TREC 2015 Total Recall Track Overview

Adam Roegiest, *University of Waterloo*
Gordon V. Cormack, *University of Waterloo*
Maura R. Grossman, *Wachtell, Lipton, Rosen & Katz**
Charles L.A. Clarke, *University of Waterloo*

1 Summary

The primary purpose of the Total Recall Track is to evaluate, through controlled simulation, methods designed to achieve very high recall – as close as practicable to 100% – with a human assessor in the loop. Motivating applications include, among others, electronic discovery in legal proceedings [2], systematic review in evidence-based medicine [11], and the creation of fully labeled test collections for information retrieval ("IR") evaluation [8]. A secondary, but no less important, purpose is to develop a sandboxed virtual test environment within which IR systems may be tested, while preventing the disclosure of sensitive test data to participants. At the same time, the test environment also operates as a "black box," affording participants confidence that their proprietary systems cannot easily be reverse engineered.

The task to be solved in the Total Recall Track is the following:

> Given a simple topic description – like those used for ad-hoc and Web search – identify the documents in a corpus, one at a time, such that, as nearly as possible, all relevant documents are identified before all non-relevant documents. Immediately after each document is identified, its ground-truth relevance or non-relevance is disclosed.

Datasets, topics, and automated relevance assessments were all provided by a Web server supplied by the Track. Participants were required to implement either a fully automated ("automatic") or semi-automated ("manual") process to download the datasets and topics, and to submit documents for assessment to the Web server, which rendered a relevance assessment for each submitted document in real time. Thus, participants were tasked with identifying documents for review, while the Web server simulated the role of a human-in-the-loop assessor operating in real time. Rank-based and set-based evaluation measures were calculated based on the order in which documents were presented to the Web server for assessment, as well as the set of documents that were presented to the Web server at the time a participant "called their shot," or declared that a "reasonable" result had been achieved. Particular emphasis was placed on achieving high recall while reviewing the minimum possible number of documents.

The TREC 2015 Total Recall Track used a total of eight test collections: three for Practice runs, three for "At-Home" participation, and two for "Sandbox" participation. Practice and At-Home participation were done using the open Web: Participants ran their own systems and connected to the Web server at a public address. The Practice collections were available for several weeks prior to the At-Home collections; the At-Home collections were available for official runs throughout July and August 2015 (and continue to be available for unofficial runs).

Sandbox runs were conducted entirely on a Web-isolated platform hosting the data collections, from mid-September through mid-November 2015. To participate in the Sandbox task, participants were required to encapsulate – as a VirtualBox virtual machine – a fully autonomous solution that would contact the Web server and conduct the task without human intervention. The only feedback available to Sandbox participants consisted of summary evaluation measures showing the number of relevant documents identified, as a function of the total number of documents identified to the Web server for review.

To aid participants in the Practice, At-Home, and Sandbox tasks, as well as to provide a baseline for comparison, a Baseline Model Implementation ("BMI") was made available to participants.[1] BMI was run on all of the

*Current affiliation: University of Waterloo. The views expressed herein are solely those of the author and should not be attributed to her former firm or its clients.

[1] `http://plg.uwaterloo.ca/~gvcormac/trecvm/`.

collections, and summary results were supplied to participants for their own runs, as well as for the BMI runs. The system architecture for the Track is detailed in a separate Notebook paper titled *Total Recall Track Tools Architecture Overview* [16].

The TREC 2015 Total Recall Track attracted 10 participants, including three industrial groups that submitted "manual athome" runs, two academic groups that submitted only "automatic athome" runs, and five academic groups that submitted both "automatic athome" and "sandbox" runs.

The 2015 At-Home collections consisted of three datasets and 30 topics. The Jeb Bush emails[2] were collected and assessed for 10 topics by the Track coordinators. The "Illicit Goods" and "Local Politics" datasets, along with 10 topics for each, were derived from the Dynamic Domain datasets [3] and assessed by the Total Recall coordinators. These collections continue to be available through the Total Recall Server to 2015 participants, and were made available to 2016 participants for training purposes.

The Sandbox collections consisted of two datasets and 23 topics. On-site access to former Governor Tim Kaine's email collection at the Library of Virginia[4] was arranged by the Track coordinators, where a "Sandbox appliance" was used to conduct and evaluate participant runs according to topics that corresponded to archival category labels previously applied by the Library's Senior State Records Archivist: "Not a Public Record," "Open Public Record," "Restricted Public Record," and "Virginia Tech Shooting Record." The coordinators also secured approval to use the MIMIC II clinical dataset[5] as the second Sandbox dataset. The textual documents from this dataset – consisting of discharge summaries, nurses' notes, and radiology reports – were used as the corpus; the 19 top-level codes in the ICD-9 hierarchy[6] were used as the "topics."

The principal tool for comparing runs was a *gain curve*, which plots recall (*i.e.*, the proportion of all relevant documents submitted to the Web server for review) as a function of effort (*i.e.*, the total number of documents submitted to the Web server for review). A run that achieves higher recall with less effort demonstrates superior effectiveness, particularly at high recall levels. The traditional *recall-precision* curve conveys similar information, plotting precision (*i.e.*, the proportion of documents submitted to the Web server that are relevant) as a function of recall (*i.e.*, the proportion of all relevant documents submitted to the Web server for review). Both curves convey similar information, but are influenced differently by prevalence or richness (*i.e.*, the proportion of documents in the collection that are relevant), and convey different impressions when averaged over topics with different richness.

A gain curve or recall-precision curve is blind to the important consideration of when to stop a retrieval effort. In general, the density of relevant documents diminishes as effort increases, and at some point, the benefit of identifying more relevant documents no longer justifies the review effort required to find them. Participants were asked to "call their shot," or to indicate when they thought a "reasonable" result had been achieved; that is, to specify the point at which they would recommend terminating the review process because further effort would be "disproportionate." They were not actually required to stop at this point, they were simply given the option to indicate, contemporaneously, when they would have chosen to stop had they been required to do so. For this point, we report traditional set-based measures such as recall, precision, and F_1.

To evaluate the appropriateness of various possible stopping points, the Track coordinators devised a new parametric measure: *recall* @ $aR + b$, for various values of a and b. *Recall* @ $aR + b$ is defined to be the recall achieved when $aR + b$ documents have been submitted to the Web server, where R is the number of relevant documents in the collection. In its simplest form *recall* @$aR + b$ $[a = 1; b = 0]$ is equivalent to R-precision, which has been used since TREC 1 as an evaluation measure for relevance ranking. R-precision might equally well be called R-recall, as precision and recall are, by definition, equal when R documents have been reviewed. The parameters a and b allow us to explore the recall that might be achieved when a times as many documents, plus an additional b documents are reviewed. The parameter a admits that it may be reasonable to review more than one document for every relevant one that is found; the parameter b admits that it may be reasonable to review a fixed number of additional documents, over and above the number that are relevant. For example, if there are 100 relevant documents in the collection, it may be reasonable to review 200 documents ($a = 2$), plus an additional 100 documents ($b = 100$), for a total of 300 documents, in order to achieve high recall. In this Track Overview paper, we report all combinations of $a \in \{1, 2, 4\}$ and $b \in \{0, 100, 1000\}$.

At the time of 2015 Total Recall Track, the coordinators had hoped to be able to implement *facet*-based variants of the recall measures described above (*see* Cormack & Grossman [3]), but suitable relevance assessments for the facets were not available in time. We therefore decided to implement such measures in a future Track. The rationale for facet-based measures derives from the fact that, due to a number of factors including assessor disagreement,

[2] `https://web.archive.org/web/20160221072908/http://jebemails.com/home`.
[3] `http://trec-dd.org/`.
[4] `http://www.virginiamemory.com/collections/kaine/`.
[5] `https://physionet.org/mimic2/`.
[6] `https://en.wikipedia.org/wiki/List_of_ICD-9_codes`.

a goal of recall=1.0 is neither reasonable nor achievable in most circumstances. However, it is difficult to justify an arbitrary lower target of, say, recall=0.8, without characterizing the nature of the 20% relevant documents that are omitted by such an effort. Are these omitted documents simply marginal documents about whose relevance reasonable people might disagree, or do they represent a unique and important (though perhaps rare) class of clearly relevant documents? To explore this question, we wanted to be able to calculate the recall measures for a given run separately for each of several *facets* representing different classes of documents; a superior high-recall run should be expected to achieve high recall on all facets. This issue remains to be explored.

In calculating effort and precision, the measures outlined above consider only the number of documents submitted to the Web server for assessment. For manual runs, however, participants were permitted to look at the documents, and hence conduct their own assessments. Participants were asked to track and report the number of documents they reviewed; when supplied by participants, these numbers are reported in this Overview and should be considered when comparing manual runs to one another, or to automatic runs. It is not obvious how one would incorporate this effort formulaically into the gain curves, precision-recall curves, and *recall* @ $aR + b$ measures; therefore, the coordinators have chosen not to try.

Results for the TREC 2015 Total Recall Track show that a number of methods achieved results with very high recall and precision, on all collections, according to the standards set by previous TREC tasks. This observation should be interpreted in light of the fact that runs were afforded an unprecedented amount of relevance feedback, allowing them to receive authoritative relevance assessments throughout the process.

Overall, no run consistently achieved higher recall at lower effort than BMI. A number of runs, including manual runs, automatic runs, and the baseline runs, appeared to achieve similar effectiveness – all near the best on every collection – but with no run consistently bettering the rest on every collection. Thus, The 2015 Total Recall Track had no clear "winner."

2 Test Collections

Each test collection consisted of a corpus of English-language documents, a set of topics, and a complete set of relevance assessments for each topic. For Practice runs, we used three public document corpora for which topics and relevance assessments were available:

- The *20 Newsgroups Dataset*,[7] consisting of 18,828 documents from each of 20 newsgroups. We used three of the newsgroup subject categories – "space," "hockey," and "baseball" – as the three practice topics in the *test* practice collection.
- The *Reuters-21578 Test Collection*,[8] consisting of 21,578 newswire documents. We used four of the subject categories – "acquisitions," "Deutsche Mark," "groundnut," and "livestock" – as the four practice topics in the *tes*t practice collection.
- The *Enron Dataset* used by the TREC 2009 Legal Track [10]. We used a version of this dataset captured by the University of Waterloo in the course of its participation in TREC 2009, modified to exclude vacuous documents, resulting in a corpus of 723,537 documents. We used two of the topics from the TREC 2009 Legal Track – "Fantasy Football" and "Prepay Transactions" – as the two practice topics for the *bigtest* practice collection. The relevance assessments were derived from those rendered by the University of Waterloo team and the official TREC assessments, with deference to the official assessments.

