On Design and Evaluation of High-Recall Retrieval Systems for Electronic Discovery

by

Adam Roegiest

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2017

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	DOUGLAS W. OARD Professor
Supervisor	GORDON V. CORMACK Professor
Internal Examiner	CHARLES L. A. CLARKE Professor
Internal Examiner	JIMMY LIN Professor
Internal-external Examiner	MARK D. SMUCKER Associate Professor

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The following is a list of publications that contain preliminary versions of material that is presented in this thesis:

- Adam Roegiest and Gordon V. Cormack, Impact of Review-Set Selection on Human Assessment for Text Classification, In *SIGIR 2016*. 4 pages.
- Adam Roegiest and Gordon V. Cormack, An Architecture for Privacy-Preserving and Replicable High-Recall Retrieval Experiments, In *SIGIR 2016*. 4 pages.
- Adam Roegiest, Gordon V. Cormack, Maura R. Grossman, and Charles L. A. Clarke, TREC 2015 Total Recall Track Overview, In *TREC 2015*. 27 pages.
- Adam Roegiest and Gordon V. Cormack, Total Recall Track Tools Architecture Overview, In *TREC 2015*. 11 pages.
- Adam Roegiest, Gordon V. Cormack, Charles L. A. Clarke, and Maura R. Grossman, Impact of Surrogate Assessments on High-Recall Retrieval, In *SIGIR 2015*. 10 pages.

The following is a list of publications related to the topics in this thesis:

- Charles L. A. Clarke, Gordon V. Cormack, Jimmy Lin, and Adam Roegiest, Total Recall: Blue Sky on Mars, In *ICTIR 2016*. 4 pages.
- Charles L. A. Clarke, Gordon V. Cormack, Jimmy Lin, and Adam Roegiest, Ten Blue Links on Mars, In *WWW 2016*. 9 pages.
- Adam Roegiest, Luchen Tan, Jimmy Lin, and Charles L. A. Clarke, A Platform for Streaming Push Notifications to Mobile Assessors. In *SIGIR 2016*. 4 pages.

Abstract

High-recall retrieval is an information retrieval task model where the goal is to identify, for human consumption, all, or as many as practicable, documents relevant to a particular information need. This thesis investigates the ways in which one can evaluate high-recall retrieval systems and explores several design considerations that should be accounted for when designing such systems for electronic discovery.

The primary contribution of this work is a framework for conducting high-recall retrieval experimentation in a controlled and repeatable way. This framework builds upon lessons learned from similar tasks to facilitate the use of retrieval systems on collections that cannot be distributed due to the sensitivity or privacy of the material contained within. Accordingly, a Web API is used to distribute document collections, informations needs, and corresponding relevance assessments in a one-document-at-a-time manner. Validation is conducted through the successful deployment of this architecture in the 2015 TREC Total Recall track over the live Web and in controlled environments.

Using the runs submitted to the Total Recall track and other test collections, we explore the efficacy of a variety of new and existing effectiveness measures to high-recall retrieval tasks. We find that summarizing the trade-off between recall and the effort required to attain that recall is a non-trivial task and that several measures are sensitive to properties of the test collections themselves. We conclude that the gain curve, a de facto standard, and variants of the gain curve are the most robust to variations in test collection properties and the evaluation of high-recall systems.

This thesis also explores the effect that non-authoritative, surrogate assessors can have when training machine learning algorithms. Contrary to popular thought, we find that surrogate assessors appear to be inferior to authoritative assessors due to differences of opinion rather than innate inferiority in their ability to identify relevance. Furthermore, we show that several techniques for diversifying and liberalizing a surrogate assessor's conception of relevance can yield substantial improvement in the surrogate and, in some cases, rival the authority.

Finally, we present the results of a user study conducted to investigate the effect that three archetypal high-recall retrieval systems have on judging behaviour. Compared to using random and uncertainty sampling, selecting documents for training using relevance sampling significantly decreases the probability that a user will identify that document as relevant. On the other hand, no substantial difference between the test conditions is observed in the time taken to render such assessments.

Acknowledgements

I would like to begin by thanking my supervisor, Gordon V. Cormack, for the time, effort, and patience he has put into getting me to this point. I am particularly thankful for Gord's continual tolerance of my wandering interests, which were not always related to my thesis. Gord has provided an excellent role model for the type of researcher that I can only hope to one day be. I have known Gord for most of my academic career, I would not have gotten here without him.

I must also thank Charles L. A. Clarke, who at times played the role of ersatz supervisor when Gord was unavailable. It has been my immense pleasure to work with Charlie over the last several years, and had circumstances worked out differently I would have been grateful to call him supervisor.

I thank Jimmy Lin for being a fun mentor and collaborator. I especially appreciate the fact that Jimmy does not shy away from being Jimmy in all circumstances.

Maura R. Grossman has been an insightful source of advice and a constant advocate for keeping me grounded in the real world. I thank her for that.

Thanks go to my examining committee: Gordon V. Cormack, Charles L. A. Clarke, Jimmy Lin, Mark D. Smucker, and Douglas W. Oard. I truly appreciate the comments on this thesis that you all provided.

As is ever the case, bureaucracy can be a bit overwhelming. At the University of Waterloo, Wendy Rush, Paula Roser, Vic DiCiccio, Jean Webster, Gang Lu, and Gordon Boerke have made it less so, and for that I thank them. I would particularly like to thank Wendy Rush for always putting in extra effort to make sure things got done correctly and as quickly as possible, and for commiserating with me when things did not go as planned.

Additionally, I would like to thank Ruth Taylor for taking the time to proofread this thesis.

During my travels for my doctorate, I have been graced with the opportunity to associate with many great researchers and would like to thank the following standouts for taking their time to chat with the new kid on the block: Shane Culpepper, Andrew Trotman, Jeremy Pickens, Jaap Kamps, Fernando Diaz, Craig MacDonald, Douglas W. Oard, Ian Soboroff, Ellen Voorhees, Carsten Eickoff, Roberto Konow, Ricardo Baeza-Yates, Arjen de Vries, Grace Hui Yang, Laura Dietz, Seamus Lawless,Yubim Kim, Hadi Hashemi, Guido Zuccon, and Julia Kiseleva.

Out of that list of researchers, I must highlight Ian Soboroff and Ellen Voorhees. Without their continual work with TREC at NIST and their corresponding contributions to the field of IR, much of this thesis would not exist. For that, I thank you both.

I would like to thank all of my friends, colleagues, and co-workers at the University of Waterloo. among them: Adriel Dean-Hall, Gaurav Baruah, Ashif Harji, Brad Lushman, Maheedhar Kolla, Nomair Naeem, Bahareh Sarrafzadeh, Luchen Tan, Aaron Moss, Gaurav Baruah, Rob Schluntz, Dan Brotherston, Sharon Choi, Cecylia Bocovich, Sasha Vtyurina, Kirsten Bradley, Michael Mior, John Akinyemi, Azin Ashkan, Matt Crane, Rob Warren, Marianna Rapoport, Jeff Luo, and Jack Thomas.

Outside of academia, I have had the pleasure to get to know a great many people and would like to thank them all for the fun times we've had over the last 10+ years. Such friends include but are not limited to: Sergey Bobkin, Adam McFarlane, Pavol and Kim Chvála, David Santos, Robert McFadden, Jennifer Taves, Scott Reid, Amber West, Leo Gimenez, Andrew Peterson, Sanya Sagar, Melissa Milley, Chantal Austin, Joanna McClintock, John Kemp, Ronny Wan, Jillian Sauder, Britney Chordash, Daniel Gimenez Paez, Sionaid Eggett, and Alex Struk.

Finally, I would like to thank my family for enduring my occasionally delayed trips back to Wallaceburg. Thanks go to my brothers, Bryan Wade and Craig Wade, their wives, Keelan Wade and Angela Wade, and all their children (Madeline, Lucas, Melody, Liam, Landon, Lily, Leo, Evan, Rita, and Lauren) for housing and feeding me when I did visit. I would also like to thank my aunts, Marie Olmsted and Irma Brunt, and my uncle, Ted Olmsted, for always making my returns warm.

To my mother, Jeanetta Roegiest, and step-father, Leonard Atkins, I thank you for supporting me in every way you could over the past years.

To my late father, Nico Pieter Dubois, I thank you for all the support you gave me and for moulding me into who I am today.

To all those listed above, there are not enough words to my express my gratitude. I leave you with two:

Thank you.

Dedication

To all those who have made this thesis possible.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

High-recall retrieval refers to tasks performed when a user desires the identification of all (or nearly all) relevant material, often with minimal effort on their part. Specific examples include electronic discovery ("eDiscovery") in civil litigation [39], systematic review in evidence-based medicine [78], and the preparation of test collections for information retrieval research [48]. These are not the only high-recall tasks but they form the basis of much of the on-going research in the field. Indeed, other applications such as spam filtering and adaptive filtering are long established; they are well studied problems that are high-recall retrieval tasks in disguise. Even "common" tasks such as literature review when writing a dissertation and vanity search are high-recall tasks. Being able to run and evaluate high-recall systems on these applications is crucial for furthering the field.

In this thesis, we focus on the evaluation of high-recall retrieval systems and the design decisions that can affect the results of such evaluation. The main contribution of this thesis is (1) a framework for replicable high-recall retrieval experimentation that was validated through its use in the TREC 2015 Total Recall track and (2) the corresponding evaluation of the systems submitted to the track, which is focused on mitigating potential sources of data leakage. Furthermore, this thesis explores the high-recall evaluation space, both at large and in the TREC context, with regards to choosing authoritative assessors, the utility of a variety of effectiveness measures, and the suitability of corpora to high-recall retrieval experiments. Finally, we present an investigation into the changes that could occur in assessing behaviour when using high-recall retrieval protocols.

1.1 Motivation and Context

The primary motivating context of this thesis is electronic discovery. The (electronic) discovery process in cases of civil litigation, such as a class-action lawsuit, allows the plaintiff ("the requesting party") the ability to require the defendant ("the responding party") to produce all documents (e.g., hard copy files, email, faxes) that are relevant to a particular "request for production."

As an example, let us consider the TREC 2009 Legal track's mock Complaint J.[1] In this mock complaint, the plaintiff (Jonas Grumby) is lodging class-action complaint against Volteron Corp. and against Jane and John Doe as officers of Volteron Corporation. This class-action complaint alleges that Volteron Corp. and its executive officers (Jane Doe and John Doe) intentionally committed fraud with respect to its shareholders on a number of issues. One of these issues related to the concealment of a lawsuit regarding an illegal football gambling ring that was run within Volteron Corporation. The request for production for this issue states that the responding party should produce "[a]ll documents or communications that describe, discuss, refer to, report on, or relate to fantasy football, gambling on football, and related activities, including but not limited to, football teams, football players, football games, football statistics, and football performance." It is worth noting that the documents to be returned are those that meet the above requirement and are "within the sole or joint possession, custody or control of the Defendants, including their agents, departments, attorneys, directors, officers, employees, consultants, investigators, insurance companies, or other persons subject to Defendants custody or control."

In the context of the United States legal system, the United States Federal Rules of Civil Procedure (FRCP) [111] govern the extent to which a request for production is deemed to be adequate and sufficient. Similar legislation exists in Canada, the United Kingdom (where it is called "disclosure"), the European Union, and Australia, and is currently being developed across Asia [58, 66, 67]. The American legal system is generally considered to be the most developed in regards to electronic discovery, so we base much of this thesis on the mandates of the Federal Rules of Civil Procedure.

Accordingly, in the American system, the defendants in our above example are required by FRCP Rule 26(g)(1) to ensure that "an attorney or party certifies that to the best of the person's knowledge, information and belief formatted after a reasonable inquiry" that the production is "not interposed for any improper purpose, such as to harass, cause unnecessary delay, or needlessly increase the cost of litigation" and is "neither unreasonable nor unduly burdensome or expensive, considering the needs of the case, prior discovery in

[1] Available from `http://plg.uwaterloo.ca/~gvcormac/treclegal/LT09_Complaint_J_final.pdf`.

the case, the amount in controversy, and the importance of the issues at stake in the action." These restrictions mean that it would be in violation of these rules to knowingly choose cost-ineffective methods for the discovery process or to produce more material than one could reasonably believe is relevant (e.g., returning the entire document collection). Additional parts of Rule 26, in particular 26(b)(2)(C)(iii), give the court the ability to limit discovery when it determines that additional retrieval of relevant information would be unduly burdensome with respect to the needs of the case. Rule 37(a)(4) requires "that an evasive or incomplete disclosure, answer or response must be treated as a failure to disclose, answer, or respond."

In essence, it is required that an eDiscovery production (the documents retrieved) must be as complete as possible but also reasonable and proportional to the needs of the case in terms of monetary and temporal costs. While such a description is reasonable for interpersonal discussion, it is difficult to quantify for scientific experimentation. To this end, we can further refine the goal of high-recall retrieval as follows: *Find as many relevant documents as practicable, using the least amount of (human) effort possible.* This definition is still not sufficiently rigorous to compare systems but it provides a basis for evaluation and encapsulates the underlying user model of high-recall retrieval.

With such a model in mind, we can see that traditional ad hoc search systems, both commercial (Google/Bing) and research (Terrier/Lucene)[2], may not immediately address the needs of high-recall tasks. Indeed, such search systems often present a snapshot of relevance with a focus on early precision (e.g., the ubiquitous ten blue links). While high-recall tasks can be accomplished through processes such as interactive ad hoc search [53], such processes may be limited by the inability of searchers to formulate enough keywords or by the belief that they have found "enough" documents when, in reality, they have not [22]. Such approaches may also require far more effort than more directed processes. Accordingly, this thesis does not focus on the high-recall retrieval tasks that are formulated around ad hoc retrieval, regardless of whether this focus was intentional or not (i.e., early TREC tasks[3]). , the focus is on tasks similar in nature to those found at later TREC tracks that placed more emphasis on "human in the loop" interaction and on systems that utilize a variety of retrieval techniques in tandem (e.g., machine learning, natural language processing, and deep learning in addition to ad hoc search. This approach also reflects current trends in various aforementioned application domains.

The TREC Legal tracks [18, 145, 109, 76, 46, 71] addressed a variety of different eDiscovery task models, ranging from boolean keyword search to interactive retrieval. On the

[2]See `http://terrier.org/` or `https://lucene.apache.org/` for more details regarding these systems.

[3]Early iterations of filtering and ad hoc tasks can be considered high-recall retrieval tasks due to their derivation from DARPA's TIPSTER project [154, Chapter 2].

other hand, the TREC Spam tracks [49, 36, 37] compared email spam filtering systems on public and private corpora by requiring a common system interface for training and classification.[4] These two particular evaluation efforts provide a basis for much of the underlying design presented in this thesis, but other public forums have investigated high-recall retrieval (e.g., the DESI workshops briefly described in Chapter 2). Recent studies have been conducted to investigate the relative benefits of using various protocols for high-recall retrieval [40, 121, 112, 163, 97, 42]. The results of these studies are not always consistent due to various differences in experimental conditions (e.g., different collection renderings, different machine learning packages, etc). An example [159] of such discrepancy comes from the blog of William Webber (a prominent eDiscovery and high-recall researcher), which describes his attempts to replicate the experiments of Cormack and Grossman [40]. Cormack, in a comment on the post, identified differences between Webber's experimental methodology and the one desribed in their publication.

It has also been suggested [75, 109, 149], in the context of the TREC Legal track, that higher interaction with a topic's "authoritative" assessor (e.g., the senior lawyer in charge), or the "Topic Authority," may have resulted in increased system performance. Whether such benefit is merely better utilization of available resources, or an attempt to "game" the experiment by seeking as much information as possible, or some combination of the two is unclear. The last iteration of the TREC Legal track [71] sought to control this interaction more precisely by limiting the types of interaction (i.e., only relevance assessing) and the amount of interaction (i.e., a fixed number of assessments). Such controlled interaction would potentially alleviate some of the variance in topic authority interaction that may have been present in previous iterations of the track.

Much of the existing information retrieval (IR) literature makes use of the Cranfield paradigm [153], which prescribes that the same task model, documents, information needs (search queries), and relevance assessments be used when conducting multiple experiments in order to test for any improvements or for any differences between algorithms. Such an approach has been seen to be reasonable for traditional ad hoc search tasks [152, 15, 146], regardless of changes in assessors or experiments involving new systems. On the other hand, there have been many studies [82, 63, 115, 86, 137, 138, 139, 119, 15, 132, 165, 10, 11] which have shown that assessing behaviour can be significantly (and often, substantially) affected by presentation order, prevalence in the documents being assessed, domain knowledge, and the instructions given to assessors to guide their judging. While these behavioural effects do not appear to have compromised the results of IR evaluation [152, 15, 146], such differences in behaviour may not be accurately reflected by "human in the loop" high-recall systems. While having a static set of assessments is preferable for an experimental set-up, due to

[4] The MIREX workshops [59, 60] also offered a similar set-up to the TREC Spam track.

associated reduction in statistical confounds, it may not necessarily reflect the fact that a given system can meaningfully affect how an assessor behaves and this change in behaviour may benefit that system over another. A single static set of assessments may hide such differences because they cannot, by definition, change.

Furthermore, many high-recall applications domains rely on some form of "topic authority" (e.g., a senior lawyer or medical doctor) to provide final determinations of document relevance. Reliance on this topic authority can lead to burdensome costs and may result in degraded system performance as fewer assessments are rendered due to the costs (time or money) associated with the authority. Several studies [112, 163, 152, 162] have suggested that this authoritative assessor can often be replaced with a "less capable" surrogate assessor with some associated (typically minor) cost. This proposal operates under the assumption that the surrogate assessor will be a cheaper alternative and result in less burdensome costs, which may not always be the case; for example, NIST assessors could be surrogates for each other but are not less costly. It has been argued [68, 5] that such replacement during the training phase of an eDiscovery retrieval effort will result in degraded performance. The potential for degradation has been reflected in the phrase: "garbage in, garbage out" [5].

If, for the moment, we sidestep the aforementioned issue and assume it has been solved or dealt with in some manner, what remains is the question of evaluation measures; namely, what does it mean to score a goal in high-recall retrieval? How do we accurately quantify how well a particular system achieves high recall? Many of the aforementioned studies dealing with high-recall retrieval have adapted existing evaluation measures, whether for historical precedence, convenience, or some other reason. But these studies typically do not investigate whether these measures are appropriate or accurate.

