TREC 2016 Total Recall Track Overview

Maura R. Grossman, Gordon V. Cormack, and Adam Roegiest
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 [3], systematic review in evidence-based medicine [6], and the creation of fully labeled test collections for information retrieval ("IR") evaluation [5]. 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 typically 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 Total Recall Track debuted at TREC 2015 [7]. The TREC 2016 track was operationally identical to the TREC 2015 Track, differing only in the following respects:

- This year, participants were *required* to "call their shot" to indicate when they believed that as many of the relevant documents as reasonably possible had been identified with proportionate effort;
- The TREC 2015 At-Home collections (as well as the TREC 2015 Practice collections) were available for testing and development;
- 34 new topics were developed for the TREC 2015 Jeb Bush dataset for the 2016 At-Home task;
- Six topics and a new Rod Blagojevich/Pat Quinn dataset, as well as four topics and a new collection of Twitter tweets [1] were introduced for the 2016 Sandbox task.

Testing and development, as well as 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 June, July, and August, 2016 (and continue to be available for unofficial runs).

Sandbox runs were conducted in September 2016, entirely on a Web-isolated platform hosting the data collections. 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.[1] BMI was run on all of the 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 2015 Notebook Draft paper titled *Total Recall Track Tools Architecture Overview.*[2]

The TREC 2016 Total Recall Track attracted five participants, including two industrial groups that submitted manual At-Home runs, one academic group that submitted only automatic At-Home runs, and two academic groups that submitted both automatic At-Home and Sandbox runs.

The 2016 At-Home collection, *athome4*, consisted of 34 topics and the same dataset of 290,000 Jeb Bush emails that was used in the TREC 2015 *athome1* collection (*see* [7]). The topics were composed by the Track coordinators, and relevance assessments were rendered by NIST assessors, with guidance and quality assurance provided by the Track coordinators. Documents were selected for assessment using a combination of interactive search and judging **[4]** and machine-learning techniques similar to those used for the TREC 2002 Filtering Track [8].

The Sandbox collections consisted of two datasets and 10 topics. The *Illinois* collection consisted of 2.1M email messages from the administration of former Illinois Governors Blagojevich and Quinn, which were provided by the Illinois State Archive. In collaboration with the University of Illinois, six topics were identified and assessed by archive and university personnel. Documents were selected for review using a combination of interactive search and judging and machine learning as described above. The *Twitter* collection consisted of 800,000 tweets, with crowdsourced relevance assessments, for four topics [1].

The principal tool for comparing runs was a *gain curve*. A gain curve 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, especially at high recall levels. The traditional *recall-precision* curve conveys similar information, plotting *precision* (*i.e.*, the fraction 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). While gain curves and recall-precision curves convey similar information, they 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. In general, Total-Recall applications tolerate a fair amount of fixed overhead in exchange for high recall; this tradeoff is more readily apparent in a gain curve.

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. This year, participants were required 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, but they had 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, in 2015, the Total Recall Track coordinators introduced a new parametric measure: *recall* @ $aR + b$, for various values of a and b. *Recall* @ $aR + b$ was defined to be the recall achieved when $aR + b$ documents had 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 and 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 identified; 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 $b \in \{0, 100, 1000\}$.

To address limitations of recall measures based on binary relevance, assessors for the athome4 and Illinois collections were asked to identify, among those documents that they assessed to be relevant, those they deemed to

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

[2] cormack.uwaterloo.ca/total-recall/overview/totalrecallarch.pdf.

be "important" (*i.e.*, "key"). An alternative version of recall was computed with respect to this set of documents; the corresponding gain curves and $aR + b$ results are shown for this alternative version of recall, as well as for traditional recall.

To address the question of how well the systems identified different *facets* of relevance (*see* [2]), the athome4 assessors were also asked to group relevant documents into folders corresponding to meaningful subcategories they identified. In addition to overall recall for each topic, recall for each facet or subcategory was computed separately, so as to assess the diversity of coverage of each topic among the submitted results.

Finally, to address the issues of assessor (dis)agreement and the completeness of the documents presented to the assessors, a stratified sample of 50 documents for each of the athome4 topics was independently assessed by three secondary NIST assessors. Alternative versions of recall were computed for each of the secondary assessors, as well as for the "majority-of-three" assessments for the documents in the sample.

In calculating effort and precision, the measures described 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 to conduct their own assessments. Participants were required to track and report the number of documents that they reviewed, and were required to submit the documents they reviewed contemporaneously to the server. However, *not all participants followed the instructions to submit all documents they reviewed to the server*. Therefore, the reader should consult the participants' descriptions of their methods; these descriptions should be considered when comparing manual runs to one another, or to automatic runs. It is not obvious whether (or how) this additional – and sometimes unaccounted for – effort is (or should be) reflected in the gain curves, and *recall* @ $aR + b$ measures; therefore, the coordinators have chosen not to try.

Results for the TREC 2016 Total Recall Track are consistent with those of the 2015 Track, showing 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 at TREC 2015 or TREC 2016 – whether manual or automatic – consistently achieved higher recall at lower effort than BMI.

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 the 2016 Practice runs, all of the Practice and At-Home collections from the TREC 2015 Total Recall Track were made available to participants.

For the TREC 2016 At-Home runs, four variants of the athome4 collection were available:

- *athome4*: The (redacted) Jeb Bush Emails,[3] consisting of 290,099 emails from Jeb Bush's eight-year tenure as Governor of Florida. We used 34 issues associated with his governorship as topics for the *athome4* test collection, shown in Table 1. For each topic, the server supplied a short topic title, consisting of one-to-three words.

- *athome4desc*: The same dataset and topics as *athome4*, but the server supplied the title as well as a short description of the topic, rather than just the title alone.

- *athome4subset*: A subset of athome4 with 12 randomly selected topics. This collection was provided for participants who lacked the resources to complete all 34 topics. Because all participants submitted results for either athome4 or athome4desc, athome4subset results are not reported here.

- *athome4descsubset*: A subset of athome4desc with 12 randomly selected topics, with both the title and a short description of the topic. This collection was likewise provided for participants who lacked the resources to complete all 34 topics. Because all participants submitted results for either athome4 or athome4desc, athome4descsubset results are not reported here.

For the Sandbox runs, we used two new datasets:

- *Illinois*: 2.1M email messages from administrations of former Illinois Governors Rod Blagojevich and Pat Quinn, supplied by the Illinois State Archive, in cooperation with the University of Illinois. Six topics supplied by the Illinois State Archive were assessed by archive and university personnel.

[3] `https://web.archive.org/web/20160221072908/http://jebemails.com/home`

Topic	Title	Description
401	Olympics	Bid to host the Olympic games in Florida.
402	Space	The space industry, space program, space travel, or space science, public and private, in Florida.
403*	Bottled Water	Extraction of water for bottling by commercial enterprises.
404	Eminent domain	Legality or morality of expropriating land for commercial development.
405	Newt Gingrich	Speaker Newt Gingrich or any entities or personnel associated with Newt Gingrich.
406	Felon disenfranchisement	Right of felons to vote, including but not restricted to voter purges and reinstatement of voter rights. Individual clemency cases are not relevant.
407	Faith-based initiatives	Grants or other initiatives to offload social services to so-called faith-based agencies. Services include but are not limited to education, prisons, and emergency relief.
408*	Invasive species	The problem of invasive species – non-native plants or animals that threaten the ecosystem.
409*	Climate change	Climate change, global warming, or carbon emissions.
410	Condos	Rules and organizations governing condominium associations and conflicts between owners and managers. Relevant documents include those concerning the establishment of the office of ombudsman, and issues relating to hiring and firing the ombudsman.
411	Stand your ground	Use of deadly force to protect one's self or one's property.
412	2000 Recount	Contested result of the 2000 presidential election.
413	James V. Crosby	James V. Crosby, including but not limited to his relationship with Gov. Bush before being appointed Secretary of Corrections, his role as Secretary, his firing, and criminal allegations against Mr. Crosby.
414*	Medicaid reform	Efforts to substantially reform Medicaid.
415	George W. Bush	Documents referring to George W. Bush, whether explicitly or by his relationship to Gov. bush.
416*	Marketing	Advertising or marketing efforts undertaken by the Governor's office or institutions of the State of Florida.
417	Movie Gallery	Investments by Florida in Movie Gallery.
418	War preparations	Preparations for the Iraq War undertaken before the March 20, 2003 invasion.
419	Lost foster child	Disappearance of Rilya Wilson and its aftermath.
420	Billboards	Rights and control of billboards. Distinct legislative efforts should be considered to be separate categories.
421	Traffic cameras	Use of unattended cameras to enforce traffic laws.
422*	Non-resident Aliens	Non-resident alien issue. Documents concerning the National Rifle Association are not relevant.
423*	National Rifle Association	The NRA, its members, and its influences.
424	Gulf drilling	Off-shore drilling for oil or gas. Water drilling is not relevant.
425*	Civil Rights Act	Civil Rights Act of 2003.
426	Jeffrey Goldhagen	Jeffrey Goldhagen's role in the administration, his firing, and reinstatement.
427	Slot Machines	Legality/licensing/definition of "slot machines."
428	New Stadiums	Construction of new sports stadiums or arenas.
429*	Cuban Child	Elian Gonzales and his status.
430*	Restraints and Helmets	Seat belt, child seat, and helmet mandates.
431	Agency Ratings	Credit ratings of Florida institutions, particularly those by Standard and Poor's, Fitch, and Moody's.
432	Gay Adoption	Gay adoption issue.
433*	Abstinence	Abstinence and abstinence-only programs to supplant birth control or sex education.
434*	Bacardi Trademark Lobbying	The Jeb Bush administration's involvement in a trademark dispute between Bacardi and the U.S. Patent and Trademark Office.