For the At-Home runs, we used three new datasets:

- The (redacted) Jeb Bush Emails,[9] consisting of 290,099 emails from Jeb Bush's eight-year tenure as Governor of Florida. We used 10 issues associated with his governorship as topics for the *athome1* test collection: "school and preschool funding," "judicial selection," "capital punishment," "manatee protection," "new medical schools," "affirmative action," "Terri Schiavo," "tort reform," "Manatee County," and "Scarlet Letter Law." Using the continuous active learning ("CAL") method of Cormack and Mojdeh [5], the Track coordinators assessed documents in the corpus to identify as many of the relevant documents for each topic as reasonably possible.

[7] http://qwone.com/~jason/20Newsgroups/.
[8] http://www.daviddlewis.com/resources/testcollections/reuters21578/.
[9] `https://web.archive.org/web/20160221072908/http://jebemails.com/home`.

Participating Team	athome1	athome2	athome3	Kaine	MIMIC II
Catres	1M				
CCRi	1A	1A	1A		
eDiscoveryTeam	1M	1M	1M		
NINJA	1M	1M	1M		
TUW	6A	6A	6A	6A	6A
UvA.ILPS	2A	2A	2A	2A	2A
WaterlooClarke	2A	1A	1A	1A	1A
WaterlooCormack	3A	3A	3A	1A	1A
Webis	2A	2A	2A	2A	2A
WHU_IRGroup	1A				

Table 1: Participation in the TREC 2015 Total Recall Track. Table entries indicate the number of runs submitted for each test collection by a particular participating team. "M" indicates manual runs, "A" indicates automatic runs.

- The *Illicit Goods* dataset collected for the TREC 2015 Dynamic Domain Track [17]. We used 465,147 documents collected from Blackhat World[10] and Hack Forum.[11] For the *athome2* test collection, we used 10 of the many topics that were composed and partially assessed by NIST assessors for use by the Dynamic Domain Track: "paying for Amazon book reviews," "CAPTCHA services," "Facebook accounts," "surely Bitcoins can be used," "Paypal accounts," "using TOR for anonymous browsing," "rootkits," "Web scraping," "article spinner spinning," and "offshore Web sites." The Track coordinators re-assessed the documents using the CAL method of Cormack and Mojdeh [5], to identify as many of the relevant documents for each topic as reasonably possible.

- The *Local Politics* dataset collected for the TREC 2015 Dynamic Domain Track [17]. We used 902,434 articles collected from news sources in the northwestern United States and southwestern Canada. For the *athome3* test collection, we used 10 of the many topics that were composed and partially assessed by NIST assessors for use by the Dynamic Domain Track: "Pickton murders," "Pacific Gateway," "traffic enforcement cameras," "rooster chicken turkey nuisance," "Occupy Vancouver," "Rob McKenna gubernatorial candidate," "Rob Ford Cut the Waist," "Kingston Mills lock murder," "fracking," and "Paul and Cathy Lee Martin." The Track coordinators re-assessed the documents using the CAL method of Cormack and Mojdeh [5], to identify as many of the relevant documents for each topic as reasonably possible.

For the Sandbox runs, we used two new datasets:

- The Kaine Email Collection at the Library of Virginia.[12] From the 1.3M email messages from Tim Kaine's eight-year tenure as Governor of Virginia, we used 401,953 that had previously been labeled by the Virginia Senior State Records Archivist according to the following four categories: "public record," "open record," "restricted record," and "Virginia Tech shooting ([subject to a legal] hold)." Each of the four categories was used as a topic in the *kaine* test collection. The runs themselves were executed on an isolated computer installed at the Library of Virginia and operated by Library of Virginia staff.

- The MIMIC II Clinical Dataset,[13] consisting of anonymized, time-shifted, records for 31,538 patient visits to an Intensive Care Unit. We used the textual record for each patient – consisting of one or more nurses' notes, radiology reports, and discharge summaries – as a "document" in the corpus, and each of 19 top-level ICD-9 codes supplied with the dataset as a topic for the *mimic* test collection: "infectious and parasitic diseases," "neoplasms," "endocrine, nutritional and metabolic diseases, and immunity disorders," "diseases of the blood and blood-forming organs," "mental disorders," "diseases of the nervous system and sense organs," "diseases of the circulatory system," "diseases of the respiratory system," "diseases of the digestive system," "diseases of the genitourinary system," "complications of pregnancy, childbirth, and the puerperium," "diseases of the skin and subcutaneous tissue," "diseases of the musculoskeletal system and connective tissue," "congenital anomalies," "certain conditions originating in the perinatal period," "symptoms, signs, and ill-defined conditions," "injury and poisoning," "factors influencing health status and contact with health services," and

[10] http://www.blackhatworld.com/.
[11] http://hackforums.net/.
[12] http://www.virginiamemory.com/collections/kaine/under-the-hood.
[13] https://physionet.org/mimic2/.

"external causes of injury and poisoning." The runs were executed on an isolated computer installed at the University of Waterloo and operated by the Track coordinators.

Table 1 shows the number of runs submitted for each test collection by each participating team.

3 Participant Submissions

The following descriptions are paraphrased from responses to a required questionnaire submitted by each participating team.

3.1 UvA.ILPS

The UvA.ILPS team [6] used automatic methods for the At-Home and Sandbox tests that modified the Baseline Model Implementation in two ways:

1. adjusted the batch size based on the number of retrieved relevant documents and stopped after a threshold if the batch contained no relevant documents;
2. one run used logistic regression (as per the Baseline Model Implementation), while another used random forests.

3.2 WaterlooClarke

The WaterlooClarke team [9] used automatic methods for At-Home and Sandbox tests that:

1. employed clustering to improve the diversity of feedback to the learning algorithm;
2. used n-gram features beyond the bag-of-words tf-idf model provided in the Baseline Model Implementation;
3. employed query expansion;
4. used the fusion of differently ranking algorithms.

The WaterlooClarke team consisted of a group of graduate students who had no access to the test collections beyond that afforded to all participating teams.

3.3 WaterlooCormack

The WaterlooCormack team [4] employed the Baseline Model Implementation, without modification, except to "call its shot" to determine when to stop. Therefore, the gain curves, recall-precision curves, and *recall @* $aR + b$ statistics labeled "WaterlooCormack" are synonymous with the Baseline Model Implementation.

Two different stopping criteria were investigated:

1. a "knee-detection" algorithm was applied to the gain curve, and the decision that a reasonable result had been achieved was made when the slope of the curve after the knee was a fraction of the slope before the knee;
2. a "reasonable" result was deemed to have been achieved when m relevant and n non-relevant documents had been reviewed, where $n > a \cdot m + b$, where a and b are predetermined constants. For example, when $a = 1$ and $b = 2399$, review would be deemed to be complete when the number of non-relevant documents retrieved was equal to the number of relevant documents retrieved, plus 2,399. In general, the constant a determines how many non-relevant documents are to be reviewed in the course of finding each relevant document, while b represents fixed overhead, independent of the number of relevant documents.

The WaterlooCormack team consisted of Gordon V. Cormack and Maura R. Grossman, who were both Track coordinators. The Baseline Model Implementation was fixed prior to the development of any of the datasets. Cormack and Grossman had knowledge of the At-Home test collections, but not the Sandbox test collections, when the stopping criteria were chosen.

3.4 Webis

The Webis team [15] employed two methods:

1. a basic naïve approach in retrieving as many relevant documents as possible;
2. a keyphrase experiment that built on the BMI system by intelligently obtaining a list of phrases from documents judged by the API as relevant and using them as new topics for ad-hoc search.

3.5 CCRi

The CCRi team [7] represented words from the input corpus as vectors using a neural network model and represented documents as a tf-idf weighted sum of their word vectors. This model was designed to produce a compact versions of tf-idf vectors while incorporating information about synonyms. For each query topic, CCRi attached a neural network classifier to the output of BMI. Each classifier was updated dynamically with respect to the given relevance assessments.

3.6 eDiscoveryTeam

eDiscoveryTeam [13] employed a manual approach for the athome1, athome2, and athome3 test collections. Eight hours of manual search and review were conducted, on average, per topic. Two of the eight hours were spent composing 25 queries (per topic, on average) and examining their results; six hours were spent reviewing 500 documents (per topic, on average), of which *only those deemed relevant were submitted to the automated assessment server*. During the search and review process, Web searches were conducted where necessary to inform the searchers.

3.7 NINJA

The NINJA team employed a manual approach for the athome1, athome2, and athome3 test collections. One hour of manual search was conducted, in which three queries were composed, on average, per topic. Wikipedia and Google searches were used to inform the searchers. A commercial "predictive coding" tool, trained using the results of the queries, was used to generate the NINJA runs. No documents from the test collection were reviewed prior to being submitted to the automated assessment server.

3.8 catres

The catres team [12] employed a manual approach for the athome1 test collection. A group of searchers independently spent one hour each investigating each topic, after which a learning tool was used to generate the run. An average of eight manual queries were used per topic, and an average of 262 documents were reviewed. Every document reviewed by the team was also submitted to the automated assessment server.

3.9 TUW

The TUW team [14] applied six variants on the Baseline Model Implementation to all of the At-Home and Sandbox test collections. The variants included the use and non-use of a BM25 ranking, the use and non-use of stop words, and the use and non-use of tf-idf weighting.

3.10 WHU_IRGroup

The WHU_IRGroup team [1] applied iterative query expansion to the athome1 test collection.