There are several contrasting elements when evaluating high-recall retrieval systems: they must achieve high recall while minimizing human effort; hence, maintain high precision to avoid wasting effort. Metrics such as recall, precision, and F1 alone ignore one or more of these aspects, most often the associated effort. Furthermore, measures (F1, for example) may combine incompatible units of measurement. Other common measures, such as effort at 75% recall [163, 40, 131, 70, 121], make use of arbitrary thresholds that may be unrealistic or unattainable without assessing the majority of the document collection (as prevalence may be low and/or the relevant material hard to find). The situation is further complicated when one must choose between set-based measures (e.g., recall and precision) and rank-based measures (e.g., Average Precision or AUC: see Chapter 2 for definitions of these measures). Set-based measures are convenient because they assume no ordering of the documents being evaluated, but tey lack measurements of the potential trade-offs that come from examining rank. For example, suppose two systems, A and B, produce

two result sets, x and y respectively, that achieve the same recall. However, system A ranks all the relevant documents in x first, while B ranks all the relevant documents in y last. If non-relevant documents are present in x and y, then x is obviously a preferable set of documents because a searcher would examine all the relevant documents first while attaining the same level of recall.

Accordingly, the gain curve [40, 42, 71, 122], which measure recall as a function of documents assessed, has become a de facto standard because it provides a visible trade-off between recall and effort without targetting and arbitrary threshold. This should not be construed to claim that the gain curve is the only or the ideal measure. The gain curve, while useful, does have problems of its own, the most obvious being that a full ranking of the collection is required for the full curve, which is not practicable outside of experimental settings (i.e., it cannot be used in the real world). Further, the gain curve is comprised of many values (it is not a single summarizing number), which makes condicting statisitical tests difficult and can require excessive space when reporting multiple curves. Averaging across topics may lead to skewed average curves due to different per-topic completion rates resulting from different numbers of relevant documents for each topic. Finally, gain curves do not inherently address the fundamental issue of when to stop. Reviewing an entire document collection is impractical and defeats the point of using technology-assisted review tools. Accordingly, having the means to determine at what point further effort will not yield tangible gains is a desirable outcome. While one can potentially infer from a gain curve when an algorithm *should* have stopped, actually determining when to stop and whether an algorithm stopped at the right time are other issues that have real-world practical concerns but are not addressed directly in this thesis.

1.2 Problem Statement

This thesis presents an investigation into the evaluation of high-recall retrieval systems from several different but complementary angles. First and foremost, we are concerned with the ability to conduct comparable and replicable high-recall retrieval experiments. This concern can be ameliorated by having a consistent architecture for experimentation but also by having meaningful performance metrics and consistent assessments. Additionally, we must understand the differences that such an architecture may elide or trivialize.

What follows is a presentation of our motivating research questions, how we approached them, and a brief summary of our results.

To what extent, if any, does the use of surrogate (i.e., less qualified, non-authoritative) assessors for training manifest the "garbage in, garbage out" phenomenon?

In order to determine the effect of using surrogate assessors for training, we trained an SVM classifier using passive supervised learning and 10-fold cross-validation with several sets of alternate assessments for three test collections. Primarily, we found that when alternating between these sets of assessments for the role of authority (i.e., the gold standard), the surrogate trained (i.e., non-authoritative) classifiers produce an inferior ranking — one must review more documents to achieve the same recall. Such a result is not surprising given that each set of assessments represents a different interpretation of relevance. Accordingly, we do not find that this is a manifestation of "garbage in, garbage out" as no single assessor (regardless of authority) is able to consistently train a classifier superior to the other assessors.

Moreover, we find that diversifying the interpretation of relevance (e.g., by combining pairs of surrogates) and liberalizing it (by training borderline documents as relevant) improves performance from that of individual surrogates and has the potential to rival the authority. This provides further evidence that errors in the training will not necessarily be magnified or amplified by the machine learning algorithm. If such errors were magnified, we would expect that in cases where the surrogate has more errors in judgment (e.g., false positives) that the classifier trained by such a surrogate would perform worse than one trained by the authority. Our results clearly show that this is not the case in all instances.

Can an evaluation architecture be designed that provides a common interface for high-recall retrieval evaluation that strongly controls the information available to systems? If so, can this architecture be designed for use with restricted-access collections?

To address this issue we took inspiration from past TREC tracks. Previously, the Spam track [49, 36, 37] required participants to create a series of shell scripts that would be called by a provided jig and run to simulate the spam filtering process. The benefit of such an approach was that the systems could be submitted to the track coordinators for execution on personal, private email collections. The downside was that getting such submitted systems to work often required extensive configuration and monitoring of the sandbox environment by the coordinators.[5]

With that knowledge in hand, we used modern Web technologies to create a Web service that systems communicate with to register, download document collections and topics, and request assessments of topic-document pairs from a pre-rendered and complete judgment set. In doing so, we effectively extended the TREC 2011 Legal track [71] topic authority interaction model to the entire collection without introducing any potential for inconsistency between assessments. Furthermore, this design allows submission of virtual machines that contain high-recall retrieval systems which can be run on a sandbox computer, potentially

[5]This information comes from discussions with Gordon V. Cormack, one of the Spam track coordinators.

air-walled, with test collections that cannot be distributed due to privacy or legal concerns. This architecture was validated during the TREC 2015 Total Recall track, where it was used for both over-the-Internet and sandbox modes of participation.

Out of the possible measures we might use to evaluate high-recall retrieval systems, is there one that is more able to distinguish different systems or produce more consistent estimates of system performance?

A good evaluation measure is crucial in determining whether one system is superior to another, in both relative and absolute terms. We first conducted baseline experiments to determine whether any of the metrics available to us did not meet either the differentiability or consistency requirements. To do this we compared relevance sampling and random sampling as they have been observed [40] to be substantially different from each other. The most surprising result was that in our first experiment, the performance measures we used were not able to detect a strong difference between these two strategies, which we posit is due to properties of the test collection used and the nature of the metrics used. Our second baseline experiment, conducted on a different collection, was intended to verify the first experiment's set-up and provide any insight into the behaviour on a more established test collection with different properties. In this experiment, we found that all measures very clearly distinguished random sampling from relevance sampling. Across both experiments we noted that a gain curve variant, which we call the *relative gain curve* as it measures recall as a function of $\frac{\text{effort}}{\text{R}}$ (R being the number of relevant documents) was able to tease apart more nuanced aspects of system behaviour.

Based upon these results, we examined the consistency of behaviour that the gain curve and its variants had on the "real" systems submitted to the 2015 Total Recall track. Contrary to our original hypothesis, the relative gain curve did not result in more consistent measurement of system performance than the standard gain curve. However, the relative gain curve highlighted different behavioural aspects that were not apparent in the traditional gain curve. Finally, we saw that the consistency of "real" system performance appears to depend, at least partially, on the interaction of the test collection with the task being conducted.

We also explored the use of facet-based evaluation for high-recall retrieval, where we define a facet to be any identifiable subpopulation of the document collection. While we found some differences between facet-based measures and binary recall, there does not appear to be substantial differences among systems for any of the measures we investigated. The underlying cause of the apparent similarity in evaluation is not clear but may be a result of the task model, or of the particular facets used, or of the metrics themselves, or some combination of the three. It may also simply be the case, as observed by Cormack and

Grossman [42], that many high-recall retrieval approaches tend to retrieve facets relatively equally throughout their execution.

Do different high-recall retrieval strategies (e.g., random, uncertainty [95, 94], and relevance sampling [40]) affect assessing behaviour?

Random, uncertainty, and relevance sampling form what might be thought of as the triumvirate of prototypical high-recall retrieval methodologies. To examine the effect that each has on assessing behaviour, we trained each on existing archival assessments and then had real humans judge the same set of documents interspersed with control documents. While such a protocol is not necessarily reflective of real-world systems, it provides a controlled experiment that reduces per-assessor variance because the assessors could not affect the documents that will appear next. We found that the "true" document relevance and the sampling strategy utilized can significantly and substantially affect the probability that an assessor will make a 'relevant' judgment. Interestingly, neither document relevance nor sampling strategy appears to substantially affect the time it takes to make a judgment. Though we do note, as in previous studies [139], that assessors take longer to make assessing errors than correct assessments.

1.3 Thesis Organization and Contributions

This section provides a brief overview of the thesis organization and the contributions of each chapter. For each of the main chapters in particular, we provide a summary of the work in that chapter and the contributions contained therein. Chapter 2 presents a review of the literature related to high-recall retrieval implementation and evaluation, in both historical and contemporary contexts, including a more formalized description of the eDiscovery technology-assisted review (TAR) problem, a set of definitions for terms used throughout this thesis, and brief descriptions of related TREC tracks. Chapters 3 through 6 form the main body of this thesis, their organization and contributions are described below. Chapter 7 describes several potential avenues for future research based upon the work described in chapters of the main body. This thesis concludes with Chapter 8, which gives a final summary of the work as a whole and a reiteration of the contributions of this thesis.

1.3.1 Chapter 3: Impact of Surrogate Assessments on High-Recall Retrieval

This chapter focuses on exploring the "garbage in, garbage out" belief that present in the eDiscovery application domain. This belief posits that training errors due to surrogate (i.e., non-authoritative) assessors will be magnified by machine learning techniques and yield inferior performance compared to the same techniques trained by the authority (i.e., a senior lawyer). To test our experimental set-up, we first replicated the earlier results of Webber and Pickens [163] which found that while one could substitute one assessor with another it may result in some loss in effectiveness.. We then trained a Support Vector Machine (SVM) using passive supervised learning and 10-fold cross-validation on three existing TREC test collections, two ad hoc collections and one Legal track collection, selecting relevance assessments for use during training and evaluation from different assessors.

The general trend was that whichever surrogate was deemed to be authoritative produced a classifier that better reflected the authoritative conception of relevance (i.e., the resulting ranked list from the authoritative classifier required less review effort than any of the surrogates to achieve the same recall). In a single case a single surrogate was better than the authority but this may be a side effect of the nature of the surrogate assessor's judging process rather than a deficiency on part of the authoritative assessor. Furthermore, we explored several techniques for diversifying surrogate assessments, by splitting the training task in half and assigning each half to a surrogate assessor, or by using the assessments of one surrogate when a second surrogate had not judged a document. These pseudo-surrogates require the same amount of judging effort but contain a more diverse conception of relevance as they combine two individual conceptions. The performance of this diversification process generally resulted in a classifier that was as good as, and sometimes better than, the better-performing constituent surrogate. An alternate approach is to liberalize the conception of relevance by taking the union of two surrogates (i.e., a document is relevant if either surrogate deems it so) or by assessing borderline relevant documents as 'relevant.' In both cases, the resultant liberal pseudo-surrogate produces a classifier that is substantially better than either constituent surrogate and, in some cases, rivals the authority-trained classifier.

These results lead us to conclude that what is being observed in our experiments is not "garbage in, garbage out" but a difference of opinion. Tefko Saracevic, in a recent article [130] on the ongoing centrality of relevance to information science, stated that "relevance is a human, not a technical, notion." With this in mind, we should not be surprised that, despite the best of efforts, no two humans are able to conceive of exactly the same conception of relevance. Since the authority is subject to this same limitation, it is

not the case that the authoritative assessor is preternaturally better at assessing relevance. It is simply the case that this particular assessor was chosen to be the final arbiter of truth.

1.3.2 Chapter 4: TREC 2015 Total Recall Track

This chapter discusses the first iteration of the Total Recall track at TREC 2015 in its entirety. One of the goals of the track, in addition to exploring the evaluation of high-recall retrieval systems and tasks, was to develop an architecture that facilitates repeatable experimentation while attempting to mitigate any potential leakage of sensitive information. To this end, the chapter contains the implementation details and design rationale of the Web-based infrastructure that facilitated such experiments. Included is a discussion of how the API's designed, which facilitates the submission of the virtual machines containing retrieval systems in order to allow evaluation using private data, and the design decisions regarding automatic system registration, the distribution of document collections via HTTP (rather than, say, sneakernet), and pre-rendering assessments for all topics and all documents in the collection. Accordingly, we report on document collections formatting, topic creation, and document assessment process at a level of detail that has not yet been produced elsewhere.

We go on to validate this experimental set-up by running actual TREC participant systems on this architecture, both over the Web and in a sandbox environment on sensitive data. We present a brief high-level overview of each of these systems and discuss the evaluation measures used in the track, including a new measure: recall @ $aR + B$ effort, where R is the number of relevant documents, and a and b are free variables representing permissiveness of non-relevant material and a fixed effort overhead respectively. This new measure attempts to summarize recall along the gain curve at equivalent points across all topics. Recall @ $aR + b$ also forms the inspiration for the relative gain curves discussed in Chapter 5. We discuss the results of these systems and find that, by and large, no system is consistently superior to the baseline. However, the baseline is not consistently superior to the best of the participant systems either.

Additionally, we labelled one of the test collections according to all possible combinations of sender and recipient. This was done in an attempt to determine whether any facet-based high-recall measure is able to distinguish between these systems. The results of the facet-based evaluation are similar to those of the general evaluation. While these results potentially provide further evidence that the depth-first approach of the baseline [42] may be a winning strategy, our observed results may stem from the general nature of the facets (i.e., that they are not topic specific) and the effectiveness measures used.

1.3.3 Chapter 5: An Exploration of Effectiveness Measures for High-Recall Retrieval

This chapter explores the high-recall evaluation space and the suitability of various effectiveness measures with respect to system consistency and ability to distinguish systems. We begin by providing an overview of existing evaluation measures and propose one new one; the relative gain curve, which measures recall as a function of $\frac{effort}{R}$ (where R represents the number of relevant documents). The intent of such a measure is to provide a more meaningful measurement of mean recall, computed at comparable points across topics. Furthermore, we propose the use of root-mean square error as a mechanism for determining the consistency of systems when averaging recall over absolute or relative effort.

We began our investigation by comparing random and relevance sampling strategies on the Reuters Collection Volume 1 (RCV1). It became quite apparent that there was a high degree of variance among system performance on a per-topic basis. Indeed, we were unable to distinguish relevance and random sampling a variety of measures even after bootstrap sampling. We posit that this is an artifact of the corpus-specific properties; namely, the overall high prevalence of topics but also the high variance of prevalence among topics (i.e., thousands to hundreds of thousands of relevant documents). This leads us to suggest that RCV1 may not be a suitable collection for *general purpose* high-recall experimentation. To confirm this, we re-ran the experiment on the TREC-6 ad hoc test collection and achieved results that are in line with the literature (i.e., that relevance sampling is substantially and significantly better) across all of the measures examined. While such a result validates our methodology, it does little to provide insight regarding which measures are more useful for evaluation. However, we did notice a trend by the relative gain curve to account for more of the per-topic variances in behaviour that had been hidden by the standard gain curve.

Based upon these results, we examined the results of the 2015 Total Recall track with respect to curve-based measures. The general outcome was much the same as our observations in Chapter 4: no system is able to beat the baseline across all topics and collections. While we found differences in the behaviour reported by the gain and relative gain curves, it appears that systems do not achieve similar levels of recall (consistency is low) regardless of the gain curve used. This counterexample to the original purpose of the relative gain curve might indicate a reason to not use the measure, but, due to the observed behavioural differences, we conclude that the relative gain curve may still be a useful metric. We found additional evidence that test collection and retrieval task (topical retrieval versus information governance) may affect how consistent a system is with respect to several of the measures used. This was clearly evident for private collections used in the Total Recall

track as systems were noticeably more consistent than on the public collections.

1.3.4 Chapter 6: Effects of High-Recall Retrieval Protocols on Assessing Behaviour

This chapter presents a user study that was conducted to examine the effects that various sampling strategies would have on assessing behaviour in high-recall retrieval scenarios. We began by using the TREC-6 University of Waterloo assessments to train three classifiers, using random, uncertainty [95, 94], and relevance sampling in the style of Cormack and Grossman [40]. To better measure assessing behaviour, we inserted control documents into the induced rankings, at fixed positions, so that we did not have to acquire an exorbitant number of judgments from the user-study participants to better account for differences among algorithms (random sampling would not necessarily produce sufficient numbers of relevant documents).

We then explored how the underlying context (i.e., sampling strategy) and underlying relevance of documents affected the assessing behaviour. Primarily we found that context and relevance can significantly affect the probability that a user will make a "relevant" assessment. In particular, relevance sampling will result in a lower probability than uncertainty and random sampling. Interestingly, we did not notice any difference between the strategies in terms of the time taken by an assessor to render a judgment. Also provided in CHapter 6 is an analysis of the results when we look at a more nuanced breakdown of user behaviour based upon all combinations of NIST and Waterloo relevance classes (i.e., how did users behave for documents that Waterloo judged as relevant and NIST judged as not relevant?). Based upon this analysis, we observed behaviour consistent with Smucker and Jethani's earlier result [139] of users taking longer to make incorrect judgments. Unlike Smucker and Jethani, we found that our users spent more time making false negative assessments rather than false positives. The chapter ends with a discussion of the implications of these results with an emphasis on being particularly aware of how assessment pools are formed, how the documents are judged, and how both of these factors can affect the results of an evaluation.

Chapter 2

Related Work

In this section, we explore the background of high-recall retrieval primarily as it relates to electronic discovery. We begin with a presentation of various approaches to high-recall retrieval, evaluation of such systems, and finally other related work that explores high-recall retrieval in various ways that are not directly concerned with the aforementioned topics. We also provide descriptions of relevant TREC tracks and a brief overview of the DESI workshops.

2.1 The TAR Problem

Technology-assisted review (TAR) is the process of conducting an eDiscovery review using computing technology. Typically, TAR utilizes human assessments (of a small subset of the document collection) to extrapolate relevance for the remainder of the collection [70, 40, 39]. It is worth noting that at this point, the underlying technology used is not rigorously prescribed by law or any other authority. Accordingly, approaches that utilize random sampling, active learning, interactive search and judging, or other statistical approaches are all valid. However, we should keep in mind that the goal of an eDiscovery review is to identify as many relevant documents as possible while ensuring reasonable costs, in terms of time and money, (see Section 1.1 for a more thorough discussion of the eDiscovery context), and some approaches may be more amenable to these goals than others [40, 44, 41].

Given the focus of this thesis and the predominance of machine learning in TAR solutions,[1] we should contrast TAR with machine learning as it is used in abductive text

[1] As above, it is worth noting that machine learning is not a necessary component in TAR. It has merely

categorization [136]. In text categorization, the document collection is typically considered to be a representative sample of an (hypothetical) infinite population of documents with similar characteristics to that sample; in TAR, the document collection is fixed but may be incrementally created.[2] It is worth noting that the machine learning aspects of TAR could be modelled in the style of transductive machine learning [88].

In text categorization, the machine learning algorithm may have prior knowledge derived from existing information sources (e.g., examples of relevant documents from a different sample of documents) or may be able to use an existing pre-labelled training subset of the collection; in TAR, there is typically no prior information about the data collection to be leveraged. Finally, the goal of text categorization is to produce the best possible classifier for the (hypothetical) infinite document collection that can be learned from the finite sample available. The goal of TAR is simply to identify as many relevant documents as possible while minimizing human effort, which in turn minimizes temporal and monetary costs.