Table 1: Topics and Topic Descriptions for the Athome4 Collection. The 12 subset topics are marked with a (*).

- *Twitter*: 800,000 tweets with crowdsourced relevance assessments, for four topics, supplied by Twitter.

3 Participant Submissions

To assist participants in completing the At-Home and Sandbox tasks, as well as to provide a baseline for comparison, a Baseline Model Implementation ("BMI") was supplied to Track participants.[4] The only change from the 2015 version of BMI was the inclusion of a default rule to call your shot: A "reasonable" result was deemed to have been achieved when m relevant and n non-relevant documents had been reviewed, where $n > a \cdot m + b$, and $a = 0.5$ and $b = 1000$, were predetermined constants. 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.

Two commercial teams (eDiscoveryTeam and catres) used manual processes; three academic teams (IMS, SFSU, and UW) used fully automated processes. All teams did the full athome4 collection; only two teams (SFSU and UW) submitted Sandbox runs.

4 Results

4.1 At-Home Task

Gain curves for the At-Home task are shown in Figure 1; $aR + b$ and "call-your-shot" results are shown in Figure 2. The gain curves plot recall (averaged over all topics) as a function of the number of documents submitted. Several of the methods – all derivatives of BMI – yielded essentially the same curve, which is superior to all other submissions. The two manual efforts (eDiscoveryTeam and catres) fall somewhat below. The first nine columns of Figure 2 show the same information in tabular form: recall when $aR + b$ documents have been submitted, averaged over all topics. BMI and sfsu yield comparable results; sfsu may have a tiny edge. The last column shows recall achieved when the system "calls its shot." The BMI-derived runs achieve recall on the order of 0.95; the manual runs, on the order of 0.75.

4.2 Sandbox Task

Figures 3 and 4 show gain curves, $aR + b$, and call-your-shot results for the Illinois collection. Only uw and sfsu participated in the Sandbox task, both achieving results comparable to BMI. Figures 5 and 6 show results for the same systems on the Twitter collection; notably, uw.knee calls its shot at lower recall (0.801), compared to other collections.

4.3 Alternative Relevance I: "Important" or "Key" Documents

Figures 7 and 8 show the results when only "important" or "key" documents are considered relevant for the purpose of evaluating recall. The calculations of the number of documents submitted and R remain unchanged. Comparison with the results that consider all relevant documents (Figures 1 and 2) shows an insubstantial difference: recall for "important" or "key" documents appears to be slightly higher, particularly at lower levels of effort. Figures 9 and 10 show a similar effect for the Illinois Test Collection, as compared to figures 3 and 4. No "important" or "key" relevance assessments were available for the Twitter Test Collection.

4.4 Alternative Relevance II: Facet or Subtopic Recall

For the Jeb Bush Test Collection, assessors were asked to categorize relevant documents into subfolders of their own choosing, reflecting meaningful facets of relevance. A total of 348 folders were created (10.2 per topic, on average). Figures 11 and 12 show recall, macro-averaged over the 348 subtopics, as a function of effort. Comparison with recall averaged over the 34 topics as a whole (Figures 1 and 2) shows no substantial difference.

[4] http://cormack.uwaterloo.ca/trecvm/.

4.5 Alternative Relevance III: Secondary Assessor and Majority-of-Three Assessments

For each topic in the Jeb Bush Test Collection, 50 documents were chosen using non-uniform sampling. Three assessors independently assessed each of the 50 documents for relevance. Recall over the entire dataset was computed using the Horvitz-Thompson estimator, as well as for four alternative versions of relevance: separately for each of the three secondary assessors, as well as a majority-of-three assessment for the three secondary assessors. The results for all of these versions of relevance is shown in Figures 13 through 20. Recall with respect to the majority vote appears somewhat lower than with respect to the original NIST assessments, as shown by comparing Figures 13 and 14 to Figures 1 and 2. Results for recall with respect to each individual secondary assessor is substantially lower; the best results achieve recall on the order of 0.7.

5 Discussion

In 2015 and 2016, a number of participants derived their systems from BMI. Other participants used manual processes involving some combination of ad-hoc search, human document review, and commercial macnine-learning tools. None of the manual participants were able to show consistently superior results to the fully automated method, BMI. This year, SFSU appeared to have a small edge in performance on the Jeb Bush Test Collection, but that edge did not manifest itself on either the Illinois or Twitter Test Collections. It is worth noting that the SFSU submission was more than ten times slower than BMI, taking more than one week to run on the Illinois collection, whereas BMI only took several hours. Similar run-time disparities were noticed in 2015: The WaterlooClarke submission, which appeared to have an edge on the At-Home Collections, took about one week to process the Kaine collection, whereas BMI took four hours.

Although the Track coordinators were not aware of any method that was superior to BMI, they were somewhat surprised that none has emerged in the two years of the Total Recall Track. One hypothesis is that uncertainty in human relevance determinations limits the ability to measure further improvements over those achieved by BMI. Proponents of manual review processes have suggested that limitations of recall may mask the inability of automated systems to find important documents, or to find documents representing uncommon facets or subtopics. To address these concerns, the Total Recall Track had the NIST assessors identify documents they felt were "important" or "key," as well as place relevant documents into subfolders reflecting the different facets or aspects of relevance. If "important" documents were being missed by the systems, or rare subtopics were underrepresented, we would have expected to see reduced recall, according to the alternate recall measures based on these criteria. In fact, we see remarkably consistent results among the different recall measures suggesting that the systems are robust in identifying "key" documents and different components of relevance.

A limitation of any Cranfield-style test [9] is the completeness and reliability of the relevance assessments. For the Jeb Bush and Illinois Test Collections, the documents were selected for review using a combination of interactive search and judging and machine-learning techniques. While these methods are state of the art, it has been suggested that they are biased in favor of similar methods. One way to investigate this issue is to use independently labeled collections. In 2015, the Kaine and MIMIC II Collections were both independently labeled; in 2016, the Twitter Collection was likewise independently labeled. The similarity of results using these independently labeled collections suggests that bias in the selection of documents is not a major factor in the results presented here.

A second way to investigate the issue of document-selection bias, and also assessor reliability, is to use independent assessments of a non-uniform random sample of documents to calculate recall using the Horvitz-Thompson estimator, which yields an unbiased estimate. When a single independent assessor is used to determine relevance, the recall results are substantially lower than those using the official relevance assessments. This result is perhaps not surprising, as a small number of false-positive assessments in the gold standard can result in substantially underestimated recall. Put another way, if the assessor has 70% precision, a perfect system (with 100% recall and 100% precision) would achieve only 70% recall, as measured with respect to the assessor's judgments. This observation is borne out by the fact that recall rises substantially when the majority-vote-of-three assessors is used to determine relevance, rather than a single assessor.

Acknowledgement

We are grateful to the Illinois State Archive and the University of Illinois for affording us access to the Rod Blagojevich/Pat Quinn dataset for the purposes of the Sandbox evaluation. A special thanks goes to David Joens,

Brent West, Joanne Kaczmarek, and their colleagues at the Illinois State Archive and the Univerity of Illinois, for the time and effort they spent assessing these documents.

References

[1] Praveen Bommannavar, Jimmy Lin, and Anand Rajaraman. Estimating topical volume in social media streams. In *SAC '16*.

[2] Gordon V Cormack and Maura R Grossman. Multi-faceted recall of continuous active learning for technology-assisted review. In *SIGIR 2015*.

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

[4] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. Efficient construction of large test collections. In *SIGIR 1998*.

[5] 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.

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

[7] Adam Roegiest, Gordon V. Cormack, Maura R. Grossman, and Charles L. A. Clarke. TREC 2015 Total Recall Track overview. In *Proc. TREC-2015*, 2015.

[8] Ian Soboroff and Stephen Robertson. Building a filtering test collection for TREC 2002. In *SIGIR 2003*.

[9] Ellen M. Voorhees. The philosophy of information retrieval evaluation. In *Evaluation of cross-language information retrieval systems*, pages 143–170. Springer, 2002.