4 Results

4.1 Gain Curves and Recall-Precision Curves

Figure 1 plots the effectiveness of the best run for each of the participating teams on the athome1 test collection. The top panel plots effectiveness as a gain curve, while the bottom panel plots effectiveness as a recall-precision curve. The gain curve shows that a number of the systems achieved 90% recall, on average, with a review effort of 10,000 documents. The recall-precision curve, on the other hand, shows that a number of the systems achieved 90%

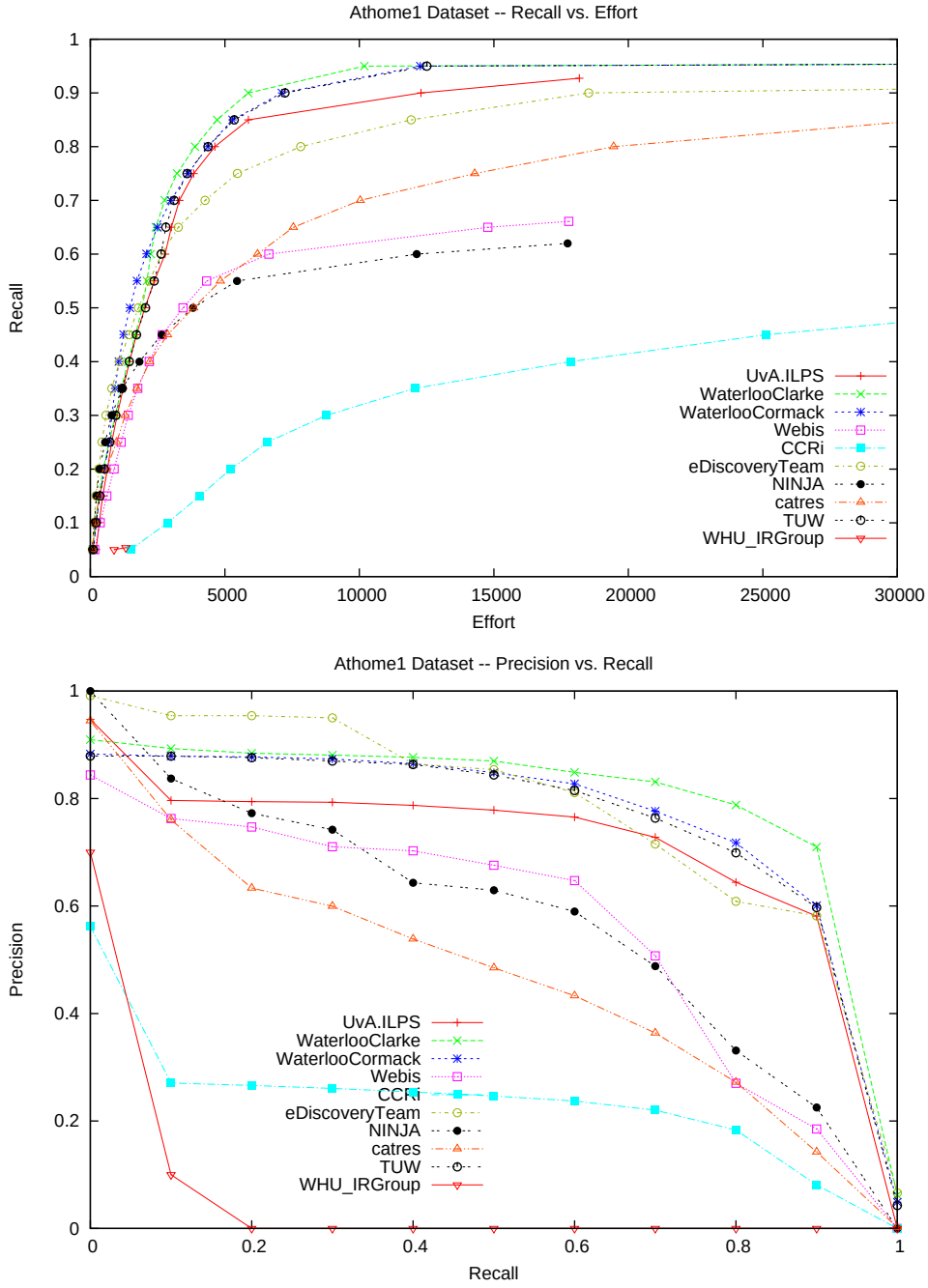

Figure 1: Athome1 Results – Average Gain and Interpolated Recall-Precision Curves.

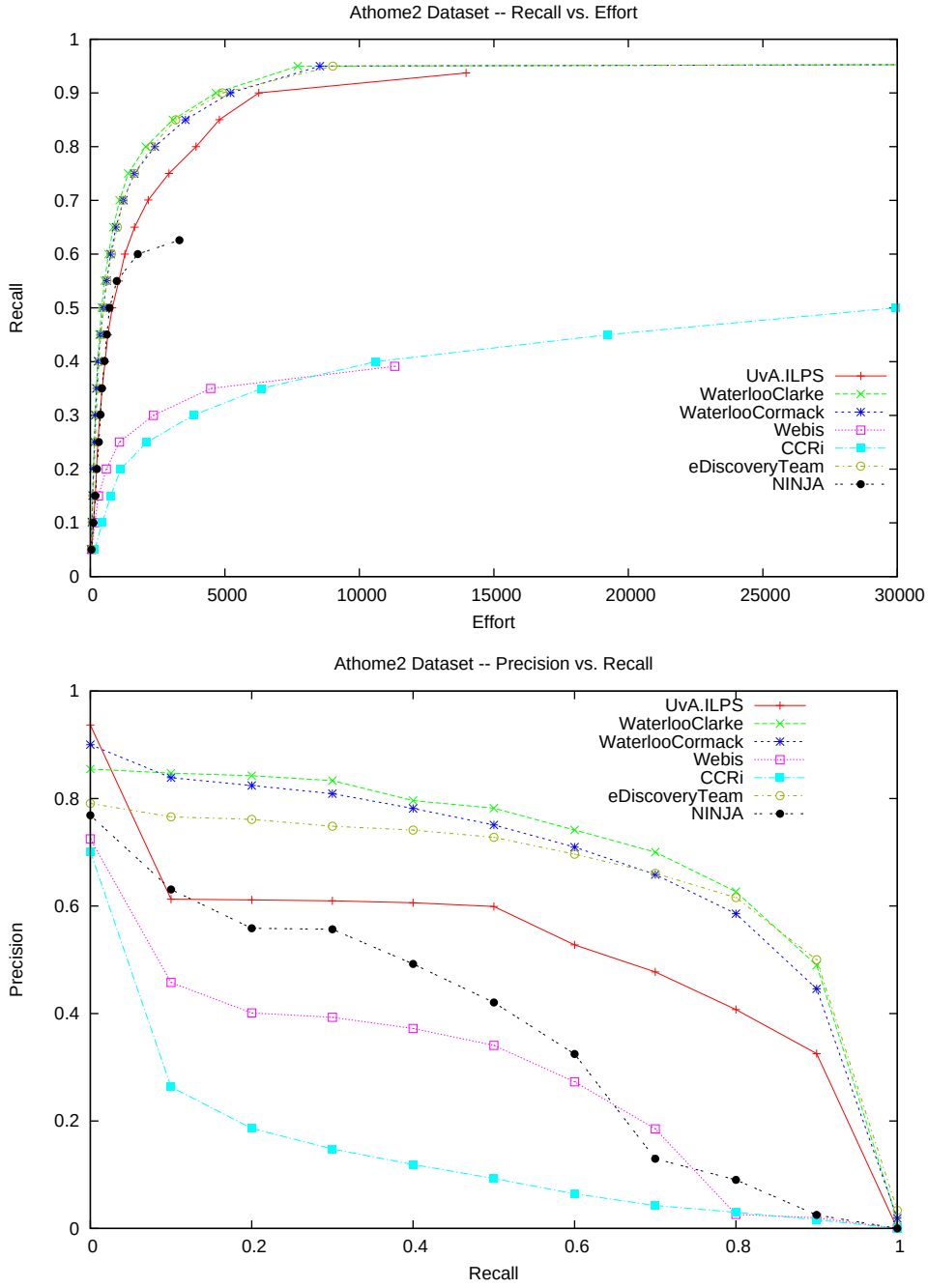

Figure 2: Athome2 Results – Average Gain and Interpolated Recall-Precision Curves.

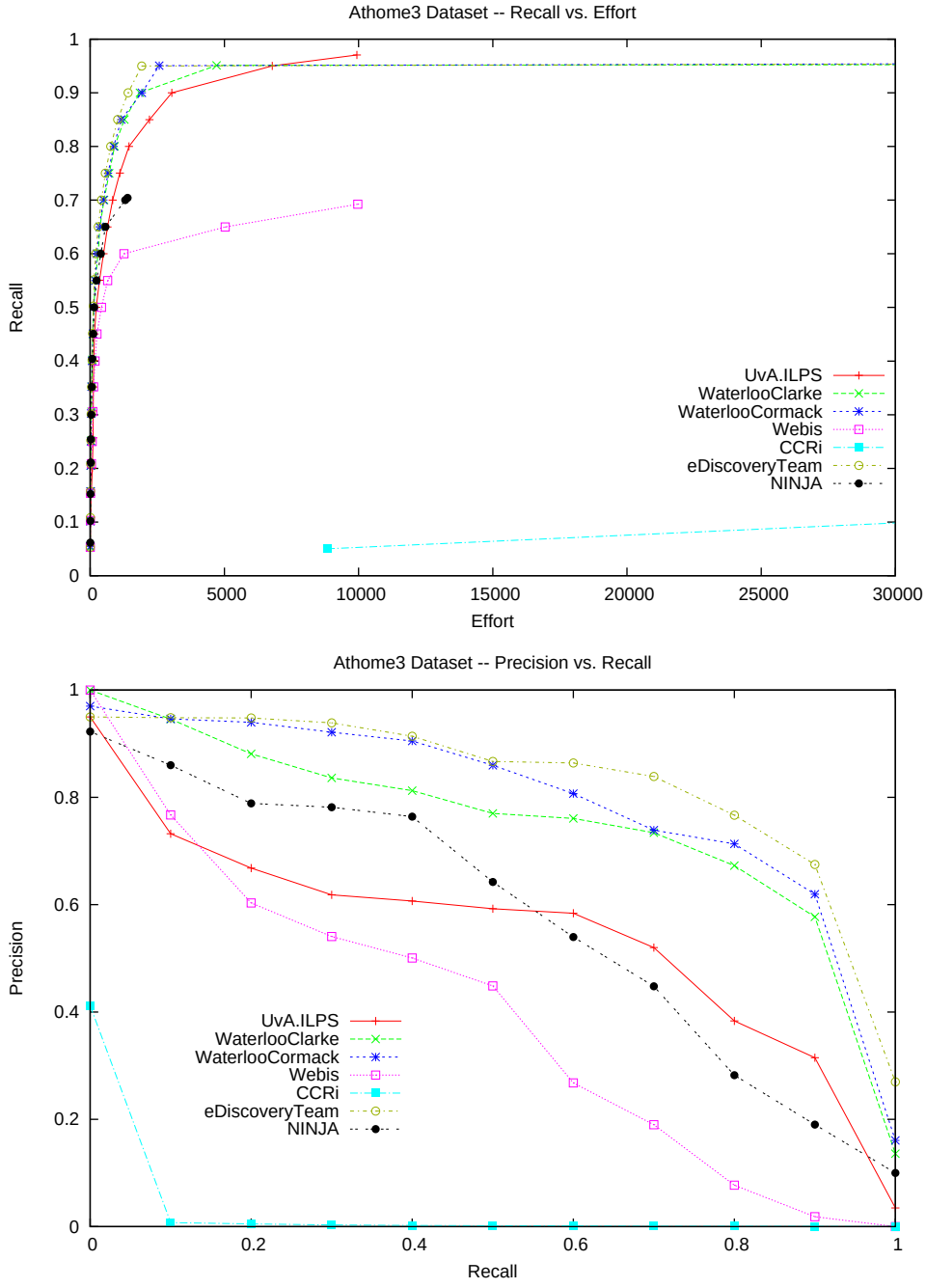

Figure 3: Athome3 Results – Average Gain and Interpolated Recall-Precision Curves.