2.2 High-Recall Retrieval Systems

2.2.1 Commercial Software

Hogan, Bauer and Brassil [80], from the commercial vendor, H5, discuss the methodology behind their technology-assisted review platform, provide insight into how their system works, and discuss some empirical validation performed at the TREC Legal track and elsewhere [19, 23, 81], which is discussed in more detail below. They highlight that their system relies on modelling the authoritative conception of relevance by trying to nail down the concepts related to relevance, how nuanced those concepts are, and the vocabulary that might be used in a relevant document.

Bauer et al. present a framework for high-recall retrieval that utilizes two components, a proxy and an assessor, to match a user's information need with a machine learner [19]. In essence, the proxy creates a user model by extensively refining the user's definition of relevance; while the assessor translates this model into something that the underlying machine learner can use to yield high precision ($> 80\%$) and recall ($> 75\%$). They present the results of such a system in the TREC Legal track, where it did yield high precision and

been observed to be an effective one.

[2] This incremental creation can arise in situations where parts of the document collection are incrementally made available due to processing bottlenecks or issues regarding access to the data.

high recall simultaneously. A follow-on paper by Brassil et al. discusses the user-modelling process employed in the aforementioned framework [23]. Their process focuses on aspects such as: the information need (what is being sought); scope (the breadth of concepts that are relevant); nuance; and linguistic variability (e.g., what words denote relevance). Further discussion of this process is presented by Hogan, Bauer, and Brassil, using a system called STIR [80, 20]. STIR utilizes these aspects to accomplish what the authors call "legal sense making," which effectively amounts to determining the criteria for relevance. Emphasis is placed on the use of a senior litigator to judge exemplar documents which are used by a multi-disciplinary team to generate queries iteratively for use in search and judging. In the TREC Legal track [109, 76], the H5 approach, discussed above, substantially outperformed other participants conducting experiments on the same topics.

In an oft-cited study regarding keyword search effectiveness, Blair and Maron report that lawyers using the IBM STAIRS software were only able to achieve 20% recall even though they believed themselves to have achieved at least 75% [22]. In this study, two lawyers created 51 information needs, from a single, real request for production, that two paralegals would use to attempt to find relevant documents in the 40,000 document collection. The process was iterative in the sense that the paralegals would formulate queries, retrieve a variably sized batch of documents, and the lawyers would review the retrieved documents. If the lawyers believed that they had sufficient evidence, which they themselves deemed to be 75% of recall, then the process would stop. If the lawyers were not, the paralegals would be given written instructions on how to improve the retrieval effort. The actual searchers, the paralegals, were familiar with the STAIRS system and had access to its full capabilities, which included proximity search (e.g., two terms should be in the same paragraph) and minor ranking capability by date, subject, author, and so forth. It is worth noting that the lawyers and paralegals did not have any mechanism to measure their performance other than their intuition. Furthermore, they had no way to determine whether they were missing any pertinent information; namely, they had no mechanism to sample the unretrieved documents in any fashion. Indeed, Blair and Maron report going to great lengths outside of just using a search system to find the relevant documents that were missed by the legal team, which led them to conclude that they had only created an upper-bound on recall as they had still likely missed relevant documents. The resulting disparity of what was achieved and what was believed to have been achieved has subsequently led to many doubts about the effectiveness of retrieval systems in the legal community. Though it is worth stressing that the lawyers involved were encouraged to continue *until* they believed they could *defend* the case, they were provided no tools to verify their belief other than the results of search sessions. In summary, Blair and Maron concluded that the high costs of full-text search (storage and creation) and poor

performance over ˜6 months indicated that manually crafted search indexes, which map terms to documents and are much smaller, were preferable because they would ensure higher-quality search results and a noticeable decrease in costs..

This issue of cost, among others, is brought up by Regard and Matzen who re-examine the Blair and Maron study in a modern context [116] and argue that the cost aspect at the time (1985) reflected the lack of high-quality automated technology to transcribe physical documents into digital formats. The more important issues that Regard and Matzen address include the following: there was no way for searchers to retrieve false negatives, they could only see false positives and, therefore, could only really improve precision; only simple boolean search was actually used by the searchers as opposed to the more complex boolean operators available to them; and the collection was not necessarily representative of the task being replicated since it resulted from a lawsuit and was likely culled of much extraneous information.

Sormunen [141], in a follow-up extension to Blair and Maron, investigated the effect of exact boolean match on three databases of different sizes and prevalence. The study began with an extensive derivation of all possible subtopics for each topic and of corresponding search terms for each subtopic. Queries were then formulated to allow variation in subtopic extent (more subtopic terms used) and subtopic exhaustiveness (more subtopics represented in the query). Sormunen found that regardless of database size and prevalence there exists a consistent drop in precision when achieving high recall ($> 80\%$) and that when the database is large and has higher prevalence this drop is much more dramatic when going from 90% to 100% recall, which Sormunen posits is due to unique and otherwise hard to find documents. However, both large databases, regardless of prevalence, had similar precision at 100% recall, indicating that those last few relevant documents will always be difficult to retrieve. It was possible to complete 94% of the 18 topics examined using only query terms for a single subtopic. However, when query exhaustiveness was increased (i.e., more subtopics were covered) performance decreased due to the presence of conjunctions, which agrees with observations made in the original Blair and Maron study. When combined with the fact that a third of the highly relevant documents (and more in the case of less relevant documents), do not contain a shared query term for all subtopics, more than one (subtopic) query is likely needed, Sormunen concludes to find all the relevant documents. This is limited, he argues, by the fact that a user would be hard pressed to figure out what these queries should be. In light of these conclusions and of the great lengths that Blair and Maron report [22] going to in order to find relevant documents to evaluate their study, the performance of Blair and Maron's legal searcher is not surprising and doing better would have likely required a great deal more effort than had already been expended.

2.2.2 Manual Review

Roitblat, Kershaw, and Oot set out to investigate whether technology-assisted review (TAR) can outperform exhaustive manual review [124]. Using an actual U.S. Department of Justice request, they collected additional manual assessments from two professional review teams, A and B, from a sample of the document collection and compared them to the original assessments created for the case. Agreement was within what previous studies have reported [152]; specifically, team A agreed with the original assessments 75.6% of the time, team B agreed with the original 72.0% of the time, and they agreed with each other 70.3% of the time. Both teams, identified sizable ($> 20\%$) proportions of the non-relevant documents to be relevant in their review, and identified approximately half, only, of the relevant documents as relevant. Two commercial TAR systems, C and D, were trained on the adjudicated assessments of teams A and B but classified the entire document collection. While both systems agreed with the original assessments more than either manual team (approximately 83% for both systems), they separately only identified 45.8% and 52.7% of the original relevant documents, and, cumulatively, identified 72.1% of the relevant documents. Out of the documents identified, by either system, as relevant, the original assessments only found 25.1% of them to be relevant. The authors found that the TAR systems were able to achieve comparable recall to the manual re-reviews, (0.518,0.527) versus (0.488,0.539), but with higher precision, (0.271,0.294) versus (0.197,0.183). Based upon these results and further analyses that accounted for different agreement rates among relevant and non-relevant documents, the authors conclude that the commercial TAR systems are at least as capable as the manual review teams.

Grossman and Cormack investigated whether technology-assisted review (TAR) is superior to exhaustive manual review through a detailed comparison of the judgments rendered by two participant systems for 5 of the TREC 2009 Legal interactive task topics. One team, the University of Waterloo, used a machine learning-based approach for 4 of these topics, while the other team, H5 (a commercial TAR vendor), used their in-house process for the other topic. The author show that across the 5 topics, technology-assisted runs had significantly superior precision and F1 when compared to the manually rendered assessments; while the difference in recall, which favoured TAR, was not significant. On the other hand, they report that on average the TREC participants reviewed only 1.9% of the collection on average, which is a fifty-fold savings compared to exhaustive manual review. From these five review efforts, a maximum (minimum) of 4.1% (0.5%) of the Enron v1 document collection (described in Section 2.5.1) was reviewed for any one topic. For each topic, about half of the documents reviewed, produced by Waterloo or H5, were relevant to the topic.

Cormack and Grossman also highlight that as part of the TREC 2009 Legal interactive task, there existed two phases of assessment: an initial phase conducted on a statistical sample of the document collection by volunteer lawyers and law students, and an appeals phase where participants could ask for documents they believed to be incorrectly labelled to be adjudicated by a senior lawyer, the topic authority. The participants achieved recall that ranged from 67.3% to 86.5% and precision which ranged from 69.2% to 91.2% with respect to the adjudicated assessments. The initial assessments, however, achieved recall ranging from 25.2% to 79.9% and precision ranging from 5.0% to 89.0%. While the initial assessments were conducted on only a sample of the document collection, they are ostensibly a stand-in for an exhaustive manual review. Accordingly, the aforementioned performance results led the authors to conclude that even though there may not be significant differences in the recall achieved by TAR systems compared to recall by human reviewers,[3] there are significant differences in precision, and thus effort, in the two approaches to reviewing those documents.

William Webber conducted his own analysis [157] of the TREC 2009 Legal interactive task data in light of the the Grossman and Cormack study [69] described above. Webber argues that a major flaw in the Grossman and Cormack study is that they did not account for poor reliability or quality control in the initial assessments and that the appeals process may have been biased towards teams with a higher proclivity to appealing. Webber highlights the fact that for any topic, an individual assessor would typically not assess more than 500 documents, so there may have been many competing conceptions of relevance in the initial assessment of topic. To identify low-quality assessments, he proposed using the proportion of documents assessed relevant as an indicator variable so that when an assessor who greatly over- or underperforms their colleagues may be considered unreliable. Using this idea, Webber examined the effect that the appeals had on these proportions per assessor and found that, for the five topics used by Grossman and Cormack, the appeals process appears to have been "reasonably complete."

In spite of this, Webber goes on to argue that comparing the manual and automatic reviews based upon the extrapolated evaluation some errors made by the initial assessors more substantially affected their performance than would be the case if the assessments had been taken at face value (i.e., not extrapolated).[4] meant that When compared using

[3]We note that on average the authors found approximately a 30% difference in recall between the two groups but did not find statistical significance.

[4]The TREC 2009 Legal interactive assessments were rendered on a stratified sample, such that a single sample document represented some fraction of the entire collection. The size of that fraction depends on its corresponding stratum's sampling rate—a smaller rate indicates the selected documents represents more documents. Accordingly, misidentifying a relevant document sampled from larger strata would result in

non-extrapolated recall and precision, the manual teams appear to have performed more competitively with the automatic teams but were not clearly superior.

Webber then presents a discussion of how to mitigate the unreliability of assessors by excluding assessors who have relevance rates that are noticeably different from those of the other assessors and by assuming that their results could be fixed by additional review of their judgments or by redistributing their assessments to other reliable assessors. The performance of the manual teams was recomputed with these unreliable assessors removed, and, unsurprisingly, (extrapolated) performance improved across the board with the exception of recall for one topic which showed a decrease. This decrease was due to one of the omitted assessors finding a great deal of relevant documents, which Webber posits may have been due to a widely cast net. Ultimately, Webber concludes that the best manual review teams are as reliable, if not more so, than automatic systems.

Throughout his analysis Webber ignores a crucial aspect of the eDiscovery problem: cost. Extrapolated evaluation means that any single assessor is judging, in the experimental setting, a far greater proportion of the corpus than 500 documents. While the best review teams may be more reliable than automatic teams, they are also much more costly than the automatic solution. Furthermore, exclusion of unreliable assessors dramatically increases review cost, since work has to be repeated. Webber's main concession to cost is that exhaustive manual review can be limited by undertaking keyword culling of the corpus o by some other custodial cleaning, which may itself remove (important) relevant documents, as highlighted by Sormunen's work [141].

Barnett and Godjevac explored the idea of the infallibility of manual review by examining the results of a pilot study conducted to select a professional review team for employment [16]. Five professional document review teams and two teams of outside counsel (i.e., trial lawyers) were contracted to label a collection of 28,000 documents (mostly Microsoft Outlook files and associated attachments) for a variety of tasks and were provided with extensive training and guidelines to do so. Documents were labelled as belonging to a single document family (e.g., email threads or parent emails for attachments belonged to the same family) such that any document in a family labelled as relevant made the entire document family relevant. Across these 7 groups, "relevant" assessment rates ranged from 23.08% to 54.16% of the collection, indicating (from the outset) that even professional reviewers provided with extensive training may disagree on relevance.

In their comparison, Barnett and Godjevac computed the overlap of pairwise assessments as the fraction of document families for which two teams agreed on relevance to the

a greater hit to recall, as the misidentified document represents a larger portion of not retrieved relevant documents.

sum of the document families reviewed by those two teams. This results in overlap values ranging from 65% to 85%, which agrees with prior work [155, 124]. They report that their results disagree with earlier results from the Voorhees study [152], discussed later in this chapter, by upwards of 30% (Voorhees reports a maximum pairwise overlap of 49.4%), but they fail to account for the fact that Voorhees calculated overlap solely with respect to the relevant assessments as opposed to both relevant and non-relevance assessments. That being said, even lawyers appear to disagree on the relevance of documents at least 25% of the time, which is evidence that even lawyers are not infallible relevance assessors.

The results of their analyses led Barnett and Godjevac to conclude that TAR is most useful for making the "easy" decisions on clearly relevant/not relevant documents and that human effort should be saved for those documents where the boundary is not so clear. Furthermore, they argue that more active quality control, including feedback from the senior lawyer about errors and sampling to identify false negatives, is necessary to more efficiently make use of human effort.

2.2.3 Automated Selection of Documents for Review

Cormack and Grossman studied three archetypal TAR protocols to determine which is able to achieve the highest recall with least effort [40]. Figure 2.1 presents a high-level look at the framework Cormack and Grossman used in their study. The general algorithm begins by selecting a set of N documents in some manner (judgmental or random sampling) and having those documents assessed (line 1). The machine learning algorithm is then trained on these initial assessments and proceeds to the classification loop (lines 2 and 3). The entire collection is classified by the machine learning algorithm, and the top K documents are selected, via some predetermined strategy, for additional assessment (lines 4 through 6). The training set is augmented by these new documents and the classifier is retrained on all of the assessed documents (lines 8 and 9). The process continue until some pre-defined stopping criterion is met (line 10). For their study, N and K were set to be 1,000.

Accordingly, the algorithm can be varied by changing lines (1),(5 and 6), and (10). In addition, the choice of classifier and features used could be varied but were held constant for this study (a Support Vector Machine and overlapping byte 4-grams). Cormack and Grossman primarily varied how line (5) was conducted, did not vary line (1), and used an oracle for line (10). Line (5) used the following techniques:

- *simple passive learning* (SPL): Select K unassessed documents randomly or judgmentally.

- *simple active learning* (SAL): Select K unassessed documents about which the classifier is most uncertain (an implementation of Lewis and Gale's uncertainty sampling [95, 94]).
- *continuous active learning* (CAL): Select K unassessed documents the classifier views to be the most likely to be relevant (e.g., extreme relevance feedback or relevance sampling).

CAL is different from the other methods in that it continues until the entire corpus is judged, or, more realistically, until it has found "enough" relevant documents, where "enough" is currently an active subject of investigation, as discussed in Section 2.3.3. In contrast, SPL and SAL require an a priori training-stopping criterion; that is, these methods require a mechanism that determines when the training set is "adequate." To this end, Cormack and Grossman chose several potential stopping points for SPL and SAL to provide an overview of possible performance. They also report results for an ideal stopping point (determined post hoc) for each method. Using 4 TREC 2009 Legal interactive task topics and four actual legal cases, Cormack and Grossman showed that CAL is able to find more relevant documents with substantially less effort than SPL. SAL is more competitive than SPL but still inferior to CAL for the majority of topics and never achieves higher recall with less effort. The authors argue that one of the chief benefits of CAL is that it targets reviewer effort on legally important documents rather than on random documents or documents that would produce a better general purpose classifier, which is the goal of uncertainty sampling. It is worth noting that Lewis and Gale [94] found that when prevalence is low, relevance sampling (CAL) approaches tend to achieve better performance than uncertainty sampling.

In a follow-up study, Cormack and Grossman proposed AutoTAR [41] which extends the CAL protocol by using an exponentially increasing batch size rather than a fixed batch size. This dynamic batching substantially outperformed the standard CAL protocol, due to the fact that the smaller batches initially allow AutoTAR to quickly identify characteristics of relevant documents without first having to go through several large batches of documents that may not accurately portray those characteristics. The authors also showed that the initial seed set of documents can be replaced by using the information need (e.g., query, request for production) as a relevant pseudo-document and a random sample of the document collection trained as not relevant.[5] This approach effectively reduces the cost of line (1) in Figure 2.1 to zero, since no documents need to be judged nor does some a

[5]The underlying assumption of the presumptively labelled random sample is that corpus prevalence is likely to be sufficiently low. Accordingly, this process will not result in many, if any, false negatives which could negatively affect the classifier.

```
seed <- select N initial seed documents and assess them
train classifier on seed
do
  classified <- classify corpus
  step <- select K documents from classified that are not in seed
          and assess them
  seed <- seed + step
  train classifier on seed
until done
```

Figure 2.1: High-level pseudocode for the TAR framework used in the Cormack and Grossman study [40]f on optimal approaches to TAR.

priori set of training documents need to be selected. This relatively simple adjustment to the initial training regiment does not appear to hamper the effectiveness of the underlying CAL and AutoTAR algorithms.

Cormack and Grossman further extended the CAL and AutoTAR protocols with a new variant called, Scalable CAL (S-CAL) [45]. As previously mentioned, the CAL algorithm operates by continuing its review until "enough" relevant material is found which may only be viable in low-prevalence but high-stakes domains such as eDiscovery. S-CAL attempts to limit labelling effort by creating a classifier that is comparable to the results of running CAL or AutoTAR on the entire collection. Effectively, S-CAL runs an AutoTAR-like algorithm but initially uses only a sample of the document collection. Additionally, for each batch of documents to be trained on, a subsample is selected rather than the entire batch (though the entire batch of documents is removed from the sampled collection—much as in normal CAL). Once S-CAL exhausts the sampled document collection, a final classifier can be trained and used to classify the entire document collection for review. At its core, S-CAL encapsulates the idea of uncertainty sampling, "produce the best classifier possible,"[6] but maintains the benefits of CAL (reviewing relevant-looking documents) while also ensuring that the review effort is minimized (with respect to both training and final review). Furthermore, this subsampling technique forms a stratified sample of the collection which can be used to estimate prevalence, recall, and precision of the review and determine whether additional rounds of labelling are necessary.