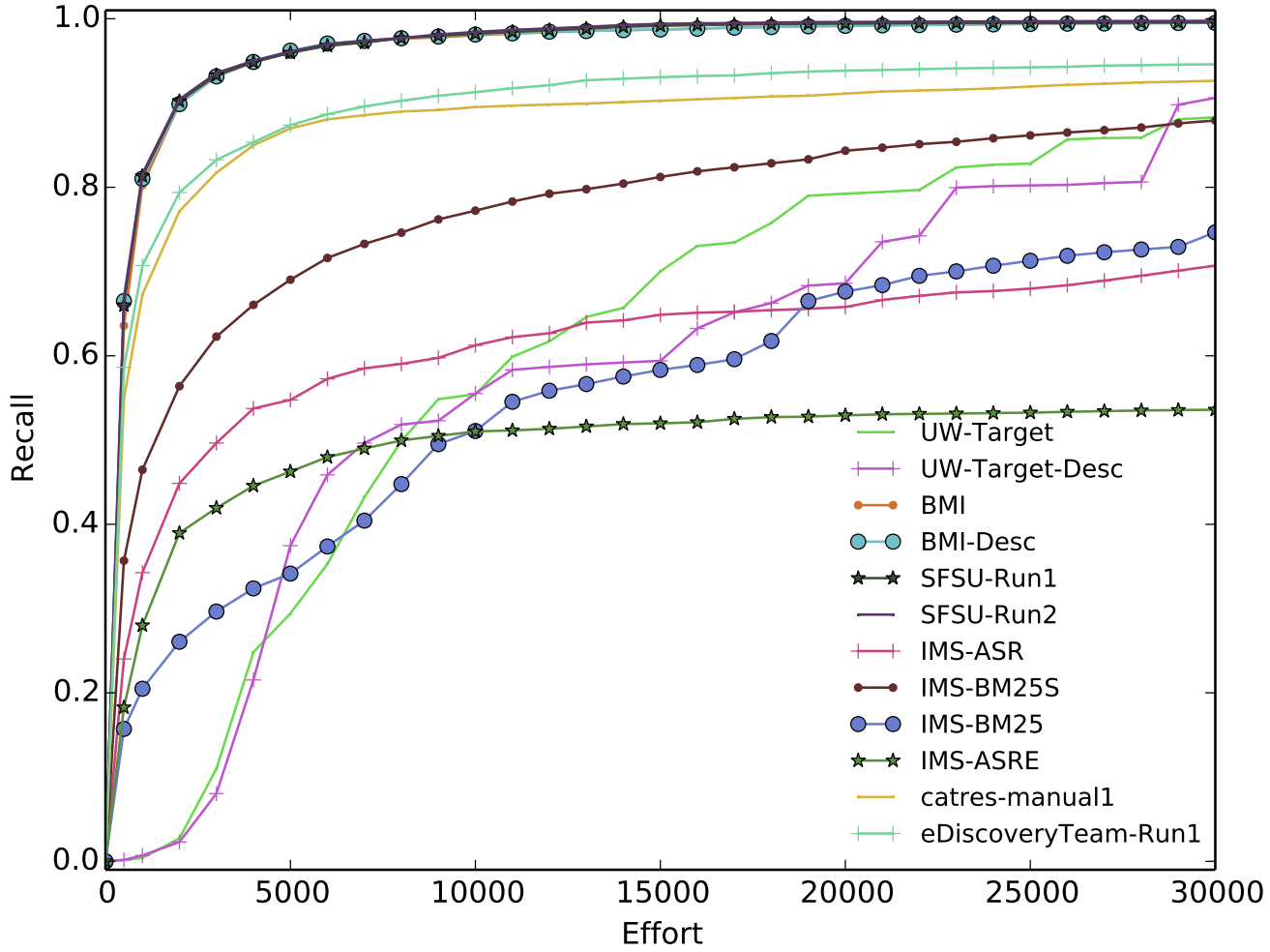

Figure 1: Gain Curves Showing Recall (Averaged Over 34 Topics) as a Function of the Number of Submitted Documents, for the Athome4 (Jeb Bush) Test Collection.

Run	Recall @									
	R	R+100	R+1000	2R	2R+100	2R+1000	4R	4R+100	4R+1000	Reasonable
BMI	.68	.79	.93	.86	.89	**.96**	.93	.93	**.97**	.945
BMI-Desc	**.71**	.82	.93	.89	.92	**.96**	**.95**	**.96**	**.97**	.951
catres	.51	.64	.79	.72	.77	.84	.83	.84	.87	.735
eDiscoveryTeam	.67	.73	.83	.79	.80	.85	.84	.85	.87	.736
ims.base	.15	.17	.25	.22	.23	.28	.28	.29	.32	.234
ims.exp	.21	.24	.38	.31	.33	.41	.37	.39	.44	.608
ims.rot	.23	.26	.40	.31	.34	.42	.38	.39	.47	.775
ims.smooth	.34	.39	.55	.44	.47	.59	.54	.56	.64	.533
sfsu_run1	.69	.82	**.94**	.88	**.92**	**.96**	**.95**	**.96**	**.97**	.969
sfsu_run2_exp	**.71**	**.83**	**.94**	**.90**	**.92**	**.96**	**.95**	**.96**	**.97**	.971
uw.desc.knee	.68	.79	.90	.87	.89	.93	.92	.93	.94	.943
uw.desc.target	.05	.05	.07	.09	.10	.16	.24	.25	.31	.924
uw.knee	.66	.78	.91	.84	.87	.93	.90	.90	.93	.949
uw.target	.04	.04	.09	.13	.13	.19	.27	.28	.31	.926

Figure 2: Recall @ aR+b for the Athome4 (Jeb Bush) Test Collection.

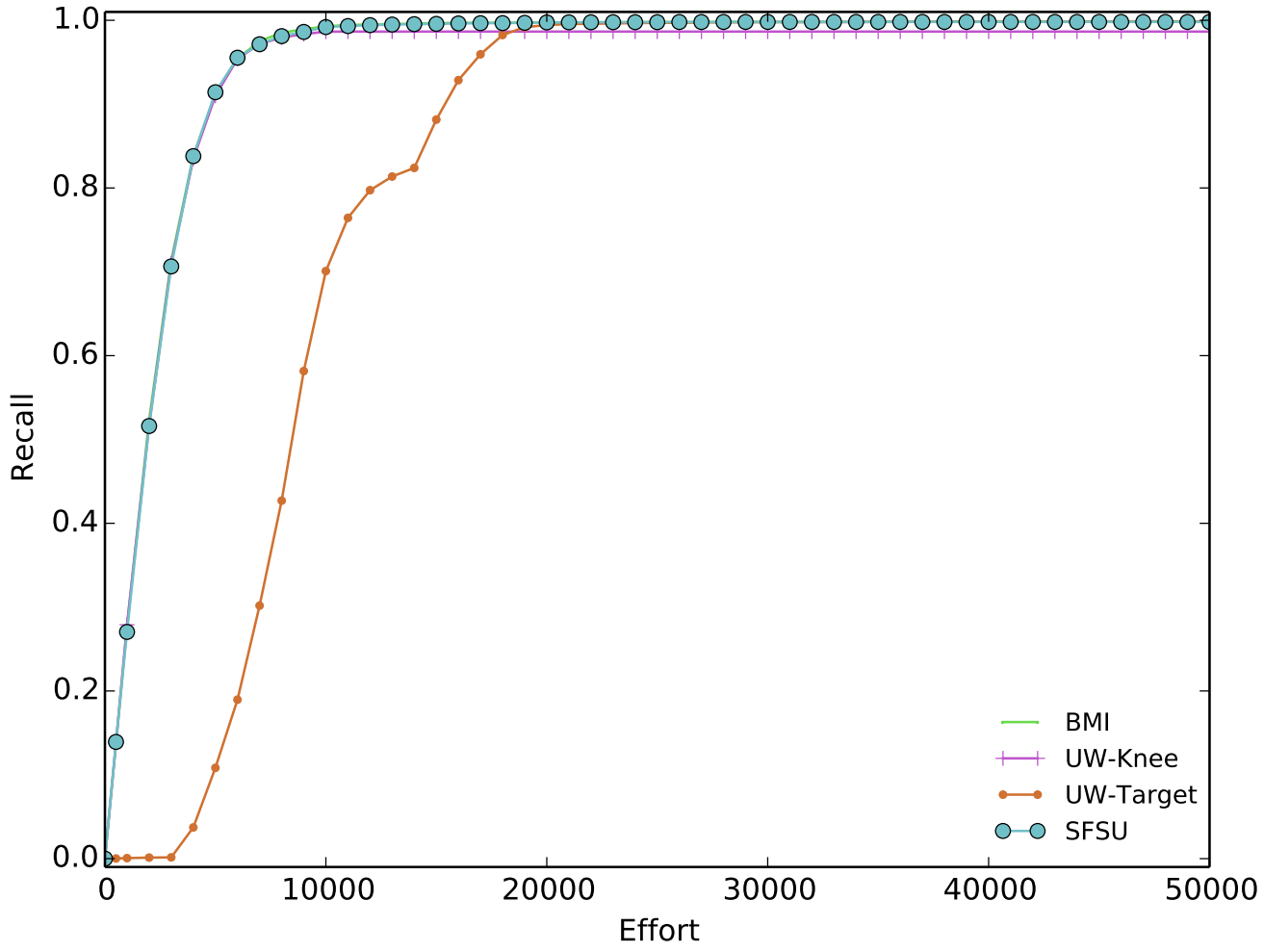

Figure 3: Gain Curves Showing Recall (Averaged Over Six Topics) as a Function of the Number of Submitted Documents, for the Illinois (Rod Blagojevich/Pat Quinn) Test Collection.