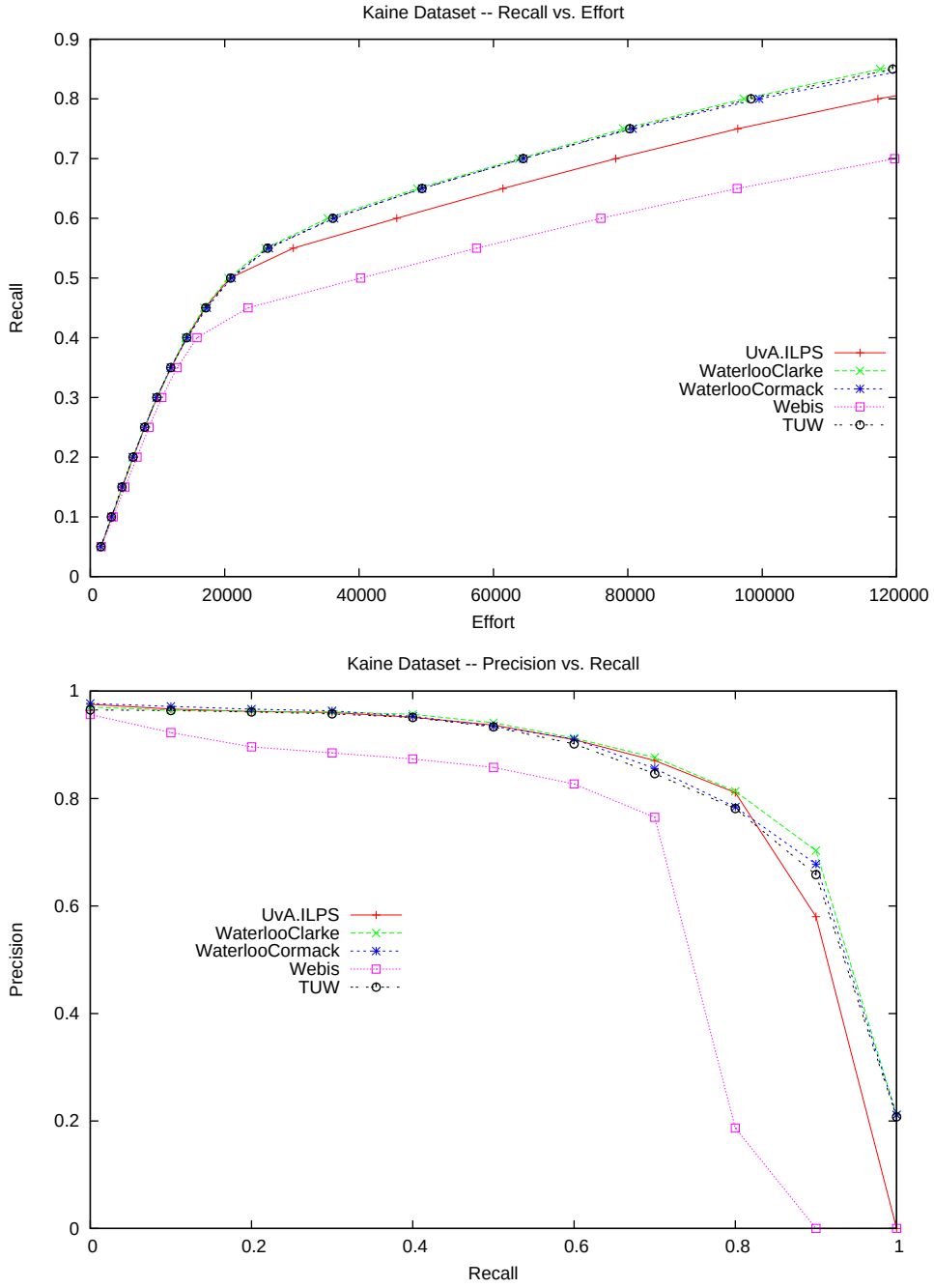

Figure 4: Kaine Results – Average Gain and Interpolated Recall-Precision Curves.

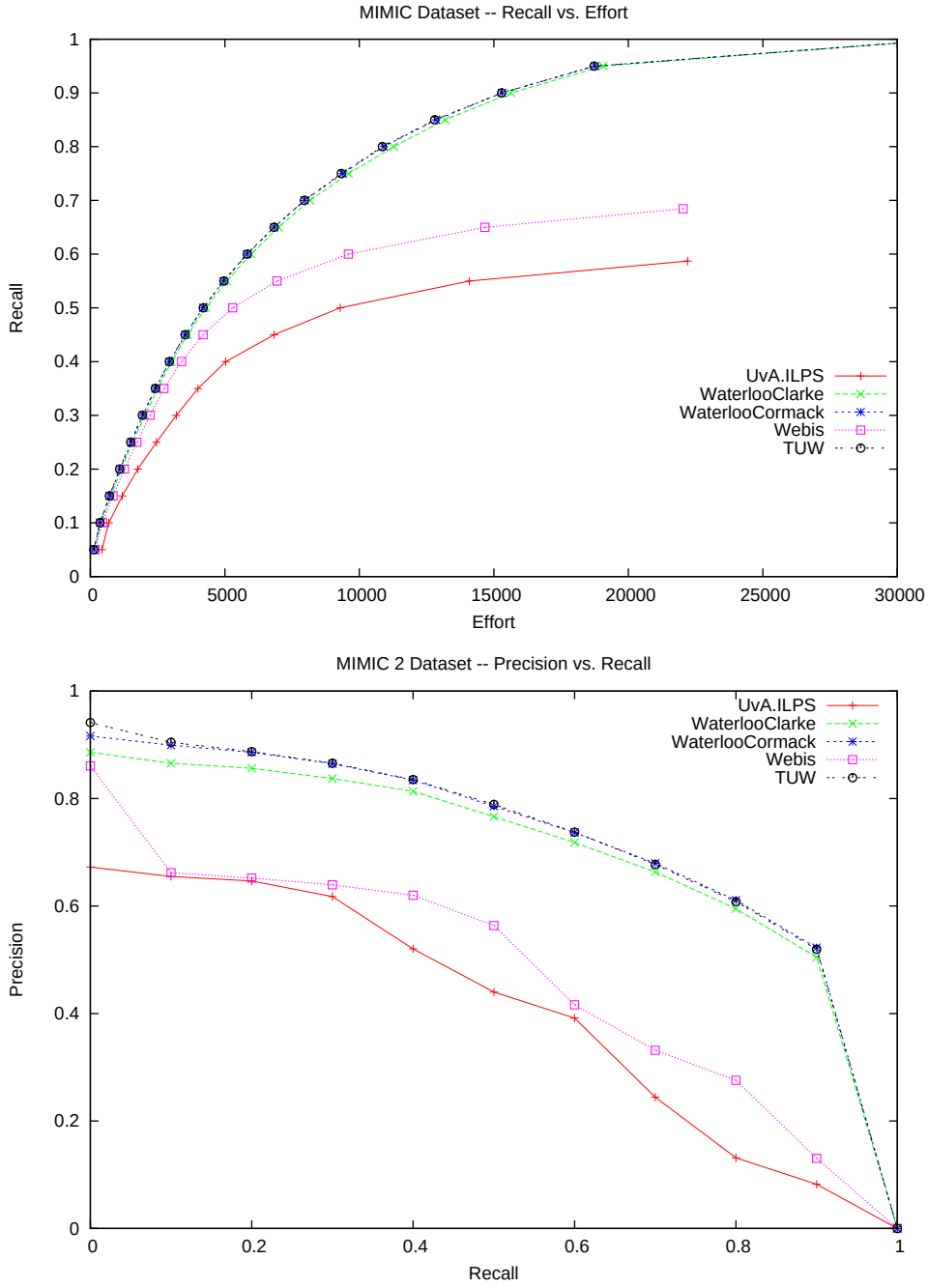

Figure 5: MIMIC II Results – Average Gain and Interpolated Recall-Precision Curves.

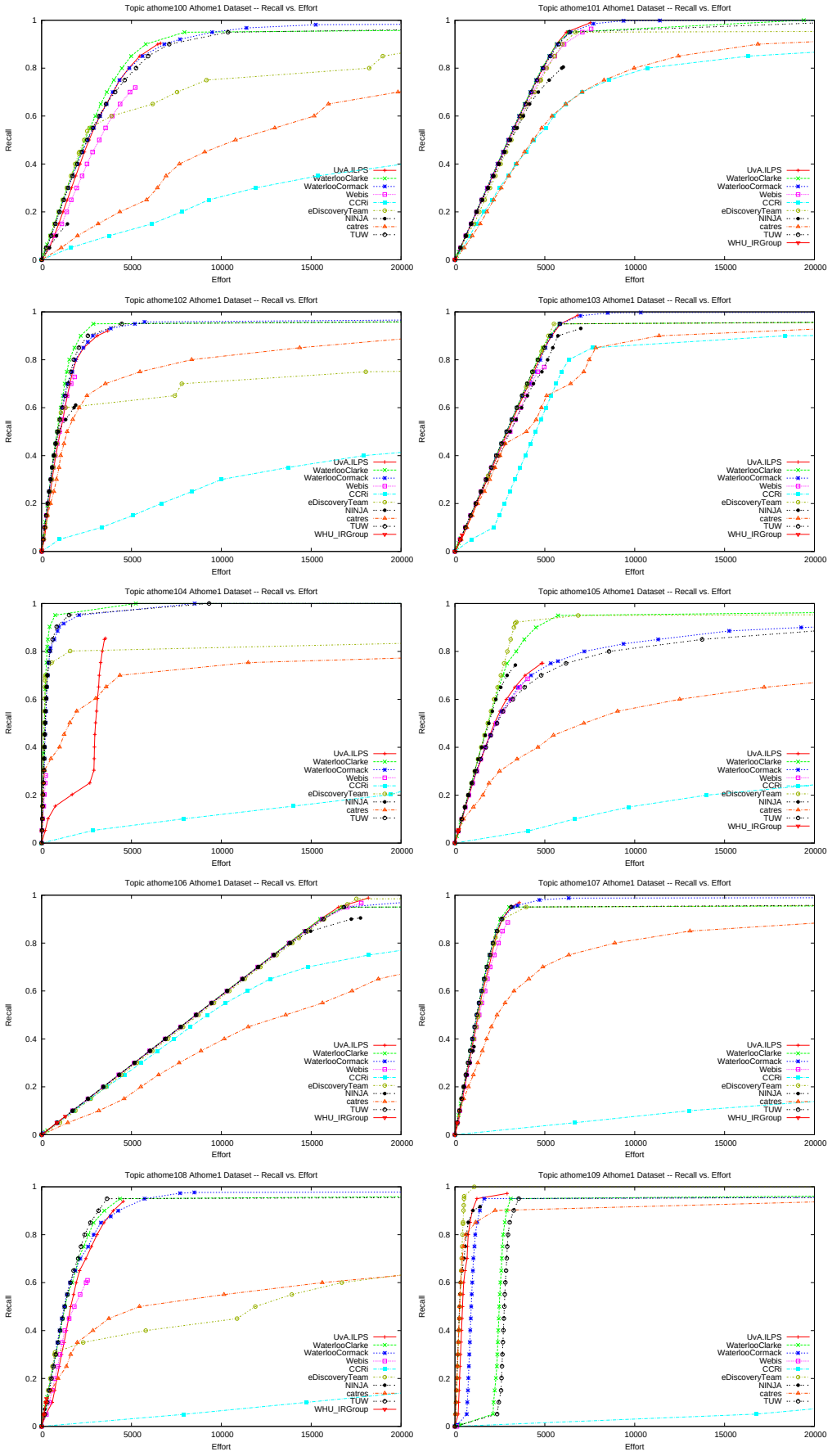

Figure 6: Per-Topic Gain Curves for the Athome1 Test Collection.