The CAL protocol can be likened to a miner: it sees a vein of relevance in some set

[6]This is the case because the entire document collection needs to be ranked for review not just the initial sample. Accordingly, producing a classifier tailored to the sample would yield inferior results overall.

of documents, mines it to completion, and moves on to a different vein. The underlying assumption is that each of these veins has a (tenuous) link to each other that allows CAL to make progress once one vein is down. It can be seen as an approximate depth-first search of the document collection. Using two test collections (TREC Legal 2009 interactive task and RCV1) and facets generated for each collection, Cormack and Grossman[42] investigated how well CAL is able to find these facets when trained only on binary relevance (i.e., topic-level relevance). While CAL may initially prioritize finding some easier-to-find facets over others, Cormack and Grossman's results indicate that when high levels of recall are achieved, equivalent levels are achieved for the majority of facets. They further report that only when the most likely relevant from all facets are found is there a precipitous drop in the marginal precision of one CAL batch to the next.

Li et al. present a TAR system they call ReQ-ReC (ReQuery-ReClassify) [97], which would largely fall under the CAL protocol of Cormack and Grossman [40] though it does mix in aspects of uncertainty sampling. Primarily, their system uses relevance feedback via Rocchio's method [118] to perform query expansion to expand the document pool. A classifier is then used to select documents for review. The authors found that when there are few judgments, Rocchio's method performs well. With more judgments, the classifier-based active learning performs better with respect to R-Precision. The recall achieved by the system was not reported nor were any electronic discovery or large high-recall-oriented test collections used, so we cannot compare the effectiveness of this technique to any others.

Relevance feedback has also been investigated by Lubell-Doughtie and Hofmann [101], who combined it with query-specific learning to rank in the context of TREC 2010 Legal track. Rocchio-style query expansion is used to create multiple expanded queries of different sizes. The ad hoc search scores of each document corresponding to these expanded queries were used as features for their learning-to-rank algorithm so that the document collection could be reordered to minimized misclassification of judged documents. They found that this method is superior to standard Rocchio-style query expansion with respect to MAP [27, Chapter 2] and nDCG [27, Chapter 12]. When they re-ran this experiment on the TREC 2011 Legal track [100], they found that there was no difference between the two on AUC [71] or Hypothetical F1 [71].

It is not outside the realm of possibility that in a review effort the collection will grow over time as more documents are processed or made available. Such collections have been termed *rolling collections* [39]. Scholtes, Cann, and Mack investigated the effect rolling collections can have on SVM classification [134]. When feature engineering utilizes complex combinations of word frequency both within and between documents and is not re-run when new documents are made available, the authors found that a dramatic decrease in classification performance with respect to F1 can occur. With one of their features sets,

they report an F1 drop from 98.7% to 30.5%. The authors also report experiments on the effect that the insertion of uniform, random training label noise can have on their SVM classifier. Due to a lack of specific experimental detail and of literature in the related field of spam filtering [135, 47], it is not clear whether these results are generalizable or applicable at large.

Barnett et al. [17] examined the effect that different document labels can have when training Xerox's proprietary machine learning software, CategoriX, which is derived from Hofmann's probabilistic latent semantic analysis [79]. Five groups of reviewers independently determined the responsiveness of 10,000 documents for a particular legal matter with responsiveness rates ranging from 39% to 58% of the collection, which is similar to rates observed in other work discussed in this chapter. When one of these review groups was selected to be the gold standard for evaluation, CategoriX was able to achieve higher recall with similar precision to whichever of the other 4 review groups was used to train it.

Barnett et al. then investigated how prevalence in training data can affect the performance of CategoriX. In particular, they tested whether it was better to maintain this corpus-wide prevalence in the training data or to bias the training set such that it contained more of a balance between relevant and non-relevant documents. The authors found that when training sets are small, it is better to use the biased training sets for training than to respect class prevalence. They report a difference of approximately 0.05 in F1 at the smallest training set size, which corresponds to a substantial difference in the amount of relevant documents identified. However, this performance difference decreases as the size of the training set increases, which follows from the fact that more relevant documents are introduced as this occurs.

Cheng et al. [33] argue that binary classifications are too restrictive for eDiscovery review processes because they may increase assessor disagreement for "arguable" or "grey" documents. They propose a soft-labelling method which uses a linear combination of first- and second-pass binary reviews to determine a document's label. Two such combinations were used: symmetrical, $\min(\alpha L_1 + (1-\alpha)L_2, 1)$ and asymmetrical, $\min(\alpha L_1 + \beta L_2, 1)$, where L_1 and L_2 are binary relevance assessments rendered by two different assessors and $0 \leq \alpha, \beta \leq 1$. Using these two combination strategies, they swept the relevant parameters (α and β) using the TREC 2010 Legal Learning task's provided seed set for training. Using the initial seed set and the same set after taking into account the track's appeals process, two sets of binary labels for 8 topics on the Enron v1 corpus were created.

A logistic regression classifier was (passively) trained on these two sets of assessments for both symmetric and asymmetric combinations, with a parameter sweep being conducted on both α and β. Surprisingly, the authors found that for 6 of the 8 topics the pre-appeals

(the initial) seed set alone trains a better classifier (according to MAP) than does the gold standard assessment, which the authors hypothesized was due to the inclusion of borderline documents that were changed from relevant to not relevant during the evaluation process. A setting of $\alpha = 0.75$ for symmetric weighting achieved the best MAP performance, and setting $\alpha = 0.75$ and $\beta = 0.9$ achieved the best MAP for asymmetric. While, a Wilcoxon signed-rank test revealed that at these settings the asymmetric achieves significantly better performance, the difference is not substantial on average (< 0.01 MAP difference). Regardless, these results indicate that solely relying on a single source of relevance assessment may not be ideal and that being overly conservative (treating borderline documents as not relevant) may not be beneficial when training a machine learning algorithm.

2.3 Evaluating High-Recall Retrieval

There have been two large-scale summaries [108, 110] of high-recall retrieval research, both having a focus on how the results are applicable to electronic discovery. Oard et al. [108] provide an overview of the legal context in which retrieval tasks are conducted and of related concepts in IR (e.g., defining relevance, evaluation). The bulk of their article is a summary of the TREC Legal track up until 2010 and is much more detailed than the summary provided earlier in this chapter. They conclude with potential future directions the community may need to consider (e.g., standardization).

The second work, penned by Oard and Webber [110], more generally conveys how information retrieval can be applied to electronic discovery. In essence, the article by Oard and Webber is a primer to help legal professionals understand how information retrieval works and, more importantly, how it can work for them. Included in the article are descriptions of evaluations measures, sampling strategies, methods to estimate recall and precision, a summary of experimental evaluation venues (similar to Oard et al. [108]), and other related topics (e.g., manual versus assisted review, keyword search). They finish with future avenues they consider need to be tread, which include: better evaluations measures, more reusable high-recall test collections, privilege review (e.g., whether a document falls under attorney-client privilege), and tools for early case assessment (e.g., for determining whether the strength of a case requires some initial retrieval effort).

2.3.1 The Cranfield Paradigm

The Cranfield Paradigm is derived from a set of experiments conducted by Cyril W. Cleverdon at the Cranfield Institute in the late 1960s to investigate the effectiveness of several

retrieval techniques [35]. In these experiments, Cleverdon used a fixed, static document collection, information needs representing a user's desired information, and relevance judgments for every document and information need pair in the collection. By using this fixed and static collection of documents, information needs, and relevance assessments, Cleverdon was able to easily recreate experiments with new techniques and recreate/repeat old results.

This controlled, laboratory style of evaluation has underpinned the vast majority of information retrieval evaluation [153, 129]. As both Voorhess [153] and Saracevic [129], as well as many others, have noted, the Cranfield Paradigm is not without its own assumptions:

- Relevance is generally considered only from a "topicality" perspective.
- Relevance is generally binary (relevant or not relevant).
- The entire collection has been completely assessed for relevance with respect to every information need.
- Relevance can and is judged for a document independent of all other documents.
- Relevance assessments are consistent (two reasonable assessors would make the same determination) and stable (they do not change over time)

While such assumptions often do not hold in practice [129, 153, 25, 126, 152], Cranfield-style experimentation and evaluation has been the predominant mechanism for furthering the field of information retrieval [153, 129]. This has been facilitated, in large part, by the use of the Cranfield Paradigm at experimental IR venues, such as the Text Retrieval Conference (TREC),[7] the Conference and Labs of the Evaluation Forum (CLEF),[8] and the NII Testbeds and Community for Information Access Research (NTCIR).[9]

The Cranfield Paradigm has formed the basis of the majority of the experiments found in this thesis. Accordingly, the remainder of this chapter discusses a variety of research that ties directly into the Cranfield Paradigm, the repercussions when the above assumptions do not hold, and how to mitigate issues arising from these assumptions.

[7] `http://trec.nist.gov`
[8] `http://www.clef-initiative.eu/`
[9] `http://research.nii.ac.jp/ntcir/index-en.html`

2.3.2 Defining Relevance

In 1975, Tefko Saracevic published a survey on the meaning of relevance, how it is measured, and how it was being used in Information Science and related areas [127]. He proposed that relevance is a fundamental aspect of human communication that is largely understood on an intuitive level. This intuitive understanding has "something to do with productive, effective communication—how well the process was conducted, how good were the results." Accordingly, defining relevance becomes a task in quantifying and formalizing this understanding. To help sort through existing definitions of relevance, Saracevic proposed a framework of different views of relevance. What follows is a brief summary of the views related to this thesis:

- *system's view*: These are the documents that the retrieval systems "thinks" are relevant. In simplest terms, this is the result of using exact matches of query terms to retrieve documents. More advanced approaches, like relevance ranking, are also examples of the system's view.

- *subject knowledge view*: In a modern sense, this is simply the topicality of a document to a particular information need.

- *subject literature view*: While related to the subject knowledge view, this view refers to the relationship between the document collection, documents contained within it, and a topic. In some sense, this view uses all aspects (metadata, citations, etc) of documents in the process of communicating information. For example, if we are interested in user-centred evaluation, we might reasonably expect articles written by Diane Kelly, a noted IR researcher in this area, to be useful for such a review.

- *destination's view*: This is at its core, the searcher's judgment of documents with respect to their information need outside of any other considerations (e.g., novelty).

- *pertinence view*: Much as the name implies, this is whether or not the searcher (the "destination") finds the information pertinent. This includes the ideas of redundancy and novelty in the information being communicated. Accordingly, this view requires consideration of aspects such as document presentation order (e.g., different orders may make some documents appear to be redundant).

- *pragmatic view*: This is the value of the information being conveyed with respect to the information need. It involves the utility and usefulness of the information

being communicated. In the context of electronic discovery, this is whether or not a document is a "smoking gun"[10] which would make the case easier to win.

Saracevic argues that any complete theory of relevance must take these views of relevance, as well as the others that have been omitted, into account. Using this argument as his measure, he concludes with a discussion of existing theories of relevance, where they succeed, and where they fall short.

Saracevic conducted a two-part follow-up survey in 2007 [128, 129] that presents his earlier discussions, along with new discoveries, in the context of 30 additional years of relevance research. Of particular importance is the idea of creation and derivation of relevance by inference: systems (or in some cases, humans) create situations or circumstances in which relevance is found by retrieving or identifying what is potentially relevant—for example, documents. In these scenarios, others humans (sometimes systems) derive relevance from these situations (the documents returned). Such a concept is not meant to be binary but a continuum as the roles of humans and systems can be intermixed in both stages. For example, in eDiscovery a junior lawyer or paralegal may identify a set of documents they believe are relevant and it turn over to the senior lawyer in charge to determine true relevance. Accordingly, we might view this relevance-creation process as a best effort to facilitate the maximization of derived relevance, which Saracevic acknowledges underpins many of the proposed theories and models of relevance.

In the the first part of this follow-up, modern-day manifestations of relevance are discussed which can be thought of as more concrete forms of Saracevic's earlier views of relevance. These include system relevance, topical relevance, cognitive relevance (pertinence), situational relevance (utility), and affective relevance, all relatively straightforward adaptations of Saracevic's original views. We note that the earlier *destination's view* of relevance aligns with affective relevance, since affective relevance fundamentally addresses the relationship between the user's motivations, goals, emotions and the information. Accordingly, Saracevic notes that affective relevance could be argued to form the basis of the other relevance manifestations, chiefly situational relevance.

The second part of the survey is mainly concerned with experimental and empirical results regarding relevance. Saracevic provides high-level and useful overviews of published experiments and synthesizes their results into further hypotheses that he considers to be in need of more rigorous investigation. Replication of all these experiments would be tedious; therefore, a brief summary of those that are related to the research in this thesis are presented:

[10] A "smoking gun" refers to the idea of someone being caught just after committing a crime (i.e., firing a gun at another person).

- Different assessors use similar criteria for judging relevance but weight individual criterion differently.
- Criteria importance changes when documents are presented in different ways (e.g., full-text versus titles and abstracts).
- Users appear more satisfied when systems incorporate their relevance feedback but they do not actively make use of such functionality.
- Expertise appears to correlate with agreement among assessors (higher expertise, higher agreement).
- Less expertise results in more lenient assessments (more likely to label a document relevant).
- Individual differences appear to be the biggest contributing factor to differences in relevance assessment.
- Topicality is an important relevance attribute but it is not the sole indicator of relevance.
- The order of the documents presented to assessors can have an effect on the judgments rendered.

Throughout these surveys and in more recent work [130], Saracevic has consistently and unequivocally stated that relevance is human in nature—that it is fundamentally a "you know it when you see it" phenomenon. To this end, he has argued that relevance will continue to be central to information science and, by association, information retrieval, since (searching and retrieving) information is the backbone of these fields.

2.3.3 Stopping Criterion

Knowing when to stop a review effort is crucial; however, there is no accepted solution currently. Using passive supervised learning, Webber et al. [160] investigated sequential testing as a means of validating a classifier's performance. In sequential testing, the classifier's performance is re-evaluated when additional labelled examples are added to either the training or the test set or both. The authors tested all three scenarios: fixed training, variable test; variable training, fixed test; and, variable training, variable test. At each iteration, they added 20 randomly selected (and labelled) documents to the variable set;

in the variable training, variable test scenario, 10 were added. The authors used an SVM classifier and for each iteration computed an estimated F1 score for the classifier and a lower one-sided 95% confidence interval for the estimated F1. Their hypothetical stopping criterion followed the rule that when the lower bound of the 95% confidence interval exceeded a target F1 value, the classifier would be accepted as "good." In all three scenarios, this stopping criterion stopped early and the actual performance of the classifier was less than the targeted performance. The worst offender was the fixed training set and a variable test set, which stopped early 31.6% of the time in their experiments. The others failed less often: 8.13% and 9.40% of the time.

The work of Webber et al. was motivated by the idea of minimizing labelling expenses (thus, the stopping criterion), but it does not quantify, outside of single topic exemplars, the difference in effort between when the algorithm should have stopped and when it did. Such knowledge might have indicated a reasonable trade-off in the case of using variably sized training sets; that is, while the classifier may stop too early a small fraction of the time, the savings would be worth it. Finally, while the authors conclude that a single, holdout test set that is used only once when training is complete will yield the most accurate estimate of performance, they do not suggest how to determine training is complete.

Bagdouri et al. [14] addressed some of the above deficiencies in a follow-up article that provides a general framework for testing when to stop training a machine learning algorithm. The goal of this work was to minimize annotation cost while ensuring classifier quality (exceeding a target F1 score). The authors adopted the variable training, variable test set size approach as in the above work by Webber et al. Their test set was used only once, however, to limit any bias in their estimates of effectiveness. They used cross-validation on their training set to estimate a minimal test set size, where their classifier would be predicted to exceed the target F1 score, if one existed. Accordingly, the total annotation budget was the existing training set size plus the minimal test set size.

Using this strategy, they proposed two stopping criteria and several hypothetical, lower-bound criteria. The two actual stopping criteria are straightforward: stop when the above method finds a test set size where the classifier is predicted to succeed; wait until w, which they vary from 1 to 200, of these stopping points have been seen. This wait-a-while strategy attempts to mitigate sequential-testing bias, where the first instance that the classifier is predicted to succeed is likely the result of an overestimate rather than an underestimate. In a third of their tests, the stop-immediately strategy failed to stop when a maximum budget of 10,000 annotations was met. In their wait-a-while strategy, this failure started at a similar rate but grew to a 70% failure to meet budget. When it does work, the wait-a-while strategy can achieve average cost savings of around 30% when 25 to 50 stopping points have been passed—the exact number is dependent on the topic tested. The stop-

immediately strategy results in a classifier that fails to actually meet the target 20% of the time, while the wait-a-while strategy approaches the authors' desired failure rate of 7% when the wait is increased. The hypothetical lower-bound criteria basically allow the user to "rewind time," choose the smallest observed test set, and proceed as if that were the point they had decided upon to stop. Further refinement of this approach allowed the user to pick the point where the classifier exceeded the target. Under this approach, the authors observed a maximum annotation savings of just under 40%, which Bagdouri et al. believe is attainable using practical strategies.

Cormack and Grossman [44] presented a system agnostic and two CAL-oriented stopping criterion with the aim of ensuring high reliability (i.e., achieving at least 70% recall). The agnostic stopping criterion, *target*, requires that the collection be randomly sampled until k (set to 10 in their experiments) relevant documents are found. The retrieval system finds as many relevant documents, with no knowledge of the randomly sampled documents, until all k are found. They showed that this method is provably reliable with a quantifiable level of additional review effort corresponding to the gathering of the random k relevant documents. The first CAL-oriented stopping criterion, *knee*, attempts to detect the knee that in the system's gain curve,[11] which they have observed corresponds to a drop in the marginal precision from one CAL batch to the next. The second CAL stopping criterion, *budget*, follows the knee method but also requires that an additional $\frac{10|C|}{R}$ documents be reviewed, which is the same amount as reviewed in the *target method*. The authors found that the *knee* method provides reliability comparable to the *target* method with less effort, while the *budget* method achieves the highest reliability with effort comparable to the *target* method. They reached this conclusion by quantifying the recall and effort lost by each particular method. The rationale behind this quantification is discussed in detail in Chapter 5.

A control set in eDiscovery is a mechanism to determine how well a review is being conducted, in essence it is Webber et al.'s [160] test set for passive supervised learning. Pickens [113] investigated the use of control sets, selected via simple random sampling, and the hypothetical F1 of a CAL protocol on such control sets as a barometer for actual performance on the remainder of the collection. Control sets are investigated from two directions: larger control sets, while having higher assessing costs may be more accurate; and selection of random samples is crucial to determining performance. Pickens found that improvement of hypothetical F1 on larger control sets generally correlates with reduced effort to reach 75% recall over the entire collection. However, the selection of the control set itself appears to play a large part in the success of this measurement. For the two topics used, out of 25 trials, a "good" control set (as defined by post hoc examination of

[11] Chapters 4, Chapter 5, and Appendix A all contain examples of such behaviour.

results by Pickens) occurred only 8% and 20% of the time. Given the experimental design used in this work, it is likely the case that much of the variation Pickens observed results from bias introduced by sequential testing, which is directly related to the work of Webber et al. discussed above.