Run	Recall @									
	R	R+100	R+1000	2R	2R+100	2R+1000	4R	4R+100	4R+1000	Reasonable
BMI	**.75**	**.77**	**.88**	**.96**	**.96**	**.97**	**.99**	**.99**	**.99**	.949
sfsu	**.75**	**.77**	.87	.95	**.96**	**.97**	**.99**	**.99**	**.99**	.962
uw.knee	**.75**	.76	.87	**.96**	**.96**	**.97**	.98	.98	.98	.986
uw.target	**.07**	.08	.14	.33	.34	.44	.65	.65	.66	.960

Figure 4: Recall @ aR+b for the Illinois (Rod Blagojevich/Pat Quinn) Test Collection.

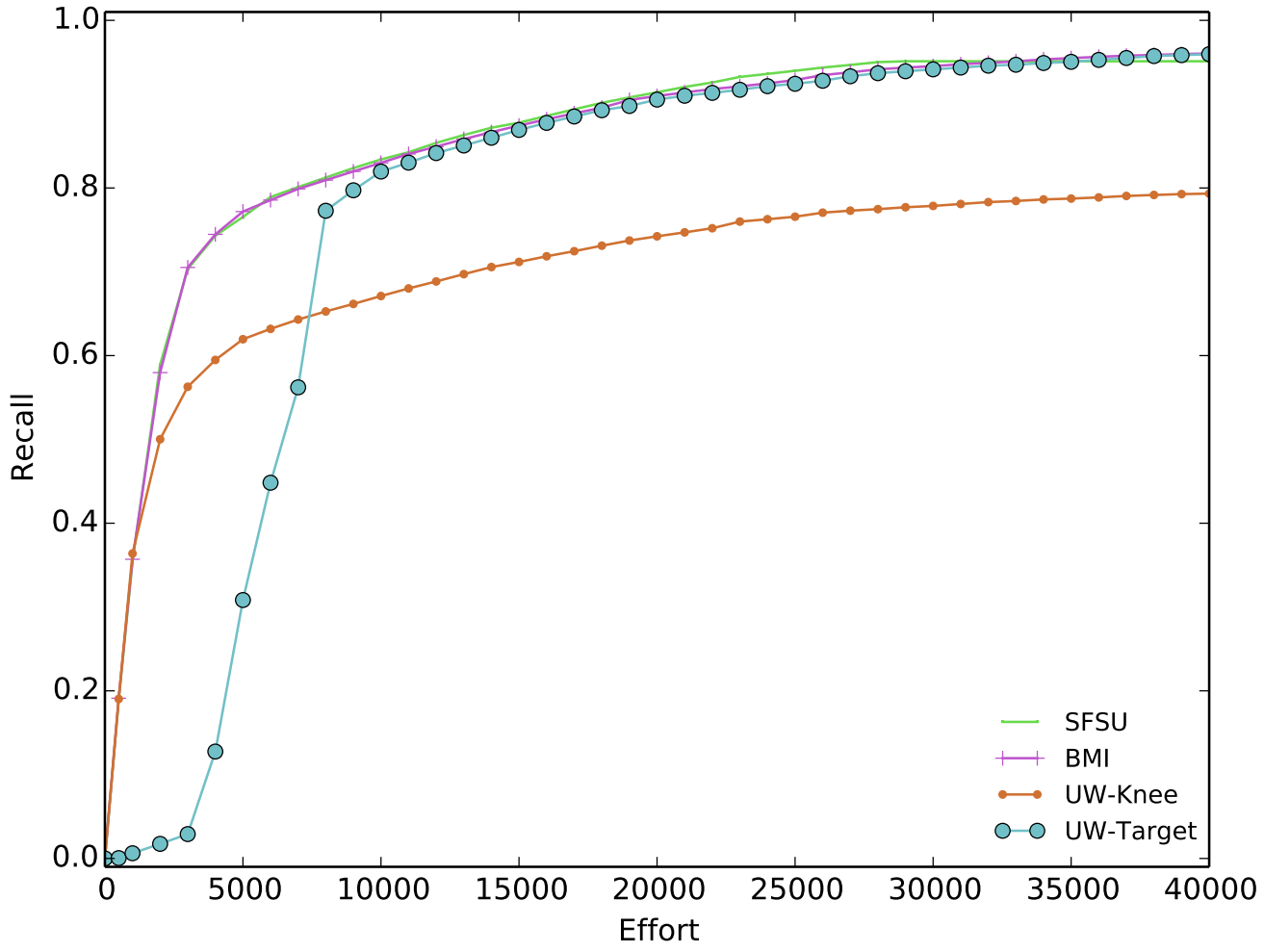

Figure 5: Gain Curves Showing Recall (Averaged Over 4 Topics) as a Function of the Number of Submitted Documents, for the Twitter Test Collection.

Run	Recall @									
	R	R+100	R+1000	2R	2R+100	2R+1000	4R	4R+100	4R+1000	Reasonable
BMI	**.80**	.81	.87	**.93**	**.94**	**.95**	**.96**	**.96**	**.96**	.927
sfsu	**.80**	**.82**	**.88**	**.93**	.93	.93	.94	.94	.94	.935
uw.knee	.71	.72	.74	.79	.79	.79	.80	.80	.80	.801
uw.target	.18	.19	.26	.51	.52	.62	.71	.71	.72	.934

Figure 6: Recall @ aR+b for the Twitter Test Collection.

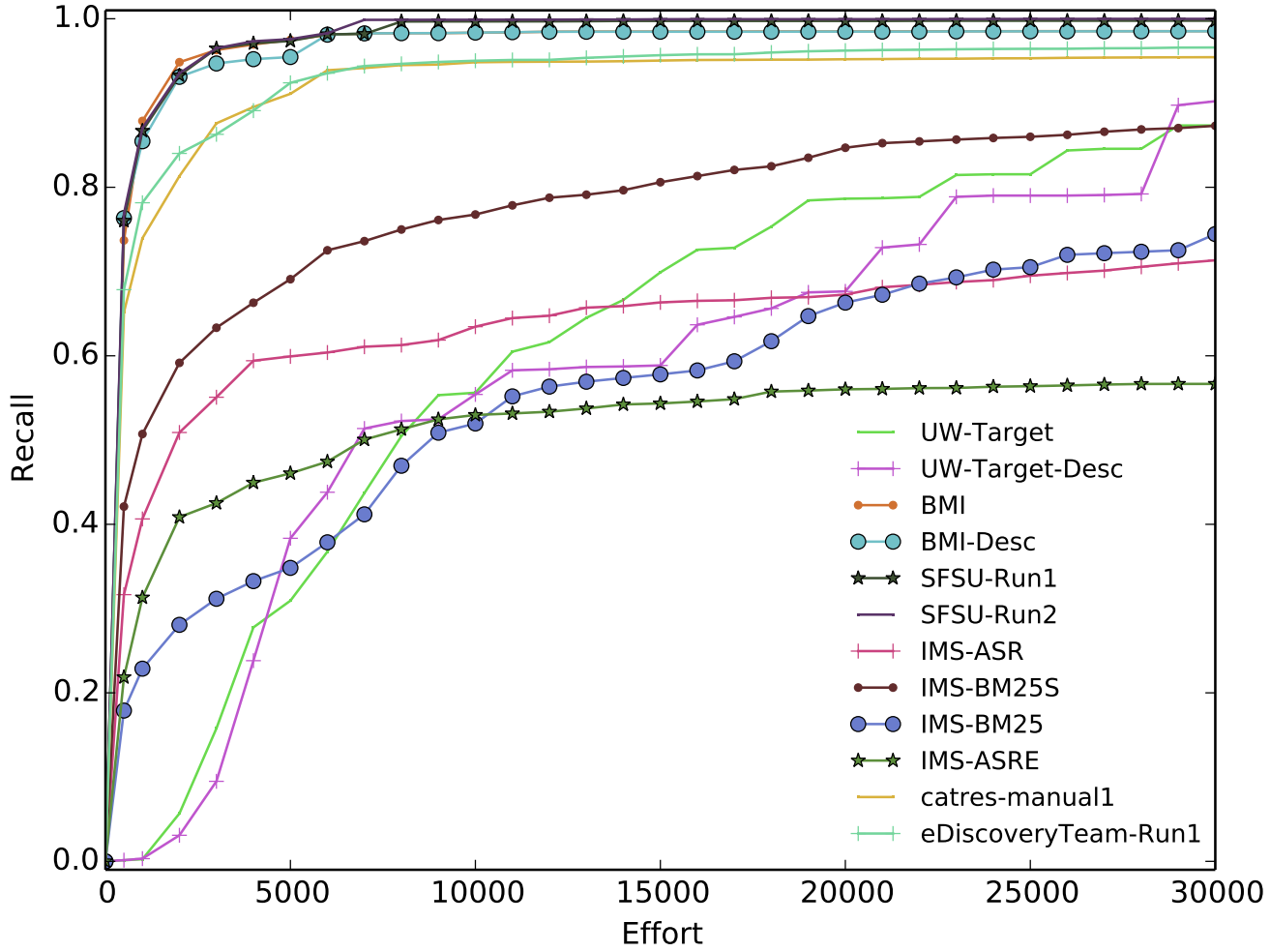

Figure 7: Gain Curves Showing "Important" or "Key" Document Recall (Averaged Over 34 Topics) as a Function of the Number of Submitted Documents, for the Athome4 (Jeb Bush) Test Collection.