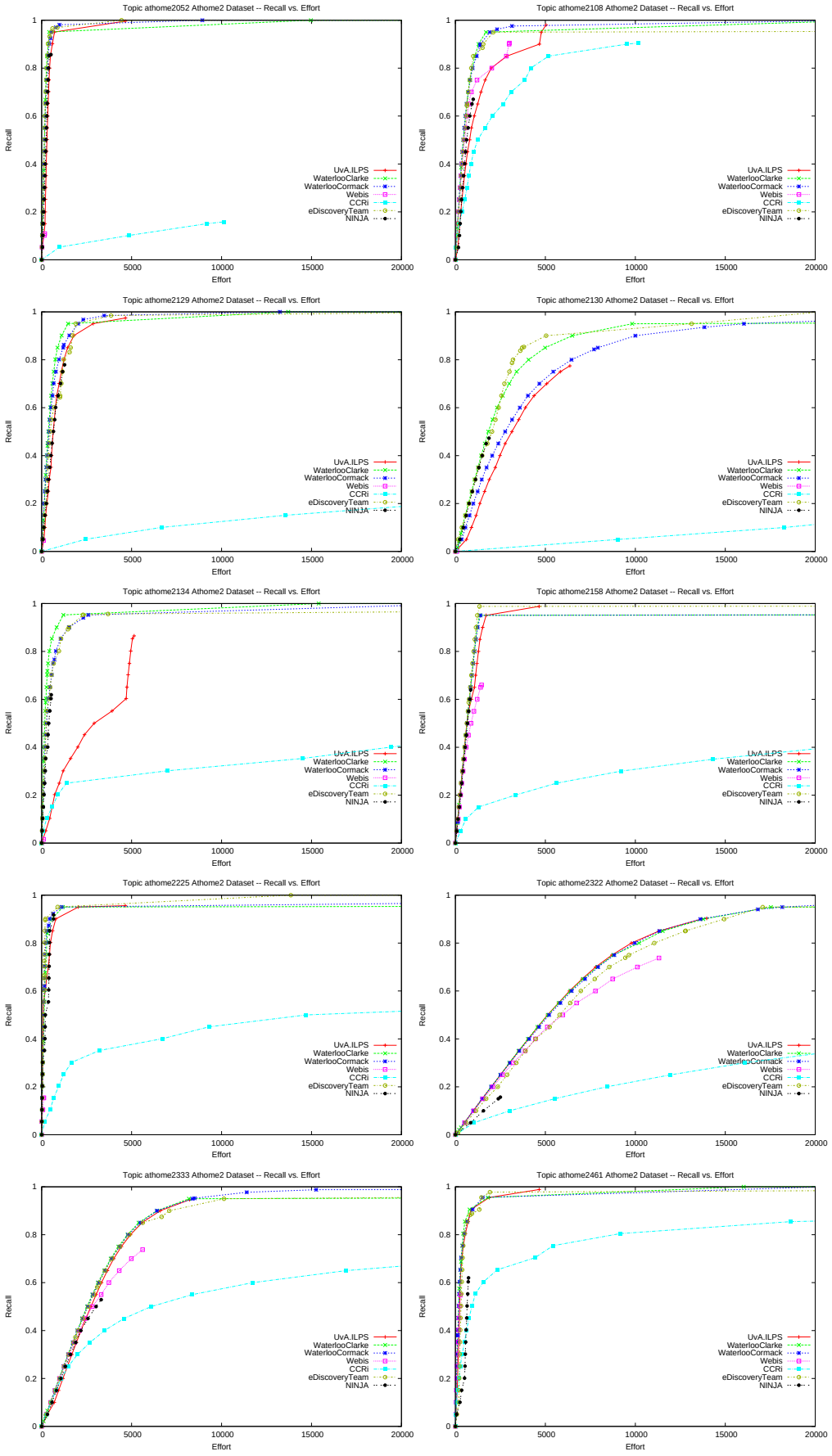

Figure 7: Per-Topic Gain Curves for the Athome2 Test Collection.

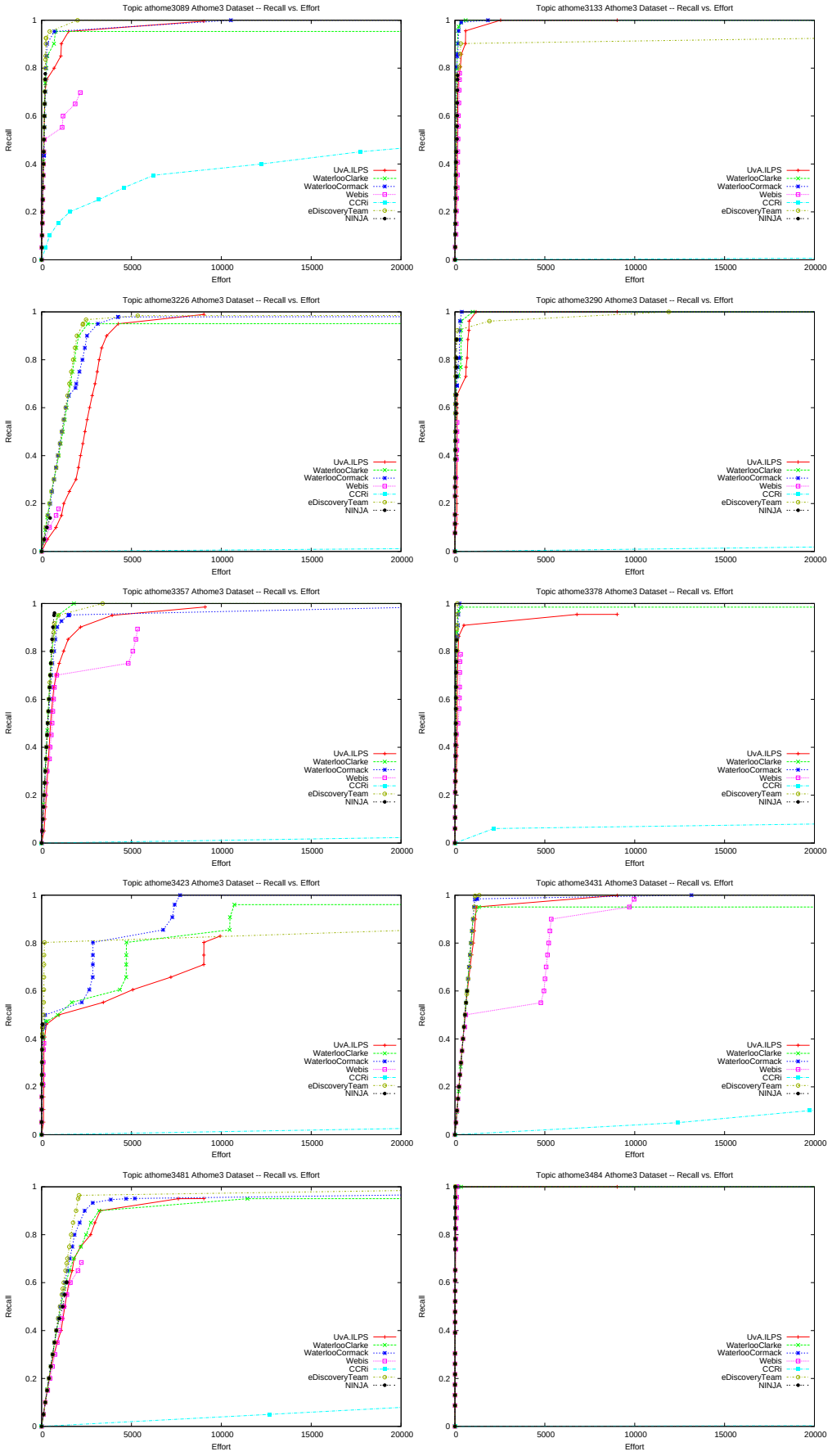

Figure 8: Per-Topic Gain Curves for the Athome3 iTest Collection.