Pickens also proposes a technique that uses the changes from the current ranking of the collection to a default ranking as a barometer for performance. This technique appears to suffer from less variance and erratic behaviour than the control sets without introducing any additional costs. Pickens is unable to specify the exact method of calculating rank change or the default ranking algorithm due to their proprietary nature, which greatly reduces the utility of the result. Ultimately, Pickens does not create a hard and fast stopping criterion using control sets or rank change and, instead, leaves that to future work.

2.3.4 Assessor Effects

Voorhees [152] investigated the effects of multiple assessors on the evaluation of IR systems in the context of the TREC-4 and TREC-6, using additional NIST assessments collected from the TREC-4 ad hoc task [73] and the University of Waterloo's assessments [38] from the TREC-6 ad hoc task [150]. Voorhees is concerned with changes in absolute effectiveness (how recall changes) and relative effectiveness of systems (how systems compare with each other) when these different assessors are used to evaluate systems. Under her experimental setting, absolute effectiveness scores can and will change when different assessors are used to evaluate systems. However, relative effectiveness is consistent across assessors when sufficient (at least 25) topics are used in the evaluation of systems. This study also caused Voorhees to put an upper limit on retrieval effectiveness of 60% recall and 60% precision since "that is the level at which humans agree with one another" when this effectiveness is measured by an independent assessor.

In an examination of the correlation between assessor disagreement and the rank of documents on TREC-4 and TREC-6 data, Webbber et al. [161] found that documents that appear lower in the ranked lists of participant systems and are returned by few systems will invariably provoke disagreement among assessors. They also report that when documents are ranked highly by systems and are found to be not relevant by the original assessor, subsequent assessors will likely disagree with the original assessor. A method of incorporating this model of assessor disagreement into a test of statistical significance is also presented.

Using the TREC-4 ad hoc data reported on by Voorhees [152], Webber and Pickens [163] concluded that using non-authoritative assessors to train a system will have a corresponding

drop in performance (F1) and effort required to reach 75% recall when compared to the same system trained by the authoritative assessor. Webber and Pickens presented these two non-authoritative assessors as conservative and liberal depending on the number of relevant assessments they made, the conservative assessor is then the one who made fewer relevant assessments while the liberal assessor made more. When looking at the assessors from this liberal-conservative perspective, Webber and Pickens did not observe any meaningful difference in the performance of the classifiers. Chapter 3 presents an alternate explanation for the results observed by Webber and Pickens: the perceived drop in performance by the non-authoritative assessors is due to differences in opinion rather than to natural superiority of the authority. Furthermore, we find differences between the conservative and liberal assessors in a more general experimental setting.

Carterette and Soboroff noticed several assessor trends when judging the TREC Million Query track [31]: non-relevant judgments are rendered more quickly than relevant ones; assessors are more likely to repeat the same judgment twice in a row; there is some small acceleration in assessment as time progresses; and, assessors can vary in how optimistic or liberal their judgments are. From these trends, the authors proposed several models of assessor type and investigated strategies to mitigate the effects these different types can have on evaluation. Such strategies include predicting problem cases for reassessment and using majority vote when reassessment is necessary.

Trotman and Jenkinson extended Voorhees's result to the use of multiple assessors for system evaluation [146]. They found that using gold standard assessments and those derived from multiple assessors did not result in substantially different rankings. Furthermore, they found that there was an almost 0% chance of swap between systems when the original MAP difference was greater than 0.05.

In what can be seen as an another extension of the Voorhees study, Bailey et al. [15] examined the differences among different assessor "qualities": gold - topic originators; silver - domain experts; bronze - non-experts. Silver assessors might be thought of as the alternate NIST assessors in the Voorhess study, while the bronze assessors might be the Waterloo assessments from the same study. They found that, with respect to precision-oriented measures, gold and silver assessors will tend to agree more and produce stable system rankings. Bronze assessors tended to substantially change system rankings, except for the top and bottom three systems, which were the most stable across assessors. These results led the authors to suggest that bronze assessors may be no more than an approximation of the gold-level assessors at best.

2.3.5 Depth Pooling and Alternatives

Depth pooling was first suggested by Spärck Jones and van Rijsbergen [89], in 1975, but was not implemented in practice until the first iteration of TREC [154]. The purpose of depth pooling is to find as many relevant documents as possible, which promotes accurate evaluation, while minimizing assessment effort and promoting reuse of the test collection. Depth pooling operates by combining the results of many ranked retrievals. It first orders each retrieval by score (i.e., rank) and then takes the union of the top k documents from each retrieval, such that only one occurrence of any document is in the pool resulting from this union for a particular topic. In the TREC context, k has ranged from 10 to 100 documents,. depending on the assessing resources available and the resultant pool size.

In a piece critical of the results of TREC [21], Blair argues that the depth pooling used at TREC has generally overestimated recall. Overestimated recall can be mitigated, Blair argues, by conducting full evaluation over the entire collection for some topics. In this particular article, Blair provides no firm details for conducting such an evaluation in a cost-effective manner.

The limitations of depth pooling [51, 25] have motivated various alternatives, including the statistical sampling method of the later Legal tracks. Cormack et al. [53] propose the use of interactive search and judging (e.g., manual review) as an effective alternate which correlates well with depth pooling but requires fewer assessments. Also proposed by Cormack et al. is move-to-front pooling which queues each participant run and traverses the first queued run in rank order until n non-relevant documents are seen, at which point it is moved to the back of the queue and the process begins with the next run. The authors also discuss per-topic and global variants, which are shown to correlate well with depth pooling, but require less effort.

An alternate view that has been presented [126, 29, 30] is the idea that more topics with fewer judgments are better for comparing retrieval systems. The crux of this argument is that shallow pools (depth pools where k is small, say 10) still find a substantial number of the most likely to be relevant documents (due to relevance ranking) while minimizing assessor effort on judging documents that are less likely to be relevant, as they appear lower in the ranked lists of systems. This currently "wasted" effort could be then spent on assessing more topics with shallow pools, resulting in more relevant documents in total being found. By having more relevant documents used for evaluation, a more precise comparison between pairs of systems will result. Accordingly, such an evaluation technique may not be a suitable evaluation technique for high-recall retrieval scenarios, since absolute performance is crucial. Answering the question "Did the system achieve high-recall or not?" requires as complete a set of assessments for a topic as possible or, at least, a sample that

can be extrapolated to the entire collection.

In the context of the TREC Million Query track, Carterette et al. investigated two different pooling techniques to facilitate the above evaluation [29]. Both pooling methods are primarily concerned with Mean Average Precision (MAP) evaluation. One method, called Minimal Test Collections (MTC), selects documents that will cause the biggest changes in MAP between systems; the other, statAP, attempts to select documents which most accurately estimate MAP. Both techniques result in estimations of MAP that agree highly with each other and less so with the gold standard, which the authors argue is indicative of a lower quality in the gold standard. For both techniques, stability of their MAP estimation is reached with a couple hundred topics and up to 40 judgments per topic. Furthermore, the authors show the MTC is likely to result in a reusable judgment pool. It is not readily apparent whether or not such techniques would scale to work in the case of high-recall retrieval, where having fully labelled collections is useful for conducting experiments with human-in-the-loop systems.

It has been observed [140, 151] that manual runs (i.e., a human involved in the retrieval process) can return many unique documents that would not be found by pooling solely automatic runs. Jayasinghe et al. [83, 84] proposed a method of combining result fusion and machine learning to identify many of the documents that were returned by submitted manual runs while using solely automatic runs to train their learning algorithm. The authors began by removing manual runs from several TREC test collections as a holdout, and then trained the machine learner on the assessed documents resulting from pooling automatic runs. Then they used ranked result fusion [27, Chapter 11] of the unjudged documents submitted by automatic runs. The most promising subset[12] of the fused result list was then re-ranked by the machine learner. From this a second judging pool is created, which Jayasinghe et al. show returns many of the same documents as the submitted manual runs.

2.3.6 Sampling and Estimation

Stratified sampling is a statistical technique that aims to provide more accurate estimation of population properties without overly burdensome costs when it comes to selecting the sample. For example, (simple) random sampling may underestimate the number of relevant documents in a collection when the sample size is too small or when a "bad" sample is selected by chance. Stratified sampling leverages the idea that some subpopulations may have variation in the desired properties to be measured. A concrete example is that the

[12] What exactly this means is not specified in either publication

most highly ranked documents returned by all systems for a particular TREC task likely has a higher incidence of relevant documents than those documents that are not returned by any system. Accordingly, in stratified sampling, the population is divided into various (mutually exclusive) subpopulations, called strata. Each stratum can then be sampled, potentially by random sampling, and the desired properties measured from this sample. Note that each document sampled from a particular stratum has an inclusion probability (or weight) which indicates how many documents in the stratum this single document "represents" in relation to the whole corpus. The sampled documents, with their inclusion probabilities and measured properties can then be used to estimate the property's value for the entire population.

The TREC Legal track used several different methods to conduct stratified sampling [145, 109, 76, 46, 71]; a brief overview of the main variants follows. The first method used was deemed the L07 method [145] due to its introduction in the 2007 iteration of the track. Three strata were used in the L07: the top 5 highest-ranked documents submitted by each run for a particular topic; the depth-B stratum, where B is the number of documents returned by a reference boolean query; the depth-25,000 stratum, which was the maximum number of documents a run could submit for a topic. These strata were sampled such that the top-5 stratum was completely sampled and the other two strata were sampled to an accuracy of at least 5 simple random sample points. The latter restriction was put in place due to a fixed budget for assessment and the top-5 documents' consuming more or less of that budget, depending on the topic.

The 2009 Legal interactive task used a more involved version of stratified sampling, since participants provided a full binary classification of the document collection. Using this strategy, strata corresponded to agreement on document relevance by submissions. For example, the documents that all submissions thought were relevant would be considered one strata and the documents that no group found to be relevant would be another stratum. Supposing there were N submissions, then 2^N strata would be created where each stratum corresponds to documents that were found relevant by a particular number of submissions. Generally, each stratum was sampled with a rate reflective of its full-population proportions, but to ensure that sufficient numbers of documents were sampled from each stratum some smaller strata were oversampled and some large strata (i.e., the all-not-relevant stratum) were undersampled. In some sense, this biasing of the strategy is done to ensure that all assessment effort did not go to the all-not-relevant stratum which presumably would have few relevant documents and thus be a waste of effort compared to strata that are smaller but more likely to contain relevant material.

In 2010, the Legal learning task [46] used a combination of stratified sampling and depth pooling. Four strata were created (100, 1000, 10000, and 1000000), which corresponded to

the traditional TREC depth-X pool with any smaller pools removed from it (e.g., the 1000-stratum was the depth-1000 pool with all stratum 100 documents removed). Strata were sampled such that the 100-stratum was sampled completely and only 2,750 assessments could be rendered. This resulted in each of the other three strata contributing similar numbers of documents but having far smaller sampling rates than their full-population proportion would otherwise indicate.

Oard et al. provided a summary of current eDiscovery practice, and its relation to information retrieval and the TREC Legal track [108]. They highlight that issues exist with the previously discussed statistical sampling employed by the Legal track, such as a larger focus on documents that participants think are relevant, which leaves some classes of documents under-represented in assessment. In spite of these issues, Oard et al. conclude that coordinators were able to yield more accurate estimates of recall and precision than would have been accomplished through traditional depth pooling. They advocate for a focus on cost and effectiveness rather than just on effectiveness, which is common in IR evaluation. They end with a call for all parties, industry and academia, to come together to develop best practices for eDiscovery.

Hedin and Oard analyzed the first year of the TREC Legal interactive tasks and shared lessons learned [75]. They observed that one topic had extensive appeals conducted on it, while others had very few. The appeals resulted in a general across-the-board increase in estimated recall. It is important to note that the topic authority's disagreeing with the first-pass assessments for any non-trivial percentage of the time would, generally, increase estimated recall, since it would either remove a document that had to be retrieved (decreasing R) or reaffirm that a relevant document should be retrieved (restoring credit to a found document). It was also observed that making extensive use of the time allotted to interact with the topic authority could substantially boost recall without negatively affecting precision.

Following a similar line of inquiry, Webber et al. [162] argued that the appeals process of the TREC Legal track is self-motivated since participants would likely only appeal documents that help themselves. The authors argued that while active and enthusiastic appeals can reduce assessor errors, appeals can never eliminate errors entirely. They acknowledge, however, that the appeals process appears to have been reasonably successful for several of the topics for which it was used. They further argue that this assessor error, when combined with the stratified sampling of the track, can mask the apparent performance of the systems arising from the raw, rather than extrapolated, recall and precision.

Due to their perceived issues with the appeals process and its costs, as well as with the effect of assessor errors, the authors propose using a method which samples the stratified

sample itself for adjudication, eliding the appeals process, to form unbiased estimates of false positive and false negative rates of the stratified sample. By incorporating these unbiased estimates of assessor errors, they argue that better estimates of system performance can be achieved with less bias and potentially less cost than that of the appeals process.

Webber presented several techniques that could be used to estimate recall confidence intervals for a (high-recall) retrieval system [158]. Webber discusses the pros and cons of using various distributions before testing several formulations on the TREC 2008 and 2009 Legal tracks. Webber found that using the Beta-Binomial distribution is superior to other distributions examined, which includes the Normal distribution used at the TREC Legal track.

Kantor et al. [90] provided several methods for estimating the number of relevant documents in a collection using the idea of mark-and-recapture population estimation.[13] They proposed three methods, two of which failed partially due to lack of independence between TREC systems, and one of which appears to work in some cases but not in all. The most successful approach appears to be a formulation of depth pooling with a varying depth, which is not too dissimilar to Cormack et al.'s move-to-front pooling [53].

Vinjumur, Oard, and Paik [149], while investigating the reusuability and reliability of the test collection created as part of the TREC 2010 Legal interactive task's privilege topic, found that teams were better than chance at identifying false negatives from the first-pass assessments but were no better than chance at identifying false positives. Combined with the relatively small amount of appeals (only 8%) and the noticeable changes in estimates of recall and precision after appeal, they argue that using the 92% of first-pass assessments, which were unappealed, to estimate recall and precision may be inadvisable and a more robust solution would be to use error rates (derived from the appeals) to produce a more accurate estimate.

In an article for the sixth DESI workshop [57], Paul Dimm argues that the F1 measure, commonly used in IR evaluation, is not suitable for eDiscovery evaluation since it is insensitive to meaningful differences in recall and precision—namely, two systems could have similar F1 scores while having markedly different precision (and recall) scores[14]. Dimm presents a mathematical formula for extrapolating the precision necessary to achieve a target recall level from a single recall-precision point mapped to the closest matching recall-precision curve (from a family of such curves). The argument in favour of this method is

[13]This is a sampling process where a portion of the population is captured, and then released. At some point later, a new sample is collected and the proportion of the second sample that is marked can be used to calculate the population size.

[14]For example, a precision of 0.7 and recall of 0.3 has the same F1 as a precision of 0.3 and recall of 0.7.

that it facilitates comparison of systems at relatively "fair" points of recall without requiring a full precision-recall curve, which may not exist if a system only produces a binary classification of documents (relevant/not relevant) rather than a likelihood of relevance. Furthermore, Dimm's other motivation for this extrapolation of precision is to estimate the amount of review effort necessary to achieve a target level of recall. This estimate review effort can then be used as a barometer for the cost of additional review and thus of the restrictions placed on the cost of an eDiscovery production.

There are several issues with Dimm's work that make the results unconvincing. The foremost is that the estimated review cost of a "final" review of the collection and excludes the costs associated with training and testing of the system. Both of those costs may very well be substantial, so the estimated review cost may not reflect the true cost of a system. It is not clear how useful this estimated cost would be in the case of a system that only produced binary judgments (i.e., that do not produce a ranking/score for the documents). Presumably such a system would require review of additional documents that were deemed not relevant, but it is not clear how such a revision should take place nor whether the resulting cost would be the same.

Additionally, it is not clear how applicable the examples that Dimm provides of his technique in use are for eDiscovery tasks. Systems, tasks, and topics are all elided and only token reference to task difficulty (easy and hard, based on prevalence) is presented. Moreover, only 2 such tasks and 6 systems are used in the evaluation of his extrapolated precision. The generality of the these results is an open question without knowing the quality of the comparison being made and without additional analysis.

2.3.7 Effects of Labels on Systems

The effect of label noise (incorrect relevance assessments) on training machine learning algorithms is an ongoing topic of investigation within the IR community [92, 47, 135], as well as in other machine learning-oriented communities [24, 65]. A more comprehensive study than appears herein was produced by Frénay and Verleysen [65] in 2013. They begin by distinguishing two important types of noise in machine learning: feature noise, which are inaccuracies in the values associated with features (e.g., an improperly computed feature weighting for a word); and class noise, where the document is associated with the wrong class (e.g., in the binary case, mislabelled as relevant when it is not relevant). The authors point out that it has been shown [169, 125] that class noise tends to be more damaging than feature noise, since there is only one class label (typically) for a document but many features and since features themselves are used differently by different algorithms (e.g.,

some algorithms may only use their presence and not the associated weight) while class label is always important. Accordingly, their review focuses on how to combat label noise effects on machine learning and leave feature noise to others, with the restriction that they do not deal with situations where label inaccuracy is intentional or malicious and, for simplicity, the assumption that label errors are independent of each other. To begin they categorize label noise into three categories: noisy completely at random, where label noise is randomly and uniformly distributed in the collection; noisy at random, where label noise is still randomly distributed but some classes may be more likely to suffer from label noise; noisy not at random, which models the situation where label noise may be more likely in particular segments of the document collection (e.g., similar documents that belong to different classes).

Frénay and Verleysen then discuss how label noise can result in deterioration of classifier performance and that label noise will generally increase the required number of training examples. Some noteworthy results include the fact that k-nearest neighbour classifiers can deteriorate rapidly in performance when label noise is introduced and k is small, that SVMs and logistic regression have been observed to fail with small amounts of label noise, that removing label noise can create learned models that are less complex, and that when test documents are polluted with label noise, their noise can pollute estimates of classifier effectiveness.

Following this is a discussion of methods dealing with label noise in which the authors emphasize that many modern machine learning algorithms (SVM, logistic regression) may not be robust to label noise, even in the uniform, random case. They conclude that most of the empirical results suggest that even noise-robust algorithms still suffer from performance decreases when label noise is present. They can, however, generally be managed by not overfitting them[15] to the training data. With respect to cleaning label noise, Frénay and Verleysen highlight the tension between removing too many examples (some of which are correct), which results in impaired performance, and removing too few noisy examples, which can increase the cost of fixing labelling errors. Label noise-tolerant algorithms are also examined but a longer discussion of it is omitted here due to its technical complexity. The salient point from the review is that while noise-tolerant algorithms are capable, they are much more complex than other solutions, which can result in overfitting and should be used with that in mind. The authors conclude that the correct choice of machine learning algorithm depends on the quality of labels (e.g., expert versus non-expert versus crowd sourced), existing knowledge of noise distribution, and whether or not label cleaning is feasible.