Run	Recall @									
	R	R+100	R+1000	2R	2R+100	2R+1000	4R	4R+100	4R+1000	Reasonable
BMI	.74	.84	.95	.89	.91	.97	.95	.95	**.99**	.967
BMI-Desc	.78	.87	.96	.92	.95	.97	**.98**	.98	**.99**	.964
catres	.62	.75	.85	.80	.84	.89	.90	.90	.91	.835
eDiscoveryTeam	.77	.82	.88	.86	.87	.90	.90	.91	.92	.822
ims_base	.17	.19	.27	.23	.24	.29	.30	.30	.33	A.255
ims_exp	.22	.24	.38	.30	.32	.42	.36	.38	.44	.642
ims_rot	.29	.33	.44	.36	.39	.47	.42	.43	.51	.797
ims_smooth	.37	.43	.59	.47	.51	.62	.59	.61	.68	.577
sfsu_run1	.78	.88	**.97**	.93	**.96**	.98	**.98**	**.99**	**.99**	.988
sfsu_run2_exp	**.79**	**.90**	**.97**	**.95**	**.96**	**.99**	**.98**	**.99**	**.99**	.990
uw.desc.knee	.76	.84	.92	.89	.91	.93	.94	.94	.95	.950
uw.desc.target	.04	.04	.07	.09	.10	.16	.26	.26	.32	.936
uw.knee	.75	.84	.94	.89	.91	.96	.92	.93	.96	.964
uw.target	.03	.03	.10	.14	.14	.21	.27	.28	.31	.947

Figure 8: "Important" or "Key" Document Recall @ aR+b for the Athome4 (Jeb Bush) Test Collection.

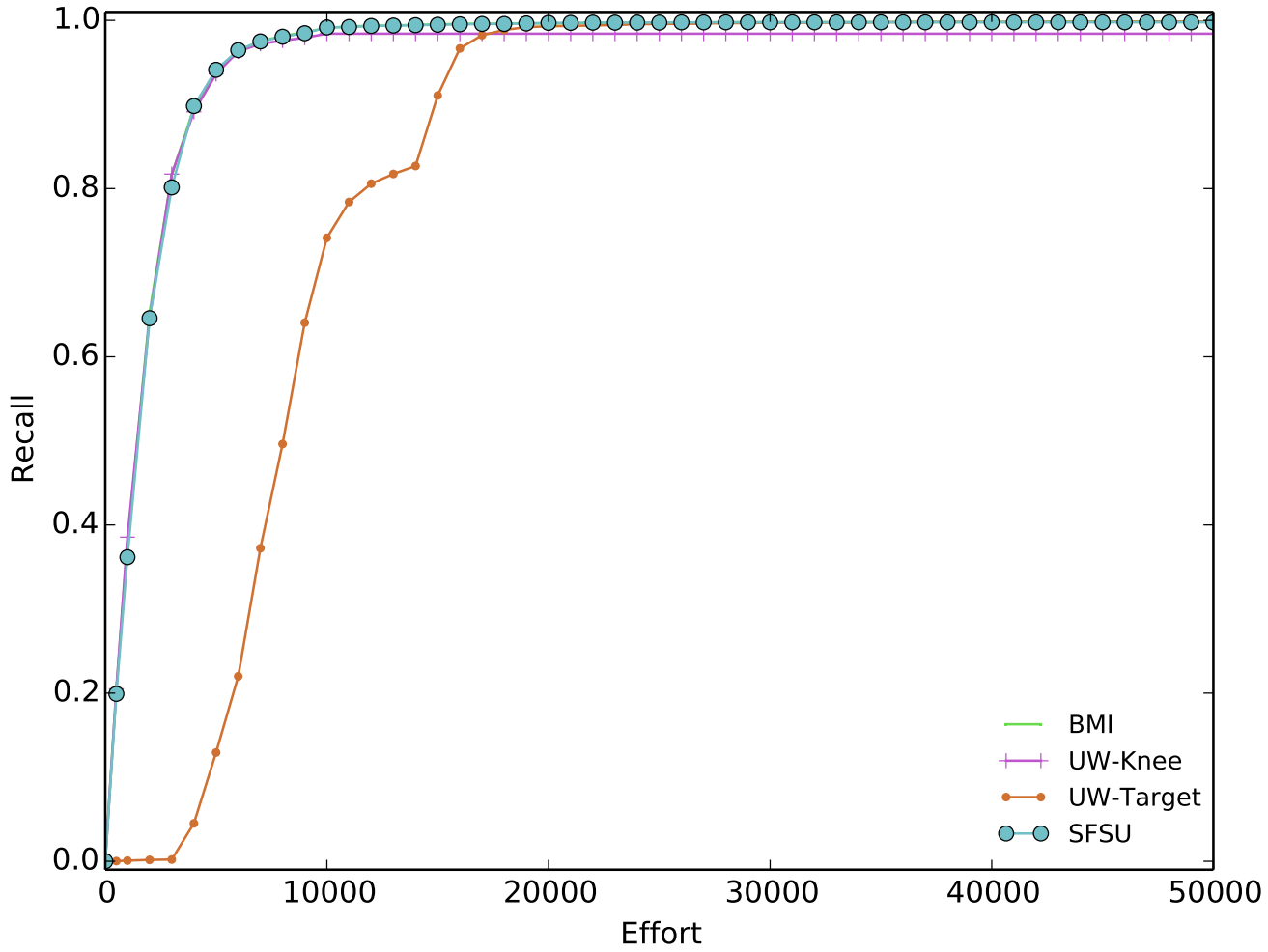

Figure 9: Gain Curves Showing "Important" or "Key" Document Recall (Averaged Over Six Topics) as a Function of the Number of Submitted Documents, for the Illinois (Rod Blagojevich/ Pat Quinn) Test Collection.

Run	Recall @									
	R	R+100	R+1000	2R	2R+100	2R+1000	4R	4R+100	4R+1000	Reasonable
BMI	.83	**.85**	**.92**	**.97**	**.97**	**.98**	.98	.98	**.99**	.963
sfsu	**.84**	**.85**	**.92**	.96	**.97**	**.98**	**.99**	**.99**	**.99**	.972
uw.Knee	.83	.84	**.92**	**.97**	**.97**	.97	.98	.98	.98	.984
uw.Target	.08	.09	.15	.36	.38	.50	.66	.66	.66	.970

Figure 10: "Important" or "Key" Document Recall @ aR+b for the Illinois (Rod Blagojevich/ Pat Quinn) Test Collection.

