effort can be expended to more carefully refine training assessments to reduce errorful R assessments. If, on the other hand, recall is low, additional efforts can be expended to find and assess additional relevant documents. Beyond the ordinary decisions regularly taken during exercise of a task, measurement can also be brought to bear to deal with extraordinary circumstances, such as the topic reinterpretation discussed in §3.3 and §6.

An important component of measurement is yield, the estimated number of relevant documents in the population. Calculation of yield is essential to establish a target for the review process and to determine progress toward that target. Yield is calculated by drawing a random sample of the entire population and assessing it according to the current interpretation of relevance. As with all aspects of relevance, however, yield is dependent on a correct interpretation of relevance, which can and does change as user modeling progresses. Yield measurements, therefore, must be interpreted with the understanding that they may change in the future, and should be repeated as relevance changes.

# 6. CASE STUDY

We present in this section an example of User Modeling requiring modification to the co-determined theory of relevance and subsequent corrections made to the training data.

## 6.1 Course correction

As mentioned in §3.3, designed into UM was a "check-in" procedure to occur during week 7 of the task. The check-in was implemented as a mechanism by which the proxy could evaluate interpretation discrepancies that might have arisen between the user and proxy, in recognition that interaction with external knowledge sources (such as the corpus) impacts knowledge states and thus might necessitate updating the co-defined theory of relevance (*cf.* [7]). During the check-in, such a discrepancy was discovered: the user presented an alternate interpretation of relevance concerning the degree of specificity required for a determination of relevance for discussions of cigarette brand promotions via media outlets. Prior to the check-in, a discussion of promoting a cigarette brand through a media outlet required a specific brand and specific media outlet be discussed for an assessment of relevant to be valid (*e.g.* A marketing budget indicating an advertisement for *Lucky Strike* being placed in *Newsweek*). Generality in either domain did not meet the definition of relevance (*cf.* Table 1).

| | Specific Media | Non-specific Media |
|---|---|---|
| Specific Brand | R | NR |
| Non-specific Brand | NR | NR |

**Table 1: Initial Definition of Relevance - Promotions and Media**

For example, `cug12d00` (Figure 3) contains a discussion of promoting `KOOL` cigarettes in various media outlets such as `True Story`, `TV Guide`, and `Us`. Because the document contains a discussion of promoting a specific brand via specific media outlets, the document was assessed as relevant.

The user's alternate interpretation allowed for non-specificity in one domain but not both (*cf.* Table 2).

| | Specific Media | Non-specific Media |
|---|---|---|
| Specific Brand | R | R |
| Non-specific Brand | R | NR |

**Table 2: Final Definition of Relevance - Promotions and Media**

Based on this change in interpretation, the definition of relevance was modified (as was the SMM and attendant materials such as the Assessment Guide). `ais35e00` (Figure 4) and `cyo18e00` (Figure 5) are examples of previously NR-assessed documents becoming R-assessed documents due to the change in interpretation.

`ais35e00` contains a discussion of promoting a specific brand `MARLBO` (Marlboro) in a non-specific media outlet `MAGAZIhC` (magazine). On the initial interpretation, the specificity of the brand was not sufficient to overcome the generality of the media outlet to trigger an R assessment. On the revised interpretation, specificity of the brand was sufficient to trigger an assessment of R even with a general media outlet. The same held true for `cyo18e00` in which running an advertisement for `Marlboro` (specific brand) in a `newspaper` (non-specific media outlet) was discussed.

The interpretation modification discussed resulted in an increase in overall yield since documents which contain discussions of placing promotional material of specific brands in non-specific media outlets like those found `ais35e00` and `cyo18e00`, constitute a fair number of the documents changed from NR to R (for further discussion of yield, see §7).

# 7. RESULTS

In the 2008 TREC Legal Track Interactive Task, we explored the application of the process described above on Topic 103. We iterated, nine times, the User Modeling process described in §3, accumulating 490 minutes of interaction with the user. The vast majority of that interaction (380 minutes; 77.55% of total time) occurred in weeks two through four in order to establish an initial definition of relevance prior to starting work on the topic. The remaining

```
BROWN d 'ILLIAMSON
ZND OUARTE21992 POSITIONING ANALYSIS
KOOL
1NSER. UNIT TOTAL REGIONAL OHIO TEST
PUBLICATION DATE SiZE PAGES pAGE Yk POSITION RATING
True Story April SP4CB 92 54-55 59 -Good
May PG4CB % 61 63 Opp. full edit Average
June SP4CB % 40-41 42 -Good
TV Guide Apil 11 2 PC lnsert 184 111 60 Opp full edit Good
April 18 2 PG Insert 208 113 54 Opp. full edit Good
Apri125 2 PG Insert 200 65 33 Opp. full edit Very Good
May 9 2 PG lnsert 212 93 44 Opp. full edit Good
May 23 2 PC Insert 152 99 65 Opp. full edit Good
May 30 2 PG lnsert 156 117 75 Opp. full edit Good
June 6 2 PG Insert 188 112A-112B 60 . Opp. full edit Good
June 20 2 PG Insert 160 140A-140B 88 Opp. full edit Average
Us April P4CB EDITION NOT RECEIVED -'MK ha''-
May P4CB % 45 47 Opp. full feature edit Very Good
June P4CB % 36 38 Opp. full cover edit Very Good
```