[15]Overfitting in this sense refers to focusing on the noise rather than other properties of the data.

Brodley and Friedl investigated the removal of noisy labels from training data [24]. They propose a two-stage process for machine learning: an initial stage consisting of one or more base learning algorithms are used to filter the data for noisy data, after which the cleaned data is fed to the desired machine learning algorithm. When multiple classifiers are used to render a single judgment by combining their judgments, the result is what is termed an ensemble classifier. Two types of ensemble classifiers are investigated: majority vote, where a training document is labelled noisy if half of the base classifiers cannot correctly label it; consensus, where a document is labelled noisy only if all base classifiers cannot correctly label the document. Brodley and Friedl discuss how the rates of false positives (correct data is labelled as noisy) and false negatives (noisy data labelled not noisy) differ in these ensemble classifiers from their rates in single base classifiers. In particular, they highlight the potential of consensus filters to make more false negative errors than would a base classifier since they require rejection by all base classifiers. Label noise was injected into their experiment in a semi-random fashion, for a given error rate (0% to 40%), a label was switched, with probability equal to the error rate, only if it belonged to a predetermined "troublesome" pairs of classes[16] identified by the authors prior to experimentation. Accordingly, some labels belonged to more of these "troublesome" pairs than others, making the noise distribution most similar to Frénay and Verleysen's noisy at random scenario, since some classes would be more likely to be switched than others overall. Empirically, this process of filtering out noisy data, even when a single base filter is used, allows baseline performance (no noise introduced) to be maintained up to the introduction of 20% noise, after which performance can drop substantially. Differences of up to 30% in accuracy are reported at the highest levels of noise when compared to no noise performance. Overall, majority vote is found to be superior to consensus, since it tends to be less conservative and eliminates more noisy data. However, the authors caution that the majority vote filter can also result in more correct data being thrown away, which may not always be desirable.

The preceding discussion has focused on label noise and machine learning at large, rather than specifically on its relation to high-recall retrieval tasks. Email spam filtering is a high-recall task, the goal being to identify and, correspondingly, not to show to the user, all spam[17] email messages. On data that has been vetted for high quality, modern spam filters have been seen to achieve close to excellent performance (e.g., AUC often exceeding 0.95) [37]. Sculley and Cormack investigated how these classifiers perform when label noise is introduced into the training data [135]. To begin their investigation, they

[16]These were pairs either of which an author could conceive of as the correct label for a document.

[17]In spam filtering, spam is the term of art for the relevant class and ham is the term for the non-relevant class.

introduced synthetic noise through the random and uniform switching of training data labels (spam to ham, ham to spam) at various levels (0 to 25%). At the highest levels of noise, the best-performing no noise algorithms (including logistic regression and SVM) had the worst performance. For example, on one data set SVM drops from a (1-AUC)% score of 0.031% to 21.680%. The authors hypothesize that these substantial decreases are due to the aggressive learning policy that these approaches take in order to learn from new spam attacks quickly. These aggressive policies result in overfitting and thus worse performance. They authors found that when overfitting is controlled for (by parameter tuning) logistic regression and SVM become more much noise tolerant at high levels of noise (SVM improved from 21.68% to 0.048%). Sculley and Cormack conclude that noise-tolerant spam filters have better performance (with respect to AUC) than noise-intolerant variants. They also note that these tolerant filters have a harder time coping with natural rather than synthetic noise.

Cormack and Kolcz [47] argue that label noise (errors) in the evaluation set of labels as well as in the training labels can mask filter performance and corresponding true error rates of the filters themselves. They propose two techniques to accommodate errors in the gold standard set of labels. The first is to use differential comparison, where a third independent adjudicating party vets (labels) disagreements between two other parties, to determine which of two sets of labels is more accurate. Accuracy is based upon their agreement with the adjudicating party. Using differential comparison, Cormack and Kolcz showed that a noise-tolerant SVM produces more accurate labels than the official TREC Spam track gold standard with respect to the third-party human labels. Based upon these results, they argue that it might be better to use the filter's labels as a source of truth. Such use would result in biased evaluations, so they suggest that a more general and less biased approach would be to use the ensemble of multiple filters.

Using results from Lynam and Cormack [104], who observed that filter fusion outperforms individual filters even when some filters are substantially better than others, Cormack and Kolcz created a pseudo-gold standard by combining their existing filter labels via logistic regression. This new pseudo-gold standard was observed to be indistinguishable from the human labels and substantially more accurate than the TREC gold standard. Cormack and Kolcz found that using these more accurate, less error-filled sets of labels yields comparable performance to the synthetic noise results of Sculley and Cormack [135] for the best spam filters and that the previously reported poor performance was inflated due to label noise.

Prior to the above work, Kolcz and Cormack investigated the efficacy of using genre (personal email, scam email, advertising email, and so on) information to improve classifier performance when noise is present [92]. Email genre is found to be a strong indicator of label

noise, and this noise was found to be higher for the under-represented class in a genre. For example, in personal emails ham has a higher prevalence and, accordingly, a lower label noise. Kolcz and Cormack observed that when genre-specific spam filters (trained only on genre-specific examples) are combined with a general-purpose filters (genre-agnostic, trained on all available examples), in a similar manner as above, the performance achieved was the best reported for the data sets tested. When the genre-specific filters were used without the general-purpose filter, performance was substantially worse, indicating that genre information alone may not be sufficient to effectively filter spam.

In a blog post [112] for the Catalyst Repository Systems website (a commercial vendor of eDiscovery software), Jeremy Pickens discusses the issues surrounding what happens when the proprietary Catalyst software is trained with wrong judgments. Using an unspecified TREC Legal test collection, he selected documents for which, during the appeals process, the initial assessment was overturned by the topic authority. Using the first-pass and second-pass assessments, he trained two versions of the Catalyst software and ranked the remainder of the corpus. In plotting the gain curve for two topics using these two different training regimes, he notes that while the initial-pass, incorrectly labelled training documents do not necessarily outperform the second-pass trained version, they are more competitive than linear (manual) review of the collection. Furthermore, at equal levels of effort, the wrong labels appear (at times) to be capable of producing a classifier that is able to achieve higher recall than the correct labels. Pickens concludes that "garbage in" does not necessarily result in "garbage out," although he does acknowledge there may be a loss in effectiveness compared to completely correct training data.

2.4 Rendering Assessments

The process of rendering assessments typically requires a single human to decide the relevance of some set of documents with respect to an information need that they may or may not have created themselves. In the Cranfield paradigm, the assessor judging relevance should be the creator of the information need but this is not necessary. As highlighted by Saracevic [129] and others[153, 126, 25, 152], the Cranfield paradigm makes several assumptions about relevance assessments and the relevance assessment process that do not always hold.[18] They include the following: assessors only render binary judgments (relevant or not relevant); assessments are stable and independent (assessors don't change their conception of relevance over time or when other documents are seen); assessors are

[18]Saracevic, in the cited work, notes that great improvements have been made to IR systems and evaluation in spite of these assumptions not always being accurate.

consistent (there is not inter- or intra-assessor variation among assessors). In this section, we focus on examining how assessors themselves can affect the assessment process which is deeply ingrained in the above assumptions.

2.4.1 Assessor Knowledge

When providing instructions to assessors for electronic discovery, Webber et al. [165] found little difference was found between detailed and simple relevance guidelines using TREC Legal track topics and corpora. It was also the case that their assessors agreed more with the topic authority than with the first-pass assessors.

Wang and Soergel explored differences in assessor background when judging documents for eDiscovery [155] by having library and information science students and law students review two previous TREC Legal topics. No significant difference was found between the two types of students, but law students tended to be more accurate and quicker at assessing documents.

Kinney et al. [91] compare domain experts and non-experts to find that generally non-experts overestimate document relevance when there are a preponderance of keywords and underestimate relevance when there is a dearth. In addition, they report that overestimation can be mitigated by giving non-experts precise query descriptions rather than topic outlines.

Scholer et al. found that how assessors are primed can significantly affect the relevance scores they render for documents [132]. Assessors who were presented with only non-relevant documents scored documents (common to all experimental conditions) significantly higher than those who were exposed to relevant or highly relevant documents in an initial priming phase. Also examined was the effect of the psychological need for cognition[19] on relevance assessments. Assessors who exhibited a low need for cognition were found to assign lower relevance scores, take less time to score document relevance, and agree less often with the expert assessor.

Al-Harbi and Smucker [10, 11] have investigated the interplay of assessor certainty and relevance judging behaviour. In their user study, they found that 7.71% of relevance judgments were made with low certainty, which indicated to them that binary relevance categories may not be sufficient to capture assessing behaviour and that systems may want to capture some idea of assessor confidence when collecting assessments. Al-Harbi and Smucker argue that such uncertainty may stem from an assessor's lack of knowledge, topics that are either too specific or too general, and documents that are difficult to process.

[19] A measure of the extent to which an individual enjoys concentrated mental activity.

2.4.2 Affecting Assessors

Graded relevance is simply the extension of binary relevance with the idea that some documents can be more relevant than others. For example, a highly relevant document may be one that *should* or *must* be returned (e.g., a "smoking gun"), while a relevant document *could* be returned but may not substantively alter the user's perception of the returned documents or their information gain.

Using 7 levels of graded relevance, Eisenberg and Barry observed noticeable differences (differences of greater than 1 relevance grade on average) between presenting documents for assessor review when documents are listed from high to low and low to high in terms of relevance[63]. In the first case, assessors were found to underestimate highly relevant documents and overestimate low-relevance documents. In the second, assessors tended to overestimate relevance across the board. Interestingly, when these two conditions (high-to-low and low-to-high) are compared to a random ranking of the documents, low-to-high ranking and random ranking produce similar relevance grades, on average, for the most highly ranked documents.

Magnitude estimation is a technique created by S. S. Stevens that uses arbitrary real numbers to describe the impact of a stimulus on an observer (e.g., the loudness of a sound) [142]. When adapted for use in IR retrieval and the rendering of assessments, magnitude estimation allows assessors to split hairs as finely as they desire since it allows them to rank documents as more or less relevant without having to pigeon hole them into a fixed set of categories or grades. When Eisenberg [62] and Eisenberg and Berry [63] used magnitude estimation in similar task models to those described above, these differences in behaviour became less pronounced and, in some cases, non-existent. More recent work [133, 105, 147] has investigated the suitability of magnitude estimation to assessment and evaluation. Generally speaking, researchers have found substantive differences between magnitude estimation and graded relevance in terms of which documents are found to be more relevant and the relative system rankings created when these different assessments are used. However, this research does not make clear whether these differences are beneficial in general.

In a related study [82], Huang and Wang observed similar (to Eisenberg and Barry) order effects when graded relevance is used and between 15-60 documents are being assessed. They did not find substantial order effects when 5 or 75 documents are used. They hypothesize that in the former case, the set is too small for any effects to take place (consistent with work by Parker and Johnson [115]), while in the latter, fatigue may start to become an issue rather than order effects.

Smucker and Jethani[137] examined the effect of ranked lists on assessor performance

with a particular focus on precision in such lists. Primarily, the authors found that in lower-precision lists, assessors will tend to have a higher true and false positive rates and in higher-precision lists, assessors have a lower true positive rate. Further analysis of this user study revealed that the majority opinion of their participants had an equivalent true positive rate to the gold standard, but a higher false positive rate, which, the authors contend appears to result from the gold standard being more conservative (i.e., marks fewer things relevant) [138].

Jethani [85] extended their previous work by examining the effect of extreme levels of precision (0.1, 0.5, 0.9) on relevance-assessing behaviour. It is worth noting that these document sets were created by randomly selected relevant and not-relevant documents (according to NIST assessments) and then randomly shuffling them together. Jethani reports that the best true positive and false positive rates are achieved at the 0.5 level, which is significantly different from the 0.9 level. In the context of Cormack and Grossman's archetypal protocols, we might reasonably expect the CAL protocol to produce more precise batches of documents, so one might expect similar results to the higher-precision batches. However, CAL forms an implicitly ranked list of the document collection, so there may be order effects that are more in line with their earlier works (higher precision, lower true positive rate).

Chandar, Webber, and Carterette report that low document cohesion (e.g., many subtopics) and documents that are easier to read can significantly affect assessor agreement [164]. They posit that the latter results from many documents in their set using simple words but covering complex topics due to their nature, being speeches and interviews.

2.5 High Recall and the Text REtrieval Conference

The Text REtrieval Conference (TREC) is an annual conference co-sponsored by the National Institute of Standards and Technology and the U.S. Department of Defense. Historically, TREC grew out of the Defense Advanced Research Project Agency's (DARPA) TIPSTER project, which sought to create a large information retrieval test collection [154, Chapter 2]. TREC aims to further research on information retrieval systems for large text collections through increased communication between academia, industry, and government. TREC supports a "track" format, where different research areas in Information Retrieval are represented: electronic discovery, spam filtering, web search, microblog search, and so forth. A track has a corresponding task(s), which is generally some domain-specific application of information retrieval. While the precise set-up of individual tracks may vary, the tracks and associated tasks tend, but are not required, to follow the same basic outline

which we now briefly describe. Organizers of a track assemble or find a suitable data set, NIST (or in some cases, practitioners) create a set of topics (usually 50 but this is also track dependent),[20] participating groups find documents[21] in the collection that are relevant to the topics and compose them into "runs," and assessors, either provided by NIST or other parties (e.g., volunteer lawyers or doctors), determine the relevance of returned results, often through the use of depth pooling or stratified sampling [154]. Evaluation is performed based upon the assessments made. NIST hosts an annual conference that allows track participants to present and discuss their results and to determine future directions for the track.

In the following subsections, we provide brief descriptions of TREC tracks that are related to the Total Recall track (described in Chapter 4) and high-recall retrieval in general.

2.5.1 TREC Legal Track

This section provides a brief synopsis of the TREC Legal track's test collections and of tasks that the track offered during its time at TREC (2006 to 2011). We provide slightly more detail here than elsewhere due to the fact that much of the work on the Total Recall track (Chapter 4) was inspired by the work done in the Legal track. Generally speaking, the Legal track was concerned with investigating retrieval systems and techniques to tasks that resemble legal discovery scenarios. Accordingly, the tasks themselves evolved over the track existence from relatively simple ad hoc search to complicated interactive learning scenarios, which reflected the move (in real-world electronic discovery) away from keyword search to more complicated solutions involving machine learning.

Test Collections

The TREC Legal 2006 track [18] used the Illinois Institute of Technology Complex Document Information Processing Test Collection version 1.0 (IIT CDIP), which was a set of documents released under the Tobacco Master Settlement Agreement[22]. The IIT CDIP 1.0 collection consists of 6.9 million document records in the form of XML elements. The

[20] Topics are task dependent but can be thought of as queries in a very general sense.

[21] *Documents* is used in a general sense and could refer to passages within a larger text, summaries of multiple pieces of text, or even images.

[22] These documents were released after the attorneys general of 46 states in the U.S.A. settled lawsuits with the four largest tobacco companies at the time.

primary content of these XML elements is text rendered using optical character recognition (OCR). This collection proved to be full of errors due to the limitations of OCR at the time and the quality of the original documents themselves. For our purposes, we will refer to this collection as the TREC Tobacco or IIT CDIP collection.

Starting in the TREC 2009 Legal track, an additional collection was added due to a perceived resemblance to more realistic corpora [76]. This collection was a version of Enron's email that had been collected from requests by the Federal Energy Regulatory Commission (FERC). The initial collection distributed was one of several produced by Aspen Systems (now Lockheed Martin) on behalf of FERC. After processing, this collection consisted of 847,791 documents (including emails and attachments). This version of the collection was used for the 2009 version of the track; a second rendering was produced by ZL systems for the 2010 and 2011 versions of the track [46, 71]. This second version originally consisted of 1.3 million documents, including emails and attachments, but was later deduplicated into 685,592 documents by the 2010 Legal track coordinators. For our purposes, we will refer to these corpora jointly as the TREC Enron collection and, when necessary, will use *v1* or *v2* to distinguish them.

Tasks

The Legal track had many subtasks over its 6-year duration (2006–2011) at TREC. The first task introduced can be characterized as a traditional ad hoc search, except that additional sets of boolean queries, mimicking those present in actual discovery cases (e.g., an initial set, some number of intermediate sets, and a final set after arbitration between both parties) were provided to participants, who could use them in any desired manner. The ad hoc task was largely kept the same until it was later merged, in 2009, with the relevance feedback task into a single "batch" task. In 2007, a relevance feedback task was introduced which provided all previously rendered ad hoc task assessments to participants and required them to identify new, previously unassessed relevant documents, a task reminiscent of earlier routing and filtering tasks. An interactive task was also introduced which allowed participants to conduct a review effort in whatever way they desired (e.g., manual search and review). The initial interactive task was limited to 100 documents per team to facilitate complete assessment of the judging pool.

In 2008, the interactive task was massively overhauled and given additional features that would persist in some form until the end of the Legal track. These four features are as follows: (1) a designated *topic authority* (a lawyer) would act as the sole arbiter for a topic, including designing the intent and scope of a topic; (2) participants were permitted to engage with the topic authority to clarify matters of relevance for a topic; (3) the

task objective was to provide, for each topic, a binary assessment for all documents in the collection; (4) an appeal and adjudication process would be conducted to correct any possible errors in assessment. The appeals process was put into place becase, due to the size of the collection and resultant pool, the topic authority could not assess the entire pool; instead, this pool was judged by volunteer legal professionals and the topic authority would guide their judging and resolved any disputes.

The batch task was replaced by a learning task in 2010. In this task, participants was given the topic (i.e., a mock legal request) and a preliminary seed set of document assessments. Participants were then required to produce a probability of relevance for every document in the collection and for each topic. The learning and interactive tasks merged in the final year of the Legal track to form a single learning task. The underlying structure of the interactive task was maintained while the goal became to produce a probability of relevance for every document in the collection with respect to each topic.

Legal track evaluation was highly volatile, changing from year to year, and included such measures as R-Precision, F1, Recall, AUC, and Hypothetical F1. Of note is the requirement in the 2008 ad hoc task which had participants submit a cutoff in their ranked list that they believed maximized F1. This influenced inclusion of the "call your shot" mechanism in the Total Recall track (Section 4).

Furthermore, the L07 sampling technique was proposed and used to help estimate recall and prevalence in the collection starting in the 2007 iteration of the track. The goal of this algorithm was to accommodate deficiencies in using only a depth pool to estimate recall and in other evaluation measures. This algorithm and subsequent variants have been described in Section 2.3.6; therefore, we do not repeat that discussion.