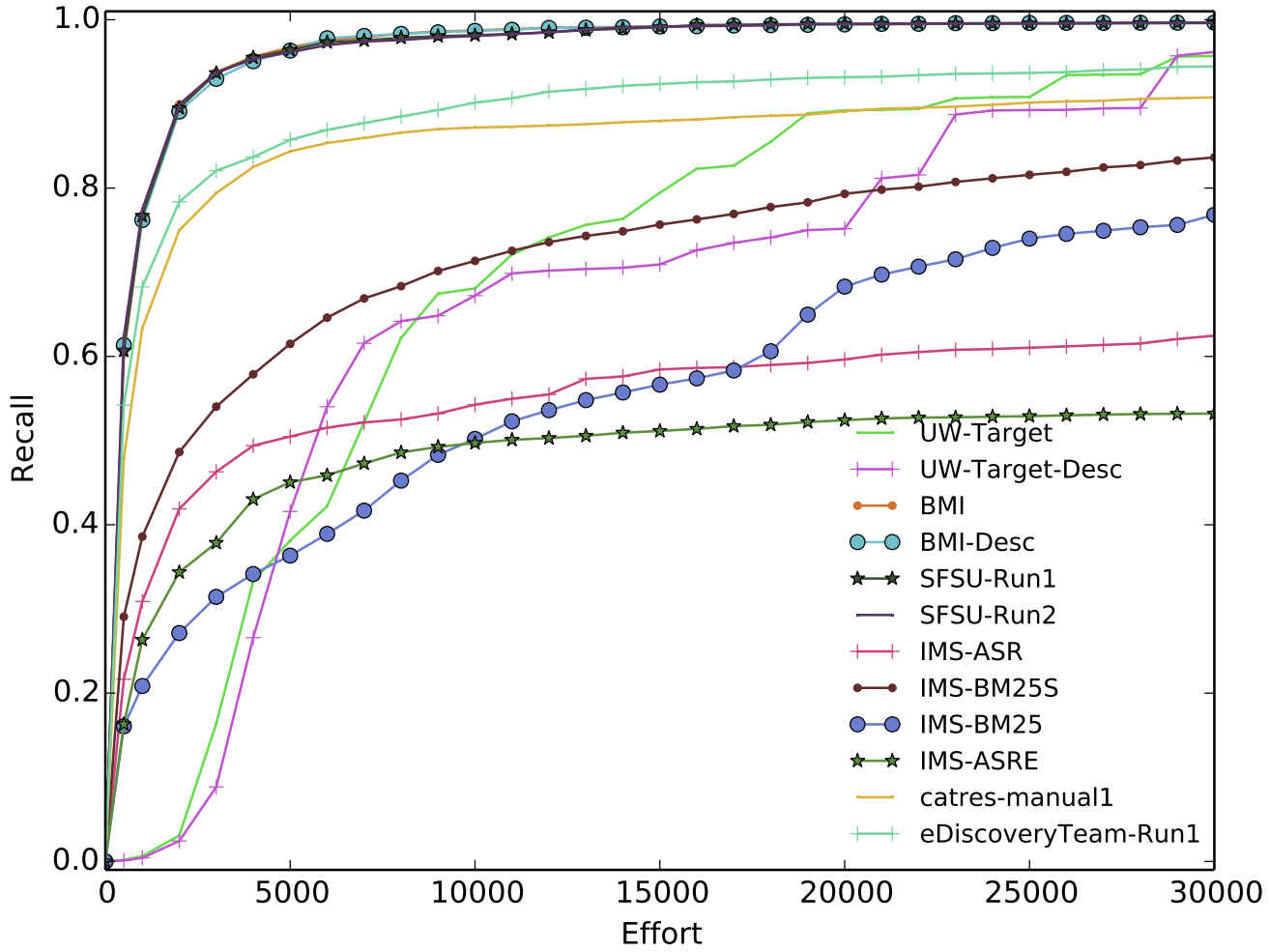

Figure 11: Gain Curves Showing Facet or Subtopic Recall (Macro-Averaged Over 348 Subtopics of 34 Topics) as a Function of the Number of Submitted Documents, for the Athome4 (Jeb Bush) Test Collection.

Run	Recall @									
	R	R+100	R+1000	2R	2R+100	2R+1000	4R	4R+100	4R+1000	Reasonable
BMI	.67	.77	.93	.86	.89	**.96**	.94	.95	**.97**	.946
BMI-Desc	**.69**	.78	.92	.88	.90	**.96**	**.95**	**.96**	**.97**	.949
catres	.52	.60	.75	.71	.74	.82	.82	.83	.86	.706
eDiscoveryTeam	.65	.71	.81	.78	.80	.84	.84	.85	.87	.735
ims_base	.17	.18	.26	.23	.25	.30	.32	.33	.36	.244
ims_exp	.19	.20	.33	.27	.28	.36	.33	.35	.39	.629
ims_rot	.22	.26	.36	.32	.33	.40	.38	.38	.44	.710
ims_smooth	.30	.33	.47	.40	.43	.52	.53	.54	.59	.454
sfsu_run1	.66	.76	**.93**	.87	**.91**	**.96**	**.96**	**.96**	**.97**	.967
sfsu_run2_exp	**.67**	**.78**	**.93**	**.88**	**.91**	**.96**	**.96**	**.96**	**.97**	.969
uw.desc.knee	.67	.77	.91	.87	.88	.93	.92	.93	.94	.950
uw.desc.target	.05	.05	.07	.10	.11	.19	.31	.32	.40	.907
uw.knee	.65	.74	.89	.85	.87	.92	.90	.91	.93	.939
uw.target	.04	.04	.10	.17	.18	.27	.34	.35	.40	.916

Figure 12: Facet or Subtopic Recall (Macro-Averaged Over 348 Subtopics of 34 Topics) @ aR+b for the Athome4 (Jeb Bush) Test Collection.

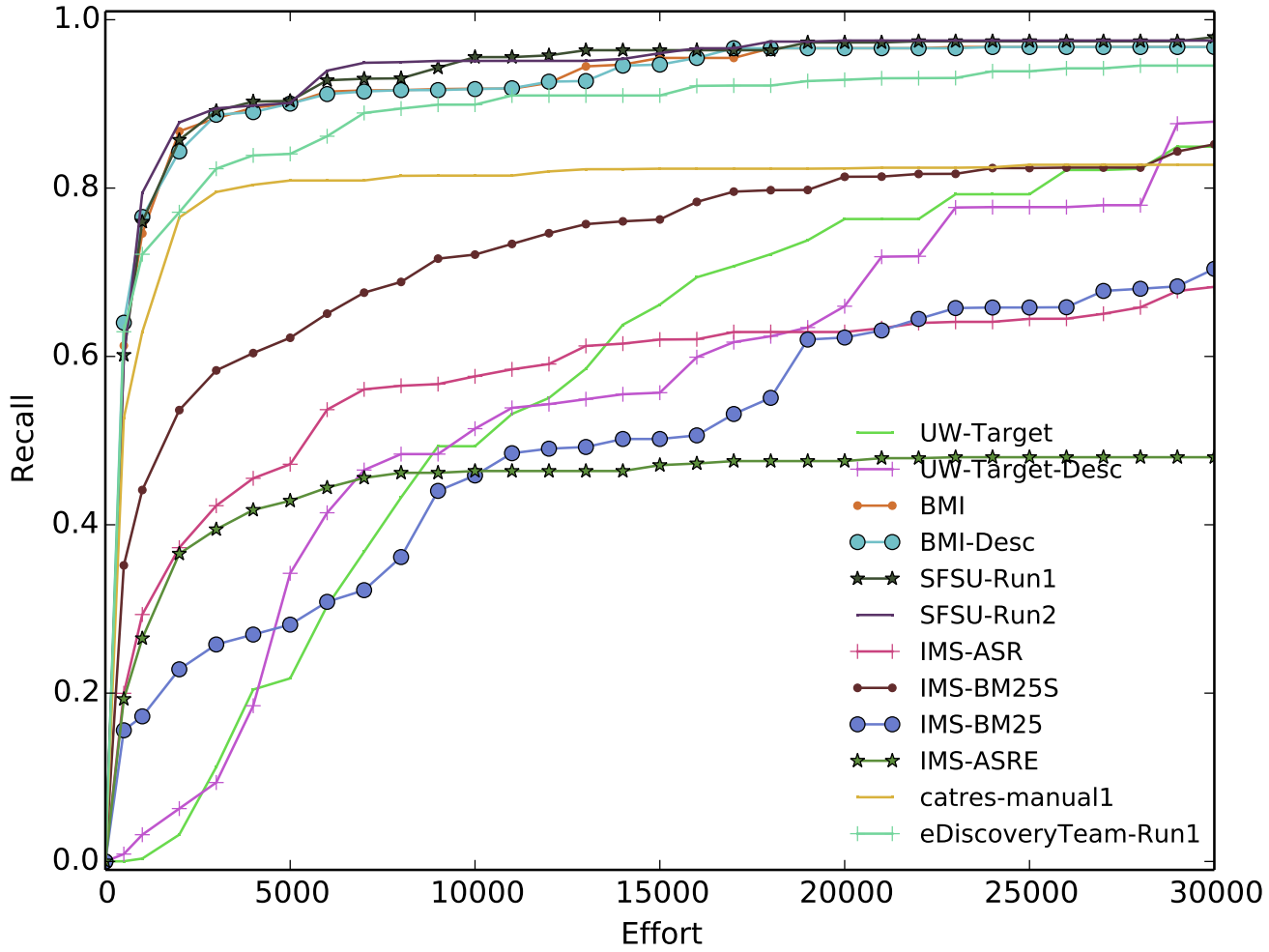

Figure 13: Gain Curves Showing Recall According to the Majority Vote of the Three Secondary Assessors (Averaged Over 34 Topics) as a Function of the Number of Submitted Documents, for the Athome4 (Jeb Bush) Test Collection.