**Figure 3:** `cug12d00`**: Initial Assessment (R) unchanged during course correction**

```
M A G A Z I h C E S T I M A T E
CLIENT PHILIP MORRIS INC PERIOD 01/01/ 77 TO 12/31/77 DATE 08/17/77
EST N,]: 7815
PRODUCT VARIOUS DESCRIPTION MARLBO
FOOT t3A RO CTHER PROMOTIONS
LL PAGE I
REPCRT PfN5o-1-1
ISSUE CLOSE ON CANCL BILL GROSS GROSS C/D TAX AD. PR

PRC-NFL ILLUSTRATED PR.OG SALE MNTH LESS C/D t R NUNBER IC
THIRC C-3VER 4/COLOR BLEED AUG 77 06/15 08/01 06/15 07 61,095.00 60,056.38 2.CC L
PCS/ED: 1577-78 SEASON
61,095.00 60,056.38 PUBLICATICK TCTAL
61,095.00 60,056.38 ESTIMATE TCTAl,
```

**Figure 4:** `ais35e00`**: Assessment changed from NR to R during course correction**

```
Run only one ad (newspaper)' during the week beginning JuIy 21, then
start general schedule the following week.
3~. The "B" market schedule will be spread out., andiw3.11 be tailored to
jibe with d'istribution, Mr. Early will get from Mr.,0!'Connor a
good'estimate of the,time lag between distribution in major cities
and! 'IBn markets.,
MTS CELLANDCIUS
Code numbers for shipping cartons wil7.' be:
Flush - 080
Recess - 083
C
The!agency is to submit a layout for a 12M shipping case that incorporates a back-
to-back replica of'the package on the face. A sample of the Marlboro soft-pack
case will be sent to the agency Monday.
```

**Figure 5:** `cyo18e00`**: Assessment changed from NR to R during course correction**

| | Recall | | Precision | | F1 | |
|---|---|---|---|---|---|---|
| | Est. | 95% CI | Est. | 95% CI | Est. | 95% CI |
| T103 Final | 0.624 | (0.579,0.668) | 0.810 | (0.795,0.824) | 0.705 | (0.676,0.734) |
| T103 Internal | 0.687 | (0.645,0.730) | 0.823 | (0.787,0.860) | 0.749 | (0.709,0.790) |

**Table 3: Final Results**

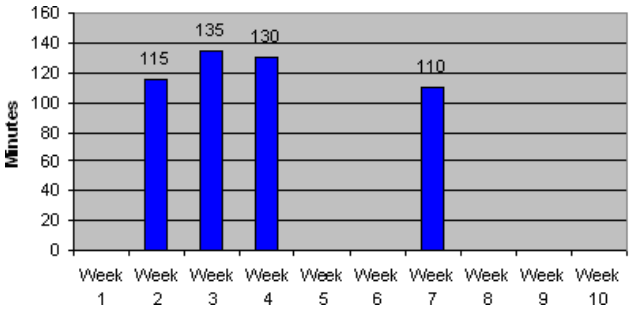


**Figure 6: Time spent with Topic Authority**

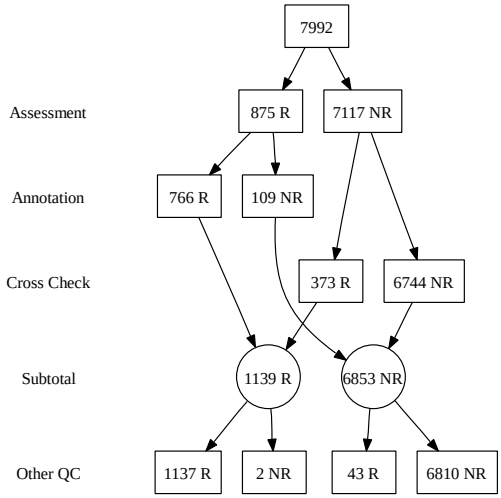


**Figure 7: Document Assessments**

110 minutes (22.45% of total time) was taken by the check-in in week seven, described in §6, above. User Modeling time is shown in Figure 6).

During our participation in the Interactive task, assessors viewed 7992 documents to provide training data for classification. Following the description in §4, Figure 7 provides a breakdown of the different assessment flows that documents took, breaking results out by number of relevant (R) and non-relevant (NR) documents. The end result was that 7992 documents were available for training: 1180 R, and 6812 NR for a training set yield of 14.76%.