2.5.2 Related TREC Tracks

The routing task was a main task for the first 6 years of TREC [154, 150]. In the routing task, a system was given a topic and a set of relevance judgments for documents in a training corpus with respect to that topic. The system is then expected to produce a ranked list of documents from a separate test corpus with respect to that topic. The filtering track was developed after several iterations of the routing task and then contemporaneously with it [154, 117]. The initial filtering task required participants to perform the routing task with the additional requirement of determining for *every* document in the collection whether it should be revealed to the user. This was followed up by batch-adaptive filtering, where any document that the system decided could be emitted would have its true relevance revealed to the system. The adaptive aspect resulted from the system's being able to learn from the

true relevance of any document that it decided to show to the user. Contemporaneously, a non-batch version of adaptive filtering was also offered. The non-batch aspect meant that there were no initial training assessments and a system started with only a topic. With the introduction of adaptive filtering, came a requirement to process the test collection in a fixed, usually time-stamped, order. Non-batch adaptive filtering would eventually be replaced by adaptive filtering with positive examples, where a small subset of random relevant documents are provided to the system with the topic statement. Modern approaches to high-recall retrieval can be most generally be seen as adaptive filtering tasks with the exact nature dependent on the application domain (e.g., eDiscovery does not typically care about time-oriented processing, though an eDiscovery collection may grow over time as additional documents are processed).

The Interactive tracks [154, 77] at TREC focused on how retrieval systems performance when driven by a human rather than ad hoc retrieval. Many issues were found in attempting to compare systems across sites, even when the use of a control retrieval system was required. Accordingly, cross-site comparison was rather limited in later iterations of the track, and generally participants compared different aspects of their own systems. Due to these types of difficulties, the track put some focus on the aspects of topics (e.g., subtopics). *Aspectual recall* is the number of aspects discussed in all documents found by a searcher, divided by the number of aspects for a particular topic (as determined by a NIST assessor). *Aspectual precision* is the fraction of found documents that contain one or more aspects. The extent to which these types of measures might benefit high-recall retrieval remains to be seen.

The High Accuracy Retrieval from Documents (HARD) [154, 12] track can be seen as extensions of the traditional TREC ad hoc task and the Interactive track. Participants were allowed to spend limited time "clarifying" the topic with an end user (i.e., the NIST assessor responsible for that topic). Typically, this clarification was done via a Web form that the assessor could fill out or use to mark document passages as relevant. In addition to the standard TREC topic information, the HARD track provided additional metadata with each topic, including related text that was on topic but not relevant and related text that was both on topic and relevant. In some sense, the formal complaints and topic authority interaction that occurred in the Legal track are reminiscent of the HARD track. Unfortunately, the Total Recall track does not at this time (or in the foreseeable future) provide either of these two features.

The Relevance Feedback track [26] focused on investigating the effect that the number and relevance of feedback documents could have on different feedback algorithms. These feedback algorithms followed a more traditional route of query refinement and expansion, which can involve discarding bad query terms and finding better ones. The end goal of

gathering relevance feedback is to improve subsequent results presented to the searcher by improving the query. High-recall retrieval often involves a feedback loop of some type; for example, active learning asks the user to assess documents to aid in producing a better classifier. Accordingly, if we consider the idea of relevance feedback more generally—to improve subsequent results presented to the user—the continuous active learning approach of Cormack and Grossman [40] could be considered extreme relevance feedback. However, the goal of the Relevance Feedback track was not necessarily to ensure high-recall but to improve the result of subsequent searches, which generally meant that systems were evaluated based upon precision rather than recall.

The TREC Spam track [37, 49, 36] can be seen as an adaptation of the Filtering track to the email spam filtering domain. Initially, the Spam track's sole task was one of immediate adaptive filtering with no prior information provided by the track coordinators—systems could use whatever existing training data they had. In this case, a spam filter would classify a message and then be trained on the gold standard. This was subsequently extended to include a delayed feedback approach, wherein the filter would classify X emails and then be trained on those same emails at some later point, and X would be randomly generated with an exponential distribution. This delayed feedback task was subsequently modified to allow for the fact that a response may never be given for some emails. An active learning task was also used to model the situation where a user would not respond to every labelling request. To simulate this scenario, a small quota is provided to the filter and the "user" would not respond once that quota was exceeded.

One of the more interesting aspects of the Spam track was the use of personal email archives for evaluation. This usage required sandboxing the spam filters to ensure that no data would be transmitted to the outside world. This submission of filters required the implementation of basic command-line interfaces. As it turned out, most filters required a great deal of manual intervention by the track coordinators to ensure they would run.[23] The Total Recall track made use of the lessons learned in the Spam track to construct its sandbox protocol.

2.6 Discovery of Electronically Stored Information (DESI) Workshops

The six DESI workshops [1, 2, 3, 4, 6, 7] have aimed to promote interaction between legal professionals and researchers regarding various topics related to the discovery of

[23] Based upon discussion with Gordon V. Cormack, one of the coordinators.

electronically stored information. Workshops take place approximately every two years, usually as part of the International Conference on Artificial Intelligence and Law. However, DESI II was held as an intermediate workshop at the University College London to provide a European point of view on the topics discussed at the first DESI workshop. Due to the wide variety of topics covered, we do not attempt to provide a summary of each workshop but instead have interspersed the discussion of relevant refereed publications at the DESI workshops throughout this chapter in the most appropriate section.

Chapter 3

Impact of Surrogate Assessments on High-Recall Retrieval

It is not uncommon in high-recall retrieval tasks to make use of supervised learning as a means of separating relevant from non-relevant documents [136]. Supervised learning uses a set of documents, which are manually assessed by a human as to their relevance, to train a machine learning algorithm, creating a classifier. This classifier is then used to classify (or rank) the documents in the collection based upon their likelihood of relevance to a particular information need. The experiments in this chapter use passive supervised learning, where the machine learning algorithm has no influence on the selection of training documents.

It is often the case, in high-recall tasks, that the opinion of a single human ("the authority") is the final arbiter of relevance. For example, the authority may be a senior lawyer in eDisocvery, or a patient examiner in the intellectual property domain, or a senior researcher in the case of diagnostic medicine. Regardless of the particular domain, authoritative opinions on relevance may be impossible to obtain or may incur substantial monetary and temporal costs for even small document sets.

As discussed in Chapter 2, several prior studies [152, 163, 112] have looked into the effect that exchanging an authoritative opinion with that of a surrogate assessor can have on evaluation and supervised learning. Under certain experimental conditions, the results of such studies suggest that a surrogate can replace the authority with only a reasonable impact on the effectiveness of an algorithm and little impact on the comparative evaluation of two systems. Accordingly, it would not be unreasonable to view surrogates as cheaper and more convenient proxies for the authority, whomever that may happen to be.

This opinion is not shared by all practitioners, especially in the eDiscovery domain. Gonsowski [68], writing in a trade publication, argued that errors in relevance assessments have the potential to be amplified by the machine learning algorithm with disastrous results. This has led [5] to the use of the label "garbage in, garbage out" when surrogate assessors train machine learning algorithms.

In this chapter, we test the hypothesis that surrogate assessors only appear to be inferior to the authority when training classifiers due to the fact that the authority is the one evaluating the resulting classifiers. That is, when the roles of surrogate and authority are swapped, similar differences in performance appear. We also test the hypothesis that diversifying the conception of relevance behind a set of training assessments will improve the trained classifier's performance relative to the non-diversified conception of relevance. The following subsection outlines the experiments conducted in this chapter and the general experimental methodology.

3.0.1 Overview of Experiments

To test our hypotheses, our general experimental protocol proceeded as follows: for multiple sets of assessments, using only the intersection of documents judged and some randomly selected documents, train a Support Vector Machine (SVM) for each of these sets and then classify the remainder of the document collection using 10-fold cross-validation. In turn, treat each of the assessment sets as the authority and evaluate the performance of each learner with respect to the amount of review the authority would have to do to achieve a particular level of recall, which is calculated using the resulting ranked list created by the cross-validation process. We used several test collections to test our primary and secondary hypotheses, specific experimental details of which can be found in Section 3.1 and in subsequent sections dealing with those collections. To validate that our experimental set-up was functioning correctly, we replicated the earlier study of Webber and Pickens [163], which compared secondary sets of assessments to the gold standard assessments in a similar manner to the protocol outlined above, the major exception being that the roles of authority and surrogate were not swapped. A brief discussion of our replication and the differences in experimental set-up is presented in Section 3.1.1.

Section 3.2 describes our experiments using the TREC-4 test collection and the supplementary assessments which were previously used by Voorhees [152] in a study of the effect that using different assessors has on IR evaluation. This data has three sets of assessments, the official and two sets of surrogate assessments, which we used to train three SVMs using cross-validation. We then took each assessor, in turn, as the authority and evaluated the

minimum rank cutoff (i.e., the minimum amount of review effort) to achieve a desired level of recall for each of the three SVMs.

To test our secondary hypothesis that diversity in the conception of relevance improves ranking performance, we explored two potential diversification strategies. For both strategies, we created three new surrogate assessors, each corresponding to a pair of the three original assessors, which conceptually represented pairs of assessors working together to judge the training set. In the first strategy, we created a merged set for each surrogate pair by randomly dividing the training set in half and used the judgments from each surrogate to judge one of the halves. Using these merged sets, we trained three additional SVMs and evaluated the resulting classifier by treating the third assessor as the authority.

For the second strategy, the union of each surrogate pair was used to create a new union set of assessments, such that a document was deemed relevant if either surrogate deemed it so. As with the merged surrogates, an SVM was trained on each union surrogate and evaluated with the third assessor as authority. Such an approach doubles the amount of surrogate assessing effort but is reasonable given that a surrogate assessor is ostensibly much cheaper than the authority. Accordingly, having two surrogates judge the same pool of documents should *in practice* cost less than having the authority judge the pool once.

An alternate strategy for diversification of relevance would be to look at assessments where the assessor could have been more or less liberal/conservative in their conception of relevance. As detailed in Section 3.3, the University of Waterloo, during the course of participating in TREC-6, rendered their own set of relevance assessments using a ternary judgment scale of relevant, not relevant, and "iffy," which denotes documents which they believed to be of borderline relevance [38]. Accordingly, if we treat only those documents that Waterloo deemed relevant to be relevant, we label the resulting set to be "conservative" in its conception of relevance. Conversely, when we treat documents labelled "iffy" to be relevant, in addition to those deemed "relevant," we produce what we call a "liberal" conception of relevance.

Using these assessments and the official NIST assessor, we train three SVM classifiers using 10-fold cross-validation and evaluated each classifier when each assessor was deemed to be the authority. If the idea of "garbage in, garbage out" were true then the conservative Waterloo assessor should produce a classifier superior to the liberal Waterloo assessor. In particular, the liberally trained classifier should be substantially worse than the conservatively trained classifier when the conservative Waterloo assessor is the authority, the result of training the liberal classifier on many false positive ("garbage in") examples, which should produce a worse ranking of the collection ("garbage out"). As we will see, this turns out not to be the case.

Section 3.4 concludes our series of experiments by looking into the applicability of our results in the legal domain as captured in the TREC 2009 Legal track. We created the following 4 sets of assessments: the initial (pre-appeals) TREC assessments created by volunteer law students and contract lawyers; the assessments generated by the University of Waterloo during participation; a combination of Waterloo and initial TREC assessments, representing a diversified conception of relevance; and, the final post-appeals assessments that were the result of adjudication by volunteer senior lawyers (the "topic authorities"). Using each of these assessments, we trained an SVM using 10-fold cross-validation and evaluated its ranking capability with respect to the final set of post-appeals assessments only to reflect the nature of the task and test collection. Based upon these results, a small follow-up experiment was conducted to test the hypothesis that augmenting the initial training set with additional documents judged by Waterloo, and thus further increasing diversity, would reduce the effort needed by the authority to reach a desired level of recall.

We conclude this chapter with a discussion of our findings, their implications, the limitations of the experiments, and some potential extensions in Section 3.5.

3.1 Experimental Methodology

As our experiments follow the Cranfield paradigm, we begin by describing the test collections, general experimental set-up, and evaluation methodology used throughout this chapter. Also included is a brief discussion of the validation of our experimental set-up through a replication of the results of a previous study by Webber and Pickens [163], which served as the inspiration for these experiments.

Table 3.1 offers summary statistics, including the number of documents and the average prevalence of topics, for each of the test collections used in this chapter. The TREC-4 and TREC-6 test collections consisted of various sets of newswire documents and speech transcripts collected during the late 1980s and early 1990s from a variety of sources, which include the Associated Press, the *L.A. Times*, the *Financial Times*, and the Congressional Record. Our first and second set of experiments makes use of secondary data from TREC-4 and TREC-6 that had been previously used by Ellen Voorhees [152] to examine the effect of different assessors on ad hoc IR evaluation.

The Legal 2009 collection consisted of the Enron email collection version 1, previously described in Chapter 2, and the four topics for which the University of Waterloo rendered assessments [52] (topics 201, 202, 204, and 207) using a combination of interactive search and judging and active learning.

Our experiments make use of assessments rendered by several groups of independent assessors to train our machine learning algorithms and determine their effectiveness. For the experiments using the TREC-4 (Section 3.2) and TREC-6 (Section 3.3), the secondary assessing data was incomplete with respect to the official NIST set of judged documents. In the case of TREC-4, the secondary assessments were a subset of the official judging pool such that at most 200 relevant and 200 non-relevant documents (as previously determined by the primary assessor) were judged by two alternate NIST assessors for each topic. For TREC-6, the secondary assessments were rendered by the University of Waterloo using an interactive search and judging system during the course of their participation [38] and thus form a set of judgments that partially overlaps with the official NIST pool. Note that the Waterloo and NIST sets are not completely independent, since Waterloo submitted their judged relevant documents (as well as some unjudged documents) as part of their participation.

Accordingly, our experiments only consider documents for which there were complete sets of assessments (i.e., all assessors judged these documents) for training and evaluation purposes. This means that any document which was not judged by all assessors was deemed not relevant for these experiments. This was done to control for the fact that some assessors rendered many more judgments than others and would then have larger training and evaluation sets, which might non-trivially affect the classifiers being produced and the resultant evaluation. This aspect is further discussed in Section 3.5.3.

This restriction was not maintained for our experiments using TREC 2009 Legal track data (Section 3.4), due to the established roles of initial assessor, topic authority, and independent assessor (the University of Waterloo) in the original task. For these experiments, only the documents present in the official TREC judging pool had their judgments used for the purposes of training and evaluation. This restriction limits the number of Waterloo assessments available for use but not those of the initial assessor or topic authority, as they ostensibly assessed the entire judging pool. However, the University of Waterloo assessments do not form a subset of the judging pool. Accordingly, we investigated two solutions to the situation when a Waterloo assessment was not present in the judging pool: (1) deem the document to be "not relevant," which corresponds to the fact that the document was not found by the independent assessor and thus would be deemed "not relevant" anyway; (2) deem the document to be the same as the initial assessment, which might be considered diversification of relevance.[1] Documents judged by the independent assessor (Waterloo) but outside of the official pool were deemed not relevant, which is the same treatment as

[1]One might argue that a third option would be to deem such unjudged documents as "relevant," but this would be an artificial and contradictory relevance, as Waterloo did not return the document and thus would have considered it "not relevant" themselves.

Corpus	Documents	Min. Prevalence	Avg. Prevalence	Max. Prevalence
TREC 4	713,049	0.002%	0.017%	0.028%
TREC 6	556,077	0.0002%	0.0097%	0.057%
Legal 2009	723,386	0.58%	0.72%	0.94%

Table 3.1: Summary statistics of all three corpora. The official NIST assessments were used as the gold standard for these statistics.

was applied in the TREC-4 and TREC-6 experiments.

A standard TREC judging pool is formed primarily from the top-ranked documents contributed by a set of runs. Accordingly, these top-ranked documents may have elements that make them appear relevant to a particular algorithm. Training a machine learning algorithm solely on these documents would result in a classifier that is trained to "split the hairs" of relevance among such documents. Keeping in mind that the goal of TAR is to separate relevant (responsive) from non-relevant documents in a particular document collection, we included an additional random sample of 1,000 documents in the training sets and presumptively labelled them not relevant. This should not introduce many, if any, false negatives, [2] due to the relatively low prevalence of the corpora used. In adding these random documents, our intent was to broaden the nature of the documents in the training set and allow the resultant classifier to more aptly distinguish relevance of documents outside the original TREC judging pool.

We used a standard machine learning package, SVM^{light}[87], with default parameters, for the purposes of training and ranking the collections. SVM^{light} is a commonly used package that has been used extensively in IR literature and is sufficient as a reputable and replicable baseline. For all collections, we performed minimal feature engineering: we began by disregarding non-alphabetic words that were stemmed using Porter and applied case-folding. Real-valued scores were generated for the features using the following tf-idf term scoring function:

$$(1 + log(\text{term frequency in document})) * log(\frac{\text{Corpus size}}{\text{\# of documents term appears in}})$$

While the experiments in this chapter attempt to investigate the impact of surrogate assessments on high-recall retrieval, we do not formulate these experiments as a high-recall

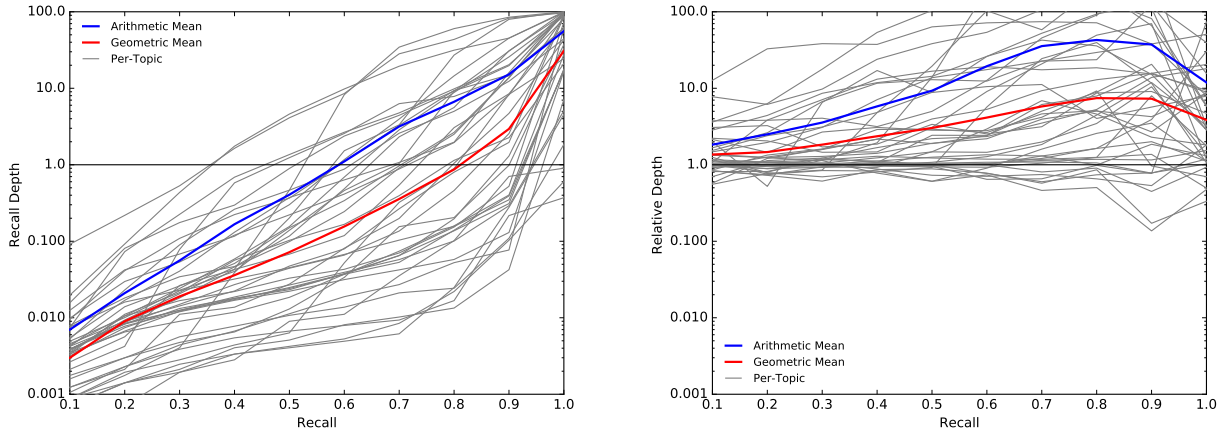
[2] A false negative would result when one of documents selected was judged as relevant by one of the surrogates but was not included in the training set, since it was not judged by all surrogates.

task but as a classic passive supervised learning task. This means that (10-fold) cross-validation can be used to simulate an independent evaluation set, since our training and evaluation sets are not disjoint. Documents appearing in both training and evaluation sets were evenly distributed (in a random fashion) among 10 folds. Recall that a training set consists of the multiply assessed documents as well as the randomly sampled documents. Documents appearing only in the evaluation set (i.e., those documents that were not judged by all assessors and were not in the random sample) were also evenly distributed among these 10 folds. For a particular fold, the learner was trained on the union of the 9 other folds and then scored each document in the fold that was not used in training. A ranking was induced for the entire corpus by sorting the documents based upon their classifier scores. We first tested our experimental methodology by replicating the Webber and Pickens study [163]. The study itself is described in Chapter 2 and our replication is discussed below in Section 3.1.1.