Run	Recall @									
	R	R+100	R+1000	2R	2R+100	2R+1000	4R	4R+100	4R+1000	Reasonable
BMI	.59	.68	.86	.74	.79	.88	.83	.83	**.92**	.883
BMI-Desc	.62	.70	.85	**.76**	.79	.89	.85	**.87**	**.92**	.862
catres	.51	.61	.73	.68	.72	.75	.74	.76	.76	.709
eDiscovery	**.66**	.73	.77	.75	.76	.79	.79	.79	.81	.742
ims_base	.13	.14	.22	.20	.20	.24	.24	.25	.27	.200
ims_exp	.20	.24	.35	.29	.31	.37	.35	.36	.41	.555
ims_rot	.19	.24	.33	.26	.28	.35	.32	.32	.40	.728
ims_smooth	.32	.37	.52	.41	.46	.55	.52	.53	.59	.485
sfsu_run1	.61	.71	.85	.73	.79	.87	**.86**	**.87**	.90	.895
sfsu_run2_exp	.61	**.74**	**.89**	**.76**	**.81**	**.91**	.82	.84	.91	.912
uw.desc.knee	.63	.72	.84	.74	.79	.86	.83	.84	.86	.861
uw.desc.target	.09	.09	.10	.11	.11	.16	.20	.21	.29	.825
uw.knee	.58	.66	.81	.71	.77	.85	.80	.81	.86	.860
uw.target	.03	.03	.10	.13	.13	.17	.20	.20	.26	.856

Figure 14: Recall @ aR+b for the Majority Vote of Three Secondary Assessors, for the Athome4 (Jeb Bush) Test Collection.

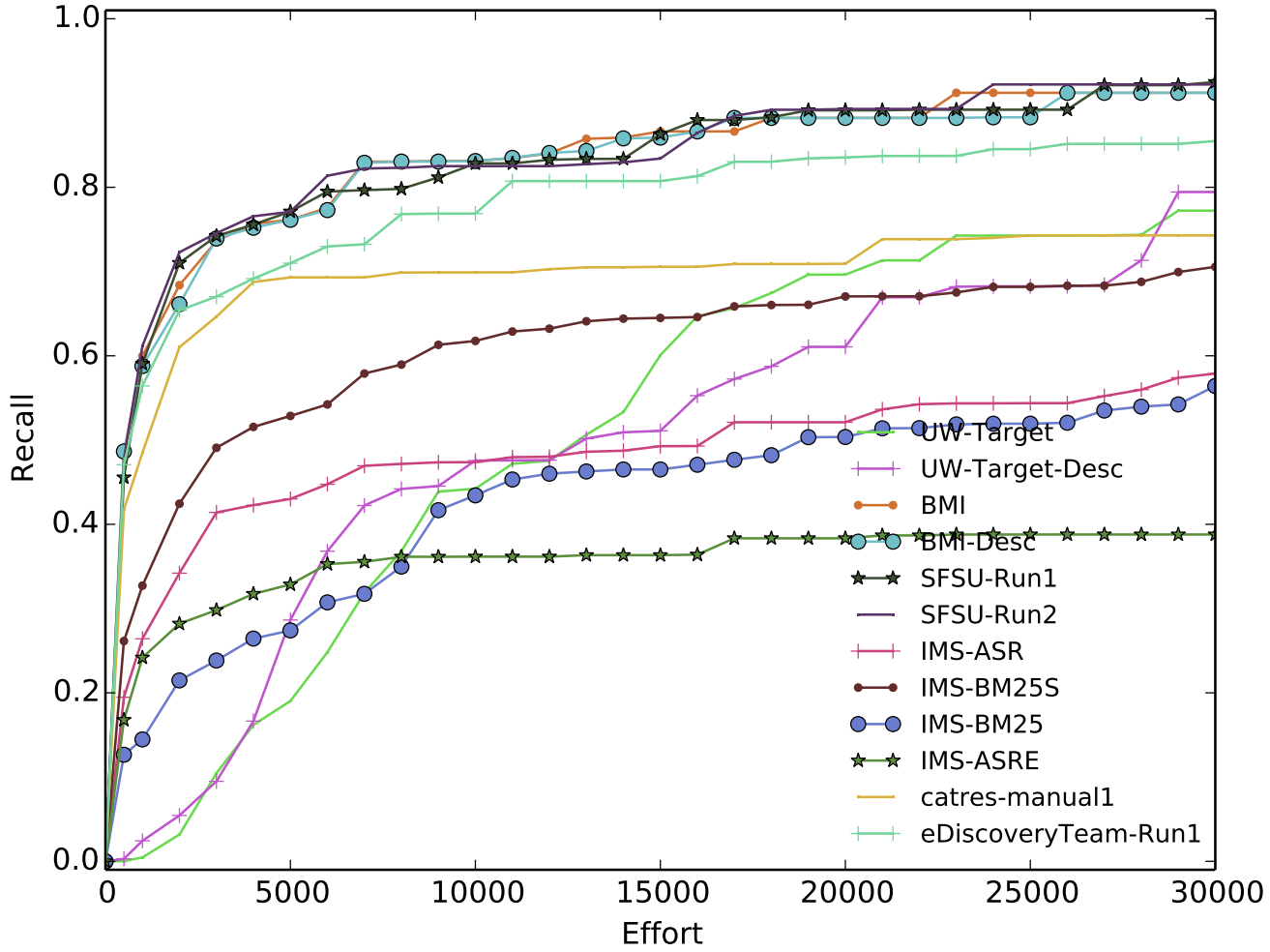

Figure 15: Gain Curves Showing Recall According to the First of Three Secondary Assessors (Averaged Over 34 Topics) as a Function of the Number of Submitted Documents, for the Athome4 (Jeb Bush) Test Collection.

Run	Recall @									
	R	R+100	R+1000	2R	2R+100	2R+1000	4R	4R+100	4R+1000	Reasonable
BMI	.47	**.55**	**.71**	.62	**.64**	**.74**	.70	.70	**.76**	.733
BMI-Desc	.48	.52	.67	.60	.63	.71	.70	.71	.73	.721
catres	.39	.50	.57	.53	.57	.62	.62	.62	.63	.572
eDiscoveryTeam	**.52**	.57	.63	.60	.60	.64	.64	.65	.69	.604
ims_base	.13	.13	.20	.18	.18	.24	.22	.23	.26	.181
ims_exp	.18	.21	.29	.24	.26	.30	.30	.31	.34	.464
ims_rot	.19	.23	.30	.26	.26	.31	.29	.30	.35	.621
ims_smooth	.26	.30	.41	.33	.36	.45	.41	.42	.49	.375
sfsu_run1	.48	.54	.68	.59	.62	.73	.71	.72	.75	.736
sfsu_run2_exp	.49	.57	.68	**.64**	.65	.73	**.72**	**.73**	.75	.737
uw.desc.knee	.49	.55	.64	.59	.62	.67	.66	.66	.67	.700
uw.desc.target	.09	.09	.09	.11	.11	.15	.17	.19	.26	.729
uw.knee	.46	.53	.66	.59	.63	.70	.67	.67	.70	.735
uw.target	.03	.03	.09	.12	.12	.16	.18	.18	.22	.681

Figure 16: Recall @ aR+b for the First of Three Secondary Assessors, for the Athome4 (Jeb Bush) Test Collection.