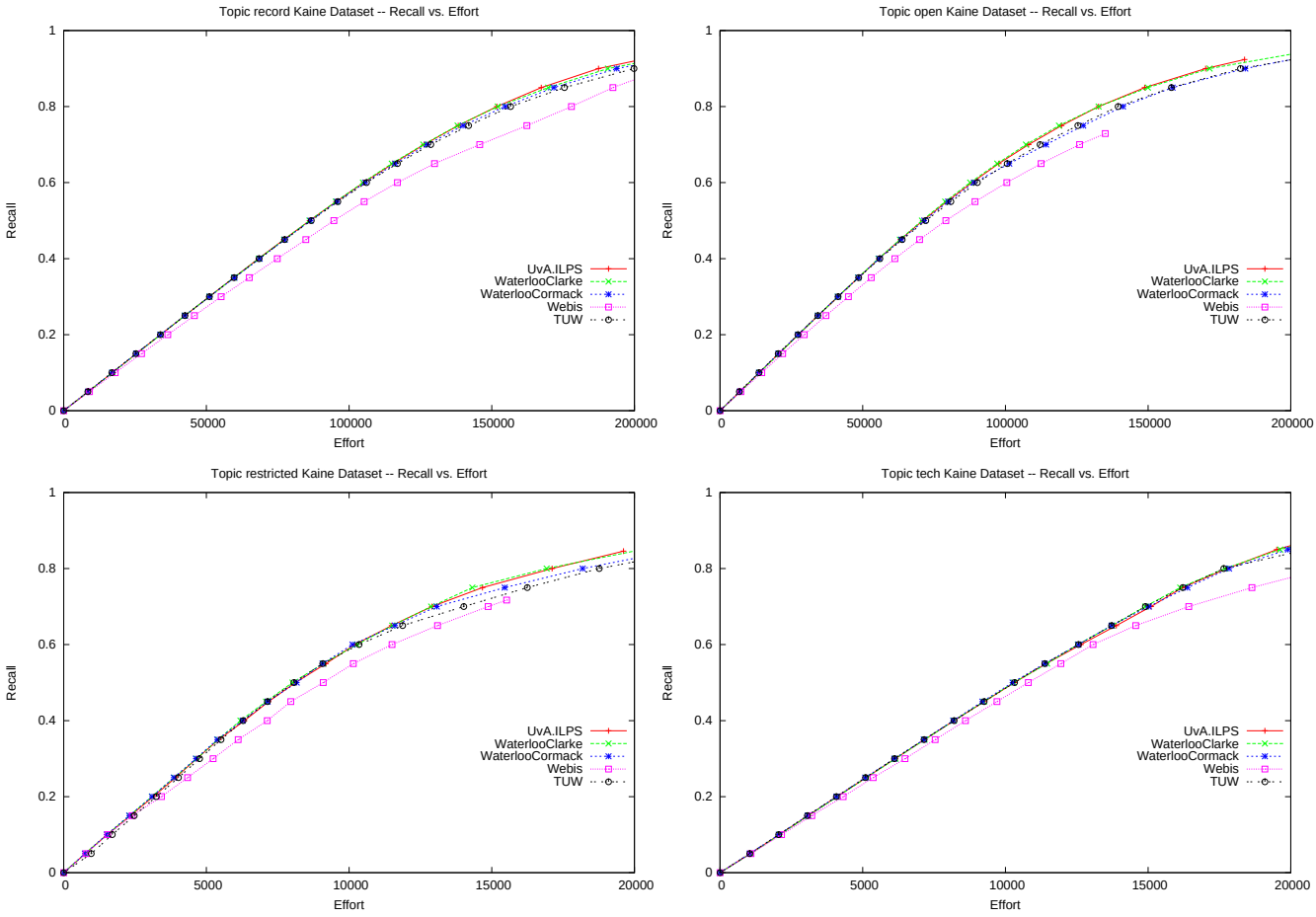

Figure 9: Per-Topic Gain Curves for the Kaine Test Collection.

recall, with higher than 50% precision, on average. It is not obvious which of these results is preferable. For a topic with only 100 relevant documents, it may be considered unreasonable to require the review of 10,000 documents; the average gain curve gives little insight in this regard. Individual per-topic gain curves (Figure 6) offer per-topic insight, but are challenging to generalize. The recall-precision curve indicates that it is possible to achieve 90% recall, with 50% precision or better; but if there are 10,000 documents to be found, is 50% good enough? Moreover, is a system that requires the review of 1,000 documents to find 100 (*i.e.*, 10% precision) inferior to a system that requires the review of 5,000 documents to find 1,000 (*i.e.*, 20% precision)? We suggest that a reasonable measure may lie somewhere in between these two extremes: The effectiveness of a system depends both on the absolute effort required (as shown by effort in the gain curve) and the effort required relative to the number of relevant documents found (as shown by precision in the recall-precision curve). Ideally, an effective system would score well on both curves.

The average gain and recall-precision curves for the remaining Athome test collections – athome2 and athome3 – are shown in Figures 2 and 3; per-topic gain curves are shown in Figures 7 and 8. Average curves for the Sandbox test collections – Kaine and MIMIC II – are shown in Figures 4 and 5. Per-topic gain curves for the Kaine collection are shown in Figure 9; per-topic curves for MIMIC-II are omitted for brevity. From these curves it can be seen that a number of systems – including the Baseline Model Implementation (denoted as "WaterlooCormack" in the curves) – achieve similar effectiveness, but no single system dominates over all topics, or over all collections. Further study is necessary to determine whether the observed differences among the top-performing systems on particular topics and test collections represent real, reproducible differences.

4.2 *Recall @ aR+b*

Tables 2 through 6 show, for each test collection, the new measure *Recall @ aR+b*. This measure quantifies the tradeoff between achieving high recall with effort proportionate to the number of relevant documents, and achieving high recall with reasonable overall effort.

4.3 When to Stop?

Participants were afforded the opportunity to "call their shot," indicating the point at which they would have recommended terminating the search because the additional effort to identify more relevant documents would have been unreasonable or disproportionate. Some participants did "call their shot," while others simply terminated their retrieval effort. In both cases, we tabulated the recall that had been achieved at that point and the effort (in terms of the number of assessed documents) necessary to achieve it. Clearly, there is a tradeoff between recall and effort; we made no attempt to quantify what was a reasonable compromise, and instead, present the raw, per-topic precision and effort results in Figures 7 through 12.

5 Discussion

In 2015, the inaugural year of the Total Recall Track, it was necessary to develop new, completely labeled datasets, a new evaluation architecture, new evaluation measures, and a Baseline Model Implementation.

For the Athome task, the original plan was to use datasets labeled by NIST assessors on a five-point relevance scale, according to a two-level hierarchy of topics and subtopics. This labeling effort was not completed within the allotted timeframe and budget, so the Total Recall coordinators decided to conduct binary relevance assessments for 30 topics across three datasets. Candidate documents for assessment were selected using a combination of ad-hoc search and machine learning. With few exceptions (*e.g.*, documents containing "schiavo" or "lethal injection"), every document labeled relevant was assessed by one of the Track coordinators. The selection and assessment of documents continued until the coordinators believed that substantially all relevant documents had been found; documents that were not selected for assessment were summarily labeled "not relevant." The Sandbox task used binary relevance assessments that did not rely in any way on the Track coordinators, or any search or machine-learning software. The Kaine collection was exhaustively labeled by the Senior State Records Archivist; the MIMIC II collection contained ICD-9 diagnostic codes which were used as relevance labels. The similarity between the Athome and Sandbox results suggests that any confounding due to the method of selecting and assessing the documents was minimal.

Run	Recall @								
	R	R+100	R+1000	2R	2R+100	2R+1000	4R	4R+100	4R+1000
catres	0.496	0.505	0.569	0.654	0.657	0.691	0.746	0.748	0.771
CCRi	0.233	0.236	0.260	0.309	0.310	0.320	0.364	0.364	0.372
eDiscoveryTeam	**0.783**	**0.792**	0.819	0.842	0.843	0.853	0.862	0.862	0.873
NINJA	0.577	0.600	0.620	0.619	0.619	0.620	0.620	0.620	0.620
TUW-1NB	0.710	0.740	0.815	0.841	0.843	0.861	0.870	0.871	0.955
TUW-1SB	0.714	0.737	0.814	0.838	0.842	0.859	0.869	0.873	0.966
TUW-1ST	0.715	0.735	0.898	0.886	0.905	0.952	0.966	0.966	0.972
TUW-6NB	0.717	0.741	0.817	0.841	0.844	0.860	0.871	0.872	0.961
TUW-6SB	0.630	0.654	0.734	0.753	0.762	0.778	0.789	0.791	0.874
TUW-6ST	0.710	0.727	0.892	0.866	0.888	0.950	0.964	0.965	0.972
UvA.ILPS-baseline	0.721	0.740	0.837	0.846	0.850	0.861	0.858	0.859	0.863
UvA.ILPS-baseline2	0.452	0.467	0.595	0.669	0.678	0.722	0.727	0.727	0.735
WaterlooClarke-UWPAH1	0.762	0.784	0.850	0.876	0.879	0.887	0.892	0.897	0.988
WaterlooClarke-UWPAH2	0.761	0.781	0.848	0.868	0.874	0.900	0.910	0.922	0.987
WaterlooCormack-stop2399	0.716	0.739	0.901	**0.903**	0.916	0.955	**0.968**	**0.968**	0.973
WaterlooCormack-Knee100	0.707	0.736	**0.904**	0.901	**0.917**	**0.956**	**0.968**	0.970	**0.974**
WaterlooCormack-Knee1000	0.715	0.738	0.900	**0.903**	**0.917**	0.953	0.967	**0.968**	0.973
webis-baseline	0.628	0.635	0.659	0.661	0.661	0.661	0.661	0.661	0.661
webis-keyphrase	0.560	0.568	0.588	0.590	0.590	0.590	0.590	0.590	0.590
WHU_IRGroup	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.038

Table 2: Recall @ aR+b for the athome1 Test Collection.

Run	Recall @								
	R	R+100	R+1000	2R	2R+100	2R+1000	4R	4R+100	4R+1000
CCRi	0.151	0.167	0.261	0.216	0.230	0.300	0.302	0.308	0.349
eDiscoveryTeam	0.647	0.732	0.893	0.859	0.891	0.951	0.942	0.948	0.969
NINJA	0.414	0.472	0.626	0.535	0.576	0.626	0.626	0.626	0.626
UvA.ILPS-baseline	0.501	0.588	0.799	0.735	0.765	0.854	0.837	0.844	0.873
UvA.ILPS-baseline2	0.260	0.279	0.468	0.429	0.450	0.617	0.615	0.630	0.733
WaterlooClarke-UWPAH1	**0.700**	**0.773**	**0.915**	**0.896**	**0.915**	**0.959**	**0.960**	**0.964**	**0.978**
WaterlooCormack-Knee100	0.670	0.735	0.883	0.856	0.877	0.939	0.941	0.947	0.971
WaterlooCormack-Knee1000	0.675	0.740	0.884	0.854	0.875	0.940	0.943	0.947	0.971
WaterlooCormack-stop2399	0.680	0.744	0.884	0.858	0.875	0.939	0.941	0.945	0.971
webis-baseline	0.328	0.356	0.377	0.377	0.377	0.381	0.382	0.383	0.391
webis-keyphrase	0.323	0.332	0.353	0.352	0.353	0.358	0.358	0.359	0.364

Table 3: Recall @ aR+b for the athome2 Test Collection.