Our chosen evaluation measure is *recall depth*, following Webber and Pickens [163], which is the size of the shortest prefix of the ranking that achieved a particular level of recall expressed as a fraction of the size of the corpus. In essence recall depth is just the transposition of a gain curve (effort is measured as a function of recall). The benefit to such a calculation is that it allows us to easily compute the relative (recall) depth of one classifier with respect to another. Relative depth is the effort of one classifier (expressed as a percentage of the corpus) divided by the effort of another classifier (expressed as a percentage of the corpus) to achieve the same recall. For example, if one classifier required 10% of the corpus to achieve 50% recall while another only required 5%, then the first classifier has a relative depth of 2. This tells us in a single measure how much additional (or how much less) effort is needed to achieve the same recall when comparing two classifiers.

For graphical results, we computed for each method the average, over all topics, of the log-transformed recall depth. To aid in direct comparison of classifier pairs, we show the average of log-transformed relative depth. In both cases, the resulting average is equivalent to the geometric mean which better captures the relative rather than the absolute differences between the various assessors and their resulting classifiers.

The use of the geometric mean is less important for recall depth, since each topic's recall depth is bounded from 0 to 1, but is extremely important for relative depth which is effectively unbounded (e.g., it can range from 2 times the effort for one topic to 1,000 times the effort for another) which has the potential to unduly influence the arithmetic average. Figures 3.1 and 3.2 provide illustrative comparisons of the arithmetic and geometric means of recall depth and relative depth for a subset of the experiments described in Section 3.2. It is immediately clear, especially for relative recall depth, that the arithmetic mean is skewed by the outlier topics, while the geometric mean appears to be more inline with the



(a) Recall Depth

(b) Relative Depth

Figure 3.1: Comparison of arithmetic and geometric means with respect to per topic results for decall depth and relative depth when trained by J2 and evaluated by J1 in Section 3.2.

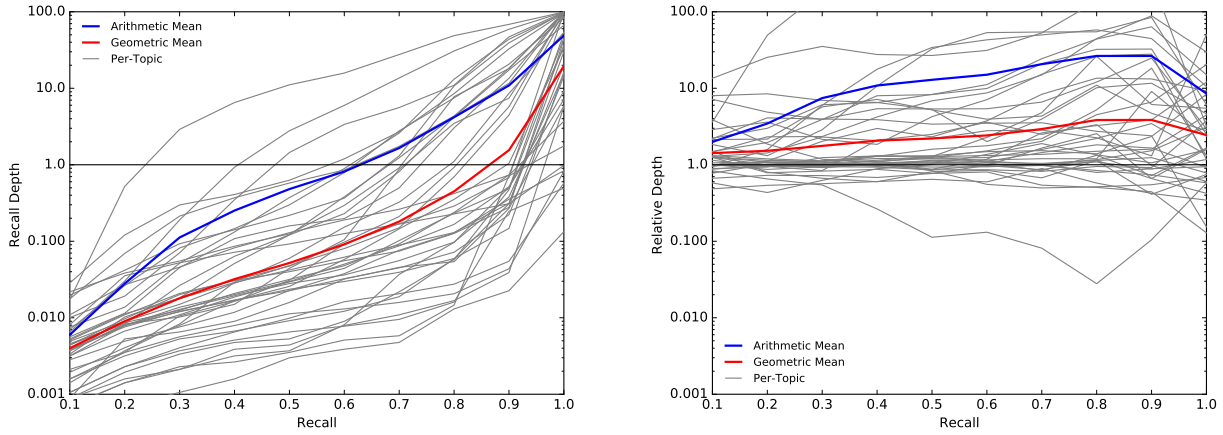
central tendency of the non-outlier topics. Based upon these examples and on the raw data, we believe that it is justified to use the geometric mean.

It is worth noting that we chose not to investigate the harmonic mean since, when dealing with positive data points as we do here, the harmonic mean is the smallest of the Pythagorean means and, while mitigating the effect of large outliers, is more so affected by small outliers tends towards those. Thus, we would expect the harmonic mean to understate differences in our comparisons. Accordingly, the choice of geometric mean is the most appropriate for our experimental setting.

For tabular results, we report recall depth for 75% recall with 95% confidence intervals, which is a previously reported recall target [8, 40, 163] and may be construed as a de facto standard. Significance of the surrogate-trained classifiers relative to the authority-trained classifier was computed by applying a t-test to the log-transformed difference. Where necessary, we use † to denote $p < 0.05$; and, ‡ to denote $p < 0.0001$.

3.1.1 Webber and Pickens Replication

To validate our experimental set-up, we replicated the Webber and Pickens study [163], described previously. The experiments described in this chapter were designed to follow the same model as that of the Webber and Pickens study. The key differences include: a



(a) Recall Depth

(b) Relative Depth

Figure 3.2: Comparison of arithmetic and geometric means with respect to per-topic results for recall depth and relative depth when trained by J3 and evaluated by J1 in Section 3.2.

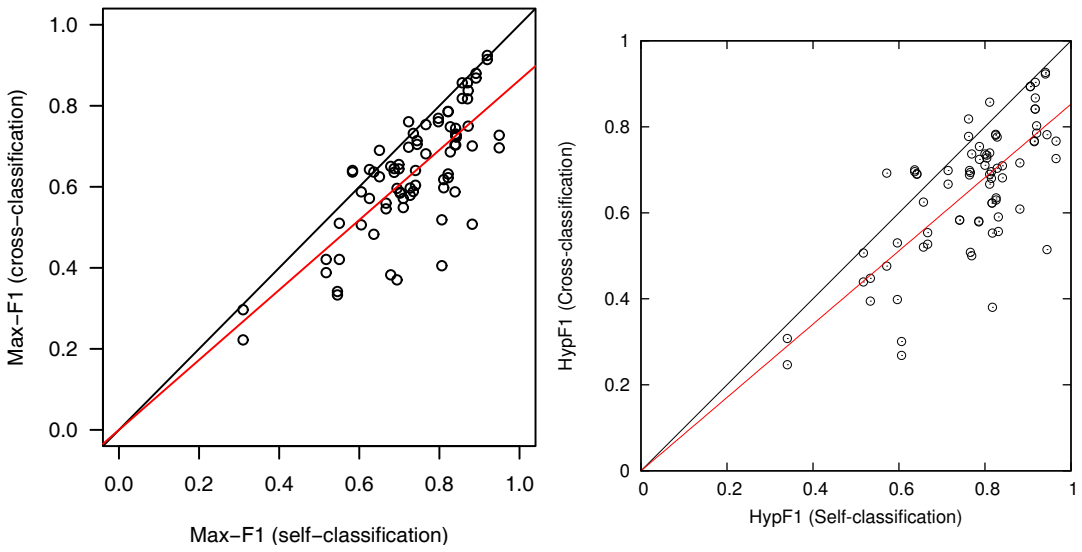
version of TREC-4 collection restricted solely to documents from the Associated Press that were judged by all three assessors for each topic; feature engineering was based upon this restricted collection; differences in the feature engineering (different tf-idf generation;[3], stop word removal, retention of all words including numbers, different stemming algorithm); and a different machine learning package (LibSVM [32] versus svm^{light}). The result is that each topic had its own associated specific document collection and set of features. Webber and Pickens considered the initial (gold standard) NIST assessment as the authority and either secondary assessment as surrogates. Thus, unlike our experiments they did not swap the roles of surrogate and authority.

Webber and Pickens used, as their primary reported evaluation measure, differences in the hypothetical F1 (HypF1), which they called Max-F1. In addition to hypothetical F1, Webber and Pickens also computed the 75% recall depth for each experiment.

Figures 3.3 and 3.4 juxtapose the original Webber and Pickens results with our replication of their results with respect to hypothetical F1 and 75% recall depth. Note that the graphs are labelled in the style of Webber and Pickens, who referred to authoritative training as "self-classification" and surrogate training as "cross-classification,"[4] which

[3]They did not specify their formula.

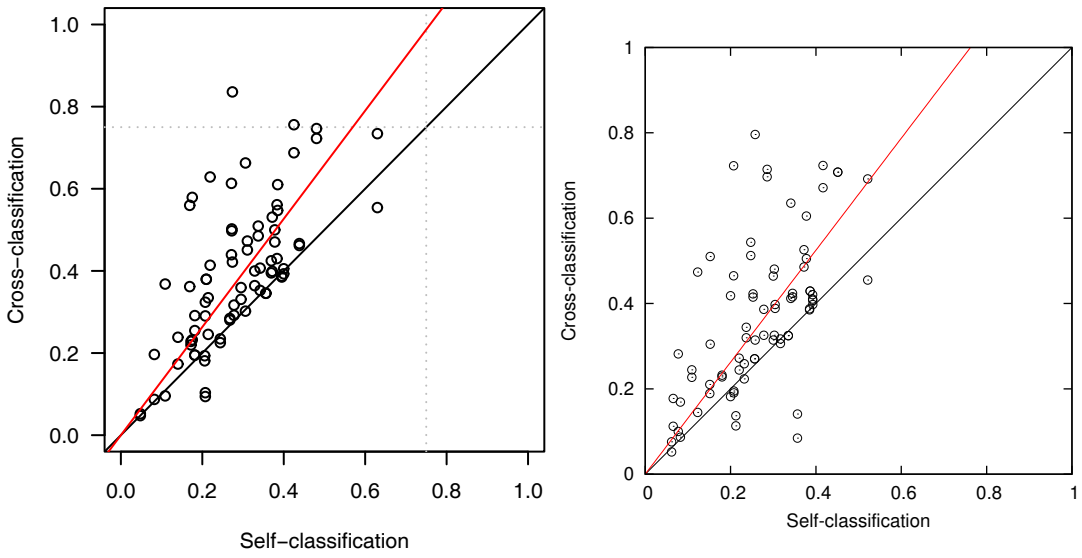
[4]Self-classification, because the training and evaluation assessor was the authority. Cross-classification, because the training assessor was a surrogate and the evaluation assessor was the authority.



(a) Webber and Pickens [163]

(b) Replication

Figure 3.3: Hypothetical F1 scores for authoritatively trained vs. surrogate-trained ranking as evaluated using the authoritative assessor, for the original Webber and Pickens study [163] and our replication.



(a) Webber and Pickens [163] (b) Replication

Figure 3.4: Comparison of 75% recall depth for authoritative training vs. surrogate training for the original Webber and Pickens study [163] and our replication.

we continue to use as an aid in comparison. From this we can see that, generally, our replication does indeed appear to have replicated the results of Webber and Pickens.

Furthermore, Table 3.2 recasts the Webber and Pickens results into a format consistent with our experimental set-up and includes our replication of their results. Their results fail within the confidence intervals of our own, and so, we can say that we have successfully replicated their results as best as we are able and within the bounds of chance. This allows us to conclude that our experimental set-up is valid.

3.2 Independent Judgments

This section describes our first set of experiments investigating the hypothesis that the apparent superiority of authoritative training for machine learning is a result of its being the evaluating assessor rather than of any inherent superiority in assessing ability. We also tested various diversification strategies for training and whether or not they improve upon single surrogate assessors.

	Training Assessor		
Metric	Authority	Surrogate Avg	
Hyp F1	0.738	0.629	
75RD	-	25% greater than Primary	
Metric	Authority	Surrogate 1	Surrogate 2
Hyp F1	0.769 (0.726 - 0.812)	0.651 (0.602 - 0.700)	66.15 (0.613 - 0.710)
75RD	0.266 (0.231 - 0.302)	0.366 (0.316 - 0.416)	0.369 (0.305 - 0.433)

Table 3.2: Webber and Pickens' results recast in our evaluation framework with our replication results. Confidence intervals, where reported, are at 95% confidence level.

These experiments were conducted using the judgments rendered for the TREC-4 ad hoc task [73], along with the two additional assessment sets of up to 200 relevant documents and up to 200 non-relevant documents for each topic. These documents were selected as subset from the official gold standard assessments and were used in several experiments regarding IR evaluation measure stability [152]. Voorhees reports that some assessors may have assessed more topics than others and may have functioned as the first surrogate for some topics and the second surrogate for others. This is due to the fact that these assessors were arbitrarily assigned to topics on a first-come, first-served basis with respect to any other assessing tasks that they were assigned. Accordingly, the two additional sets of assessments may have overlapped in assessors, but the assessors were independent for any single topic.[5] For our purposes, we label the three sets as J1, J2, and J3, where J1 is the subset of the assessments used for the official TREC evaluation. The experiments in this section treated these three sets equally and used each as the "authority" or gold standard, while treating the others as surrogates. To maintain consistency with the study by Webber and Pickens [163] and our replication, we considered only those topics for which each of the three assessors found at least 8 relevant documents. The goal of such a restriction was to mitigate excessive variance that could be caused by very low prevalence topics and, in the case of one topic, to elide issues with J2's finding no documents to be relevant.

For experimental purposes, J1, J2, and J3 are each considered to reflect a single, independent conception of relevance, though we acknowledge that this is not strictly true given the original assessment process.[6] We have proposed that a more varied conception of relevance would yield a more effective training. To test this hypothesis, we began by

[5]Thus, it is likely the case that the same assessor judged different topics under each of the three possible assessors at different times.

[6]Namely, J2 and J3 judged documents were selected based upon J1's assessments, but neither J1 or J2 were explicitly aware of those assessments.

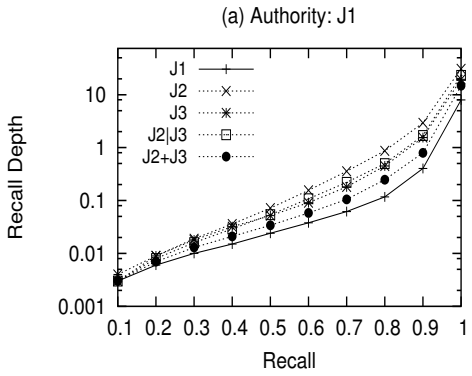
splitting the training set in half randomly and, for each pair of assessors, assigning one of these halves to each of the pair and merging the combined relevance assessments. These new assessors are denoted by J1|J2, J1|J3, and J2|J3, and referred to as merged surrogates for convenience. A classifier was trained by each of the merged surrogates and was then evaluated, in turn, by J3, J2, and J1, respectively. It is worth noting that the effort required for each of these pseudo-assessors is, in total, equal to that of a single assessor; we have merely split the work between two individuals. These pseudo-assessors are intended to reflect a more diverse conception of relevance, since no attempt was made to reconcile any potential disagreements or inconsistencies in the judging behaviours of the merged surrogates.

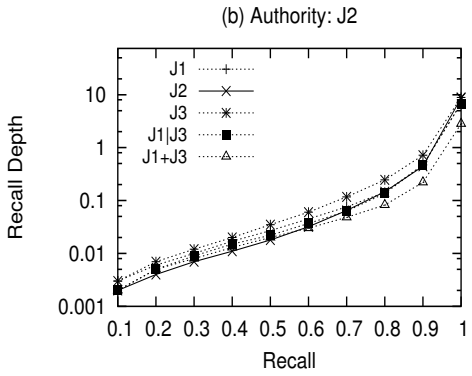
We have hypothesized that a more liberal interpretation of relevance is more effective for training than a conservative one, by taking the union of each pair of surrogates, denoted by J1+J2, J1+J3, and J2+J3. By union, we mean that a document (in the training set) is relevant if either constituent assessor has deemed it to be relevant. This introduces no additional documents to the training set but does double the effort, as we have two assessments per document.

3.2.1 Results

Based upon the results of Figures 3.5 and 3.6, which plot the recall depth and relative depth of all experimental conditions, we can easily see that training a classifier with a single-surrogate assessor is generally inferior to training by the authoritative assessor. In several cases, there is substantially more effort required by a single-surrogate to achieve the same level of recall as the authority. This is most evident at high levels of recall, which is our particular area of interest. According to 75% recall depth (Table 3.3), the difference in percentage of corpus reviewed is significant when J1 and J3 are used as the authority. This provides further evidence that there are meaningful differences in conceptions of relevance, regardless of the choice of authority.

In contrast, when J2 is the authority there does not appear to be any substantive difference between surrogate and authority with respect to 75% recall depth. Indeed, Figures 3.5 and 3.6 also indicate that the J1-trained classifier is better at ranking the collection than the J2-trained classifier (i.e., the authority). While it is possible that this has occurred due to chance, it may also be a side effect of the assessment process. Remember that documents assessed by J2 (and J3) were a subset of those assessed by J1 (the NIST official assessor), for which prevalence was likely much higher than for the official pool. This higher prevalence would be due to the enforced limits on the number of non-relevant documents in the subset. Furthermore, it has been observed that prevalence and





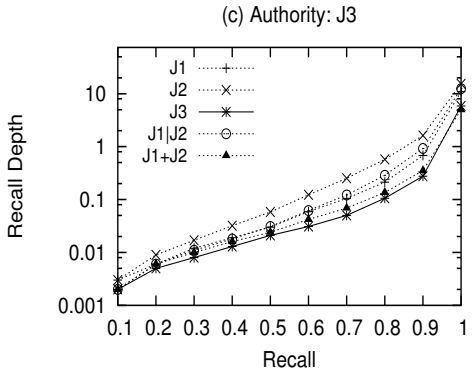
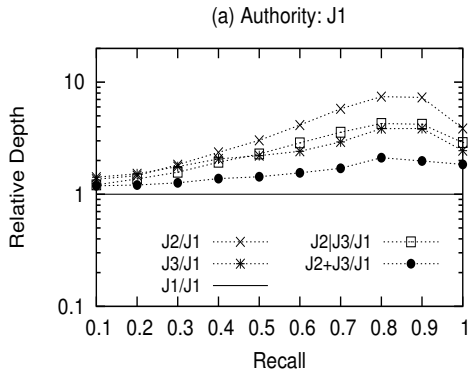