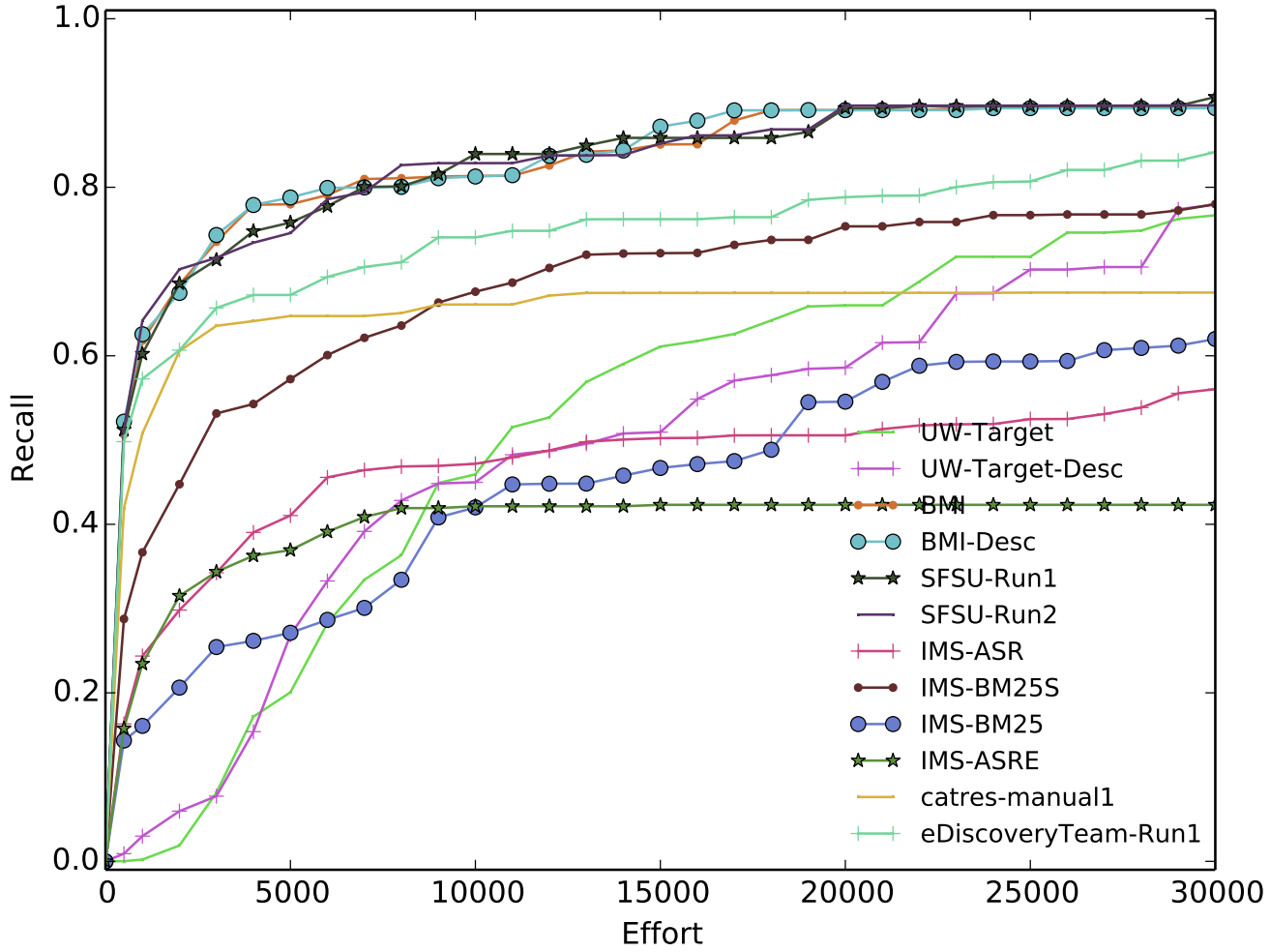

Figure 17: Gain Curves Showing Recall According to the Second of Three Secondary Assessors (Averaged Over 34 Topics) as a Function of the Number of Submitted Documents, for the Athome4 (Jeb Bush) Test Collection.

Run	Recall @									
	R	R+100	R+1000	2R	2R+100	2R+1000	4R	4R+100	4R+1000	Reasonable
BMI	.47	.52	**.69**	.57	.60	.73	.65	.67	.76	.717
BMI-Desc	.49	.54	**.69**	.60	.63	**.74**	**.69**	**.72**	**.77**	.701
catres	.39	.47	.58	.52	.56	.59	.60	.61	.61	.565
eDiscoveryTeam	**.51**	.57	.63	**.61**	.61	.64	.64	.64	.65	.579
ims_base	.12	.12	.20	.17	.18	.22	.22	.23	.26	.191
ims_exp	.16	.18	.30	.23	.24	.32	.27	.27	.35	.511
ims_rot	.14	.18	.26	.18	.20	.27	.24	.24	.31	.594
ims_smooth	.28	.33	.43	.35	.40	.46	.43	.43	.52	.404
sfsu_run1	.49	.56	.66	.58	.63	.69	**.69**	.69	.75	.710
sfsu_run2_exp	.50	**.58**	.68	**.61**	**.64**	.71	.67	.68	.75	.726
uw.desc.knee	.50	.56	.67	**.61**	.65	.71	.67	.68	.71	.713
uw.desc.target	.06	.06	.07	.07	.07	.10	.14	.15	.25	.705
uw.knee	.48	.52	.67	.58	.62	.72	.65	.66	.73	.734
uw.target	.02	.02	.08	.08	.08	.11	.17	.17	.23	.694

Figure 18: Recall @ aR+b for the Second of Three Secondary Assessors, for the Athome4 (Jeb Bush) Test Collection.

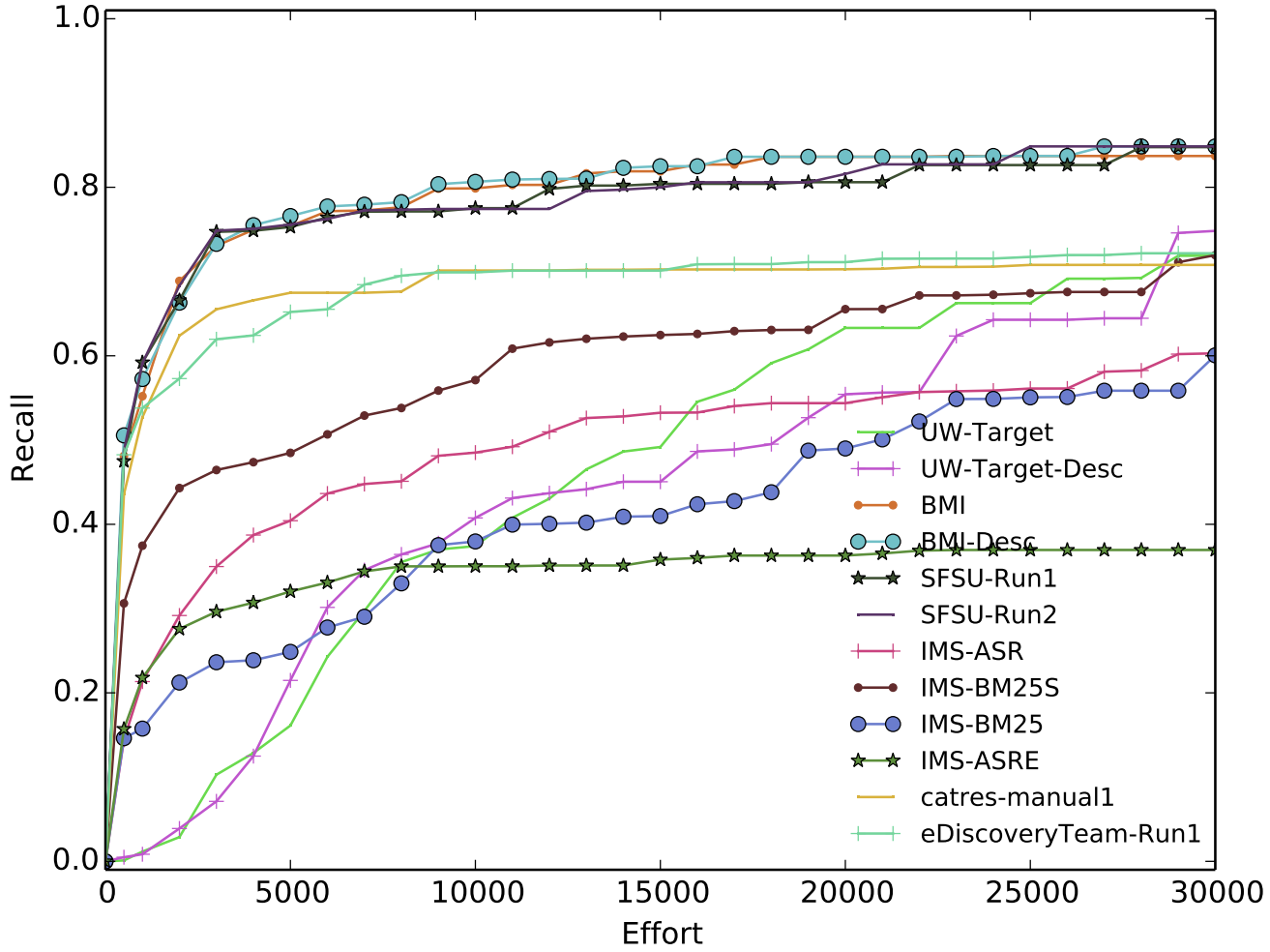

Figure 19: Gain Curves Showing Recall According to the Third of Three Secondary Assessors (Averaged Over 34 Topics) as a Function of the Number of Submitted Documents, for the Athome4 (Jeb Bush) Test Collection.

Run	Recall @									
	R	R+100	R+1000	2R	2R+100	2R+1000	4R	4R+100	4R+1000	Reasonable
BMI	.42	.51	.65	.57	.61	.69	.66	.67	**.74**	.699
BMI-Desc	.44	.53	.65	.58	.61	.71	.70	.70	**.74**	.668
catres	.37	.46	.59	.54	.57	.63	.61	.62	.65	.599
eDiscoveryTeam	**.50**	.54	.56	.55	.56	.59	.61	.61	.61	.587
ims_base	.11	.12	.21	.19	.19	.22	.22	.23	.24	.187
ims_exp	.15	.18	.24	.20	.22	.27	.26	.26	.30	.458
ims_rot	.16	.19	.25	.21	.22	.27	.27	.28	.32	.654
ims_smooth	.25	.28	.42	.34	.36	.44	.40	.42	.46	.404
sfsu_run1	.45	.54	.68	.58	.62	.72	**.69**	**.71**	.73	.728
sfsu_run2exp	.45	.55	**.69**	**.60**	**.63**	**.73**	.66	.70	.74	.735
uw.desc.knee	.45	**.56**	.65	.58	.62	.69	.67	.68	.69	.692
uw.desc.target	.05	.06	.06	.08	.08	.12	.13	.13	.18	.655
uw.knee	.40	.49	.60	.53	.59	.66	.62	.63	.66	.692
uw.target	.03	.03	.08	.10	.11	.13	.14	.14	.17	.712

Figure 20: Recall @ aR+b for the Third of Three Secondary Assessors, for the Athome4 (Jeb Bush) Test Collection.