Run	Recall @								
	R	R+100	R+1000	2R	2R+100	2R+1000	4R	4R+100	4R+1000
CCRi	0.009	0.012	0.023	0.016	0.017	0.026	0.024	0.024	0.031
eDiscoveryTeam	**0.823**	**0.923**	**0.962**	**0.934**	**0.957**	**0.963**	**0.958**	**0.963**	**0.966**
NINJA	0.602	0.704	0.704	0.670	0.704	0.704	0.704	0.704	0.704
UvA.ILPS-baseline	0.469	0.720	0.880	0.721	0.833	0.919	0.802	0.872	0.926
UvA.ILPS-baseline2	0.270	0.324	0.522	0.394	0.429	0.640	0.550	0.567	0.714
WaterlooClarke-UWPAH1	0.736	0.851	0.927	0.873	0.892	0.932	0.902	0.902	0.934
WaterlooCormack-Knee100	0.776	0.843	0.935	0.877	0.904	0.942	0.914	0.922	0.948
WaterlooCormack-Knee1000	0.793	0.857	0.936	0.889	0.909	0.942	0.914	0.922	0.948
WaterlooCormack-stop2399	0.786	0.859	0.935	0.886	0.913	0.943	0.914	0.922	0.948
webis-baseline	0.432	0.558	0.619	0.513	0.587	0.620	0.571	0.613	0.670
webis-keyphrase	0.412	0.444	0.456	0.435	0.452	0.458	0.457	0.459	0.513

Table 4: Recall @ aR+b for the athome3 test collection.

Run	Recall @								
	R	R+100	R+1000	2R	2R+100	2R+1000	4R	4R+100	4R+1000
TUW-1NB	0.787	0.788	0.799	0.962	0.963	0.965	0.990	0.990	0.990
TUW-6SB	0.798	0.799	0.808	0.967	0.967	0.968	0.991	0.991	0.991
TUW-6ST	0.803	0.805	0.816	0.969	0.969	0.970	0.992	0.992	0.992
UvA.ILPS-baseline	**0.812**	**0.813**	**0.824**	0.921	0.921	0.921	0.921	0.921	0.921
WaterlooClarke-UWPAH	**0.812**	**0.813**	**0.824**	**0.972**	**0.972**	**0.973**	**0.993**	**0.993**	**0.993**
WaterlooCormack-stop2399	0.795	0.797	0.808	0.968	0.969	0.969	0.991	0.992	0.992
Webis-baseline	0.736	0.738	0.748	0.781	0.781	0.781	0.781	0.781	0.781
Webis-keyphrase	0.556	0.558	0.568	0.603	0.603	0.603	0.603	0.603	0.603

Table 5: Recall @ aR+b for the Kaine Test Collection.

Run	Recall @								
	R	R+100	R+1000	2R	2R+100	2R+1000	4R	4R+100	4R+1000
TUW-1NB	0.688	0.702	0.748	0.894	0.896	0.912	0.973	0.974	0.978
TUW-1SB	0.691	0.704	0.749	0.894	0.896	0.911	0.973	0.973	**0.977**
TUW-1ST	0.695	0.706	0.752	**0.896**	0.899	0.913	**0.974**	**0.974**	**0.977**
TUW-6NB	0.681	0.695	0.741	0.889	0.891	0.907	0.971	0.971	0.975
TUW-6SB	0.685	0.697	0.743	0.890	0.891	0.906	0.971	0.971	0.975
TUW-6ST	0.689	0.700	0.744	0.891	0.893	0.907	0.972	0.973	0.976
UvA.ILPS-baseline	0.453	0.478	0.489	0.490	0.490	0.490	0.490	0.490	0.490
UvA.ILPS-baseline2	0.516	0.518	0.570	0.554	0.565	0.587	0.587	0.587	0.587
WaterlooClarke-UWPAH1	0.684	0.695	0.742	0.889	0.892	0.908	0.970	0.971	0.975
WaterlooCormack-stop2399	**0.697**	**0.707**	**0.753**	**0.896**	**0.898**	**0.913**	0.973	**0.974**	**0.977**
webis-baseline	0.624	0.637	0.662	0.680	0.681	0.684	0.684	0.684	0.684
webis-keyphrase	0.513	0.526	0.539	0.559	0.559	0.559	0.559	0.559	0.559

Table 6: Recall @ aR+b for the MIMIC II Test Collection.

	Topic (R) – Athome1 Collection									
	100 (4542)	101 (5836)	102 (1624)	103 (5725)	104 (227)	105 (3635)	106 (17135)	107 (2375)	108 (2375)	109 (506)
eDiscoveryTeam	0.5374 2536 +651*	0.7493 4771 +6841*	0.5800 1071 +1493*	0.8351 4817 +7203*	0.6916 199 +1091*	0.9224 3418 +674*	0.9847 17516 +2226*	0.8236 2259 +1164*	0.3091 746 +696*	0.9605 510 +753*
NINJA	0.1497 1437	0.8045 6025	0.6108 1895	0.9310 6998	0.7005 327	0.7431 3359	0.9048 17743	0.3676 1056	0.0720 171	0.9170 1396
UvA.ILPS-baseline	0.9036 6618	0.9896 7538	0.9206 3679	0.9857 6823	0.8546 3551	0.7508 4876	0.9879 18180	0.9689 3582	0.9390 4543	0.9723 2901
UvA.ILPS-baseline2	0.7431 10274	0.9467 9719	0.6552 3392	0.9275 8024	0.1674 2908	0.6526 7330	0.9720 22204	0.7937 5510	0.6830 6859	0.9091 2914
WaterlooClarke-UWPAH1	0.0638 307	0.0526 315	0.1921 315	0.0552 315	0.8194 315	0.0869 315	0.0184 315	0.1301 315	0.1221 308	0.0059 308
WaterlooClarke-UWPAH2	0.0553 314	0.0526 315	0.1921 315	0.0552 315	0.7885 315	0.0869 307	0.0180 315	0.1309 308	0.1217 308	0.0059 303
WaterlooCormack-Knee100	0.9813 15256	0.9988 11396	0.9581 5729	0.9974 10336	0.9163 1232	0.8856 15257	0.9977 27254	0.9878 6325	0.9764 8498	0.0059 130
WaterlooCormack-Knee1000	0.9830 16810	0.9991 11396	0.9483 5188	0.9976 10336	0.9119 1105	0.8809 15257	0.9978 27254	0.9861 5729	0.9764 8498	0.9802 3144
WaterlooCormack-stop2399	0.9194 7702	0.9979 10336	0.9360 4251	0.9963 9373	0.9692 2841	0.7590 5729	0.9973 24749	0.9857 5729	0.9389 5188	0.9802 3144
Webis-baseline	0.7191 5213	0.9639 7578	0.7285 1837	0.7696 4967	0.2819 231	0.6856 4044	0.9669 17786	0.8863 2938	0.6097 2554	0.0000 110
Webis-keyphrase	0.0007 105	0.9627 7351	0.7174 1817	0.7686 4858	0.3172 237	0.6875 4039	0.9665 17767	0.8842 2925	0.5912 2438	0.0020 110

Table 7: Set-Based Results (Recall Over Effort) For the athome1 Test Collection. The top number in each cell indicates the recall achieved when the submission indicated by "calling its shot" that a "reasonable" result had been achieved, or when the submission terminated its run. The second number indicates the number of documents submitted to the automated assessment server at this point. (*) indicates the number of additional documents reviewed by a manual review team, and not necessarily submitted to the automated assessment server.

	Topic (R) – Athome2 Collection									
	2052 (265)	2108 (661)	2129 (589)	2130 (2299)	2134 (252)	2158 (1256)	2225 (182)	2322 (9517)	2333 (4805)	2461 (179)
eDiscoveryTeam	0.9698 871 +2325*	0.8850 1505 +2101*	0.9847 3854 +94*	0.8534 3802 +285*	0.9563 3689 +19*	0.9881 1339 +1335*	0.8956 201 +205*	0.8512 12799 +195*	0.8745 6670 +228*	0.9777 1918 +32*
NINJA	0.8566 516	0.6702 990	0.7793 1276	0.4728 1867	0.6191 534	0.6393 850	0.9176 665	0.1567 2497	0.5290 3310	0.6201 728
UvA.ILPS-baseline	0.9962 4651	0.9803 5038	0.9745 4657	0.7747 6363	0.8651 5141	0.9881 4651	0.9561 4652	0.9032 13969	0.9463 8224	0.9888 4653
UvA.ILPS-baseline2	0.9660 4672	0.9183 4746	0.8693 4736	0.7234 8472	0.6865 4722	0.6226 5704	0.8681 4715	0.8880 25613	0.7832 11549	0.8883 4692
WaterlooClarke-UWPAH1	0.8491 307	0.3858 315	0.4414 324	0.0761 308	0.7183 315	0.2508 315	0.8407 308	0.0302 315	0.0651 313	0.6872 320
WaterlooCormack-Knee100	0.9811 989	0.9758 3144	0.9847 3478	0.9622 20402	0.9405 2316	0.0892 130	0.9231 630	0.9841 30010	0.9881 15257	0.3799 130
WaterlooCormack-Knee1000	0.9849 1104	0.9788 3144	0.9796 3144	0.9604 20402	0.9405 2316	0.9881 2841	0.9560 1105	0.9862 33042	0.9875 15257	0.9385 1372
WaterlooCormack-stop2399	0.9925 2840	0.9818 3478	0.9830 3144	0.7086 4697	0.9524 2841	0.9881 4251	0.9835 2841	0.8800 12562	0.9534 8498	0.9888 2841
Webis-baseline	0.1094 173	0.9047 3002	0.0459 110	0.0000 80	0.0159 111	0.6600 1457	0.1539 123	0.7376 11313	0.7378 5611	0.5475 271
Webis-keyphrase	0.0868 172	0.8684 3096	0.0492 110	0.0000 111	0.0159 111	0.6147 1229	0.1044 111	0.7354 11152	0.7590 5915	0.4078 278

Table 8: Set-Based Results (Recall Over Effort) For the athome2 Test Collection. The top number in each cell indicates recall achieved when the submission indicated by "calling its shot" that a "reasonable" result had been achieved, or when the submission terminated. The second number indicates the number of documents submitted to the automated assessment server at this point. (*) indicates the number of additional documents reviewed by a manual review team, and not necessarily submitted to the automated assessment server.