Over the running of the task, measurements were conducted at regular intervals, as described in §5. Yield measurements (estimated number of relevant documents) are shown in Table 4. Note that while relevance varied from a 9.2% to 10.7%, this reflects a substantial number of documents in a population of this size. After initial uncertainty during early user modeling (weeks 1–2), yield settles on a downward trend, reaching a low of 9.2% in week 5, due to increasingly strict relevance definition. Following the check-in with the Topic Authority described in §6, however, relevance expanded, and this is reflected in the measurements with yield eventually rising to 10.5%.

| | # Relevant | % Relevant |
|---|---|---|
| Week 2 | 693,693 | 10.03874 |
| Week 3 | 744,515 | 10.77420 |
| Week 4 | 666,828 | 9.64996 |
| Week 5 | 636,587 | 9.21233 |
| Week 6 | 673,329 | 9.74403 |
| Week 7 | 720,810 | 10.43115 |
| Week 8 | 729,099 | 10.55110 |
| Week 9 | 729,099 | 10.55110 |
| **Final** | 787,762 | 11.4 |

**Table 4: Estimated Yield over Time**

Table 3 shows final, post-adjudicated results reported by TREC, as well as final internal estimates. Although the internal estimates are slightly higher than the final TREC results, the difference is well within the confidence interval.

## 8. CONCLUSIONS

We have presented a novel approach to addressing the task of large-scale information retrieval in an interactive task where relevance is defined primarily by the judgments of a single individual. The triad of User Modeling, Document Assessment and Measurement combine to provide a shared understanding of relevance, a means for representing that understanding to an automated system, and a mechanism for iterating and correcting such a system so as to converge on a desired result.

The problem of how external notions of relevance are converted into a computerized representation is deserving of further research, with consequences not only for the Legal community but for all areas of human endeavor with massive, comprehensive Information Retrieval problems.

## 9. REFERENCES

[1] J. Allan. Hard track overview in trec 2003: High accuracy retrieval from documents. In *Proceedings of TREC 2003*, page 24, 2003.

[2] D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2:343–370, 1988.

[3] J. R. Baron, B. Hedin, D. W. Oard, and S. Tomlinson. Trec-2008 legal track interactive task — guidelines. Available online at: http://trec-legal.umiacs.umd.edu/2008InteractiveGuidelines.pdf.

[4] N. Belkin. Anomolous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5:133–143, 1980.

[5] N. J. Belkin. Interaction with texts: Information retrieval as information-seeking behavior. In *Information Retrieval '93: Von der Modellierung zur Anwendung*, pages 55–66, Konstanz, September 1993. Universitaetsverlag Konstanz.

[6] N. J. Belkin, R. N. Oddy, and H. M. Brooks. ASK for information retrieval: Part I. background and theory. *Journal of Documentation*, 38(2):66–71, 1982.

[7] J. S. Brown. *A Symbiotic Theory Formation System*. PhD thesis, University of Michigan, 1972.

[8] M. K. Buckland and C. Plaunt. On the construction of selection systems. *Library Hi Tech*, 1994.

[9] K. W. Church and R. L. Mercer. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24, 1993.

[10] W. Hersh and P. Over. Trec-8 interactive track report. In *Proceedings of TREC-8*, page 57, 1998.

[11] M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45:983–1006, 1998.

[12] M. Kearns and M. Li. Learning in the presence of malicious errors. In *Proceedings of the twentieth annual ACM symposium on Theory of computing*, pages 267–280, 1988.

[13] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas. Data preprocessing for supervised leaning. *International Journal of Computer Science*, 1(2):111–117, 2006.

[14] M. Maamouri, A. Bies, and S. Kulick. Enhanced annotation and parsing of the arabic treebank. In *6th International Conference on Computers and Informatics, INFOS2008*, 2008.

[15] J. Roschelle and S. Teasley. The construction of shared knowledge in collaborative problem solving. In C. O'Malley, editor, *Computer-supported collaborative learning*, pages 69–77. Springer-Verlag, Heidelberg, Germany, 1995.

[16] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part ii: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, 58(3):1915–1933, 2007.

[17] T. Saracevic, A. Spink, and M.-M. Wu. Users and intermediaries in information retrieval: What are they talking about? In *User modeling. Proceedings of the Sixth International Conference, UM97*, pages 43–54, New York, 1997. Springer.

[18] S. Tomlinson, D. W. Oard, J. R. Baron, and P. Thompson. Overview of the trec 2007 legal track. In *Proceedings of TREC 2007*, 2007.

[19] Y. Zhu, L. Zhao, J. Callan, and J. Carbonell. Structured queries for legal search. In *Proceedings of TREC-2007*, 2007.