	Topic (R) – Athome3 Collection									
	3089 (255)	3133 (113)	3226 (2094)	3290 (26)	3357 (629)	3378 (66)	3423 (76)	3431 (1111)	3481 (2036)	3484 (23)
eDiscoveryTeam	0.9255 250 +834*	0.7699 97 +49*	0.9842 5347 +18*	0.8846 95 +306*	0.9173 701 +920*	0.8939 106 +200*	0.4474 40 +92*	0.9973 1121 +272*	0.9646 2077 +367*	1.0000 23 +73*
NINJA	0.7765 201	0.7699 143	0.1395 474	0.8846 92	0.9603 714	0.8485 99	0.4605 45	0.5995 673	0.6027 1392	1.0000 23
UvA.ILPS-baseline	0.9961 9024	1.0000 9024	0.9895 9031	1.0000 9024	0.9857 9100	0.9546 9024	0.8290 9936	0.9991 9024	0.9509 9026	1.0000 9024
UvA.ILPS-baseline2	0.9765 9057	0.9735 9097	0.9231 9032	0.6923 9106	0.9142 9103	0.9546 9091	0.4474 9116	0.9937 9046	0.8144 9325	1.0000 9084
WaterlooClarke-UWPAH1	0.8510 314	0.9912 330	0.1423 315	0.7692 309	0.4706 315	0.9848 332	0.4737 309	0.2844 315	0.1552 315	1.0000 330
WaterlooCormack-Knee100	0.4353 129	0.9912 343	0.9790 4251	0.6923 111	0.9523 1526	0.8636 130	0.4605 151	0.9847 1232	0.9514 5188	1.0000 111
WaterlooCormack-Knee1000	0.9961 1104	0.9912 1105	0.9733 3846	1.0000 1105	0.9634 1883	1.0000 1105	0.5263 1105	0.9892 1232	0.9504 4251	1.0000 1105
WaterlooCormack-stop2399	0.9961 2840	1.0000 2566	0.9852 5188	1.0000 2566	0.9841 3144	1.0000 2566	0.6184 2566	0.9991 3846	0.9504 5188	1.0000 2566
Webis-baseline	0.6980 2156	0.7788 261	0.1777 942	0.5385 119	0.8935 5338	0.7879 314	0.3816 130	0.9829 9971	0.6842 2215	1.0000 110
Webis-keyphrase	0.8275 1954	0.0620 106	0.1380 1066	0.2692 111	0.8347 5212	0.5000 296	0.1974 111	0.9748 10175	0.6680 6235	0.8696 110

Table 9: Set-based Results (recall over effort) for the athome3 test collection. The top number in each cell indicates the recall achieved when the submission indicated by "calling its shot" that a "reasonable" result had been achieved, or when the submission terminated its run. The second number indicates the number of documents submitted to the automated assessment server at this point. (*) indicates the number of additional documents reviewed by a manual review team, and not necessarily submitted to the automated assessment server.

	Topic (R) – Kaine Collection			
	Open (131698)	Restricted (14341)	Record (166118)	VA Tech (20083)
UvA.ILPS-baseline	0.9237 183797	0.8458 19616	0.9599 228712	0.9558 27456
WaterlooClarke-UWPAH1	0.0020 304	0.0191 304	0.0013 304	0.0126 304
WaterlooCormack-stop2399	0.6605 104421	0.7440 15259	0.8801 185276	0.9507 27256
Webis-baseline	0.7291 134983	0.7171 15523	0.8858 205524	0.7934 20854
Webis-keyphrase	0.0001 73	0.7214 15806	0.9007 210913	0.7916 20785

Table 10: Set-Based Results (Recall Over Effort) For the Kaine Test Collection. The top number in each cell indicates the recall achieved when the submission indicated by "calling its shot" that a "reasonable" result had been achieved, or when the submission terminated its run. The second number indicates the number of documents submitted to the automated assessment server at this point.

	Topic (R) – MIMIC II Collection [Part I]									
	C01 (5881)	C02 (3867)	C03 (15101)	C04 (7826)	C05 (6123)	C06 (5081)	C07 (19182)	C08 (11256)	C09 (8706)	C10 (8741)
UvA.ILPS-baseline	0.6910 5960	0.7210 3528	0.0031 319	0.6360 7520	0.5692 5061	0.4718 4146	0.9585 20584	0.0022 328	0.0030 317	0.6608 7520
UvA.ILPS-baseline2	0.5581 5702	0.6289 4332	0.9364 20458	0.6254 8904	0.4023 4569	0.3340 3815	0.9766 22203	0.7831 12381	0.6230 8459	0.6984 9989
WaterlooClarke-UWPAH1	0.0452 301	0.0706 302	0.0195 302	0.0235 302	0.0483 302	0.0382 302	0.0157 302	0.0263 302	0.0326 302	0.0257 302
WaterlooCormack-stop2399	0.8109 8497	0.8836 6981	0.9548 20402	0.7313 9373	0.7013 7703	0.6215 6325	0.9949 27254	0.9087 15257	0.8256 11396	0.8396 11396
Webis-baseline	0.6817 6630	0.7277 3709	0.9533 21110	0.6606 9082	0.5078 5098	0.4399 4360	0.9722 22035	0.8047 12790	0.6149 7580	0.7603 10727
Webis-keyphrase	0.6667 6387	0.7199 3609	0.9464 20725	0.0017 47	0.5166 5283	0.0010 48	0.9691 21848	0.8189 13195	0.6347 7973	0.0008 48

Table 11: Set-Based Results (Recall Over Effort) for the MIMIC II Test Collection, Part I. The top number in each cell indicates the recall achieved when the submission indicated by "calling its shot" that a "reasonable" result had been achieved, or when the submission terminated its run. The second number indicates the number of documents submitted to the automated assessment server at this point.

	Topic (R) – MIMIC II Collection [Part II]								
	C11 (180)	C12 (2579)	C13 (3465)	C14 (2143)	C15 (5143)	C16 (8047)	C17 (11117)	C18 (16827)	C19 (6828)
UvA.ILPS-baseline	0.9500 359	0.0016 315	0.0026 315	0.5684 2019	0.9117 5481	0.4977 6710	0.6980 10436	0.4561 7966	0.5053 4319
UvA.ILPS-baseline2	0.6778 683	0.2404 1650	0.0205 388	0.3985 1569	0.8707 5249	0.3759 5699	0.7603 13813	0.7688 18688	0.4782 4979
WaterlooClarke-UWPAH1	0.9444 302	0.0733 302	0.0390 302	0.0854 302	0.0428 302	0.0229 302	0.0247 302	0.0180 302	0.0395 302
WaterlooCormack-stop2399	0.9889 2841	0.6758 4697	0.5957 5188	0.7485 4697	0.9994 9373	0.5469 7703	0.8239 13845	0.8585 20402	0.6995 8498
Webis-baseline	0.9556 797	0.4362 2594	0.3417 2761	0.6122 2861	0.8991 5264	0.5144 7753	0.7857 13487	0.8271 21000	0.5082 4449
Webis-keyphrase	0.8333 301	0.4017 2390	0.0023 47	0.5469 2484	0.9078 5380	0.5211 7880	0.8093 14073	0.8126 20449	0.5069 4416

Table 12: Set-Based Results (Recall Over Effort) for the MIMIC II Test Collection, Part II. The top number in each cell indicates the recall achieved when the submission indicated by "calling its shot" that a "reasonable" result had been achieved, or when the submission terminated its run. The second number indicates the number of documents submitted to the automated assessment server at this point.

6 Conclusions

The 2015 Total Recall Track successfully deployed a new evaluation architecture, using five new datasets. Three of the datasets will be publicly available, while two datasets must remain private. The results appear to be consistent across all datasets: A fully automated Baseline Model Implementation achieved high recall for all topics, with effort proportionate to the number of relevant documents. Several manual and automatic participant efforts achieved higher recall with less effort than the baseline on some topics, but none consistently improved on the baseline. Some teams appeared to be able to predict when a "reasonable" result had been achieved; however, further work is needed to derive appropriate measures to evaluate what is "reasonable."

Acknowledgement

We are grateful to the Library of Virginia for affording us access to the Kaine email collection for the Sandbox evaluation. We give particular thanks to SusanGray Paige, Roger Christman, Rebecca Morgan, and Kathy Jordan for their invaluable assistance with and unwavering support for this work.

References

[1] Wei Lu Chuan Wu and Ruixue Wang. WHU at TREC Total Recall Track 2015. In *Proc. TREC-2015*, 2015.

[2] Gordon V. Cormack and Maura R. Grossman. The Grossman-Cormack Glossary of Technology-Assisted Review. *Fed. Cts. L. Rev.*, 7(1), 2013.

[3] Gordon V Cormack and Maura R Grossman. Multi-faceted recall of continuous active learning for technology-assisted review. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 763–766. ACM, 2015.

[4] Gordon V. Cormack and Maura R. Grossman. Waterloo (Cormack) Participation in the TREC 2015 Total Recall Track. In *Proc. TREC-2015*, 2015.

[5] Gordon V Cormack and Mona Mojdeh. Machine Learning for Information Retrieval: TREC 2009 web, relevance feedback and legal tracks. In *Proc. TREC-2009*, 2009.

[6] Evangelos Kanoulas David van Dijk, Zhaochun Ren and Maarten de Rijke. The University of Amsterdam (ILPS) at TREC 2015 Total Recall Track. In *Proc. TREC-2015*, 2015.

[7] Vivek Dhand. Efficient semantic indexing via neural networks with dynamic supervised feedback. In *Proc. TREC-2015*, 2015.

[8] Maura R Grossman and Gordon V Cormack. Comments on "The implications of Rule 26(g) on the use of technology-assisted review". *Fed. Cts. L. Rev.*, 8(1), 2014.

[9] Yipeng Wang Charles L.A. Clarke Haotian Zhang, Wu Lin and Mark D. Smucker. WaterlooClarke: TREC 2015 Total Recall Track. In *Proc. TREC-2015*, 2015.

[10] Bruce Hedin, Stephen Tomlinson, Jason R. Baron, and Douglas W. Oard. Overview of the TREC 2009 Legal Track. In *Proc. TREC-2009*, 2009.

[11] Julian PT Higgins, Sally Green, et al. *Cochrane handbook for systematic reviews of interventions*, volume 5. Wiley Online Library, 2008.

[12] Bayu Hardi Jeremy Pickens, Tom Gricks and Mark Noel. A Constrained Approach to Manual Total Recall. In *Proc. TREC-2015*, 2015.

[13] Ralph Losey, Jim Sullivan, and Tony Reichenberger. e-Discovery Team at TREC 2015 Total Recall Track. In *Proc. TREC-2015*, 2015.

[14] Mihai Lupu. TUW at the first Total Recall Track. In *Proc. TREC-2015*, 2015.

[15] Magdalena Keil Olaoluwa Anifowose Amir Othman Matthias Hagen, Steve Göring and Benno Stein. Webis at TREC 2015: Tasks and Total Recall Tracks. In *Proc. TREC-2015*, 2015.

[16] Adam Roegiest and Gordon V Cormack. Total Recall Track Tools Architecture Overview. In *Proc. TREC-2015*, 2015.

[17] Hui Yang, John Frank, and Ian Soboroff. Trec 2015 dynamic domain track overview. In *Proc. TREC-2015*, 2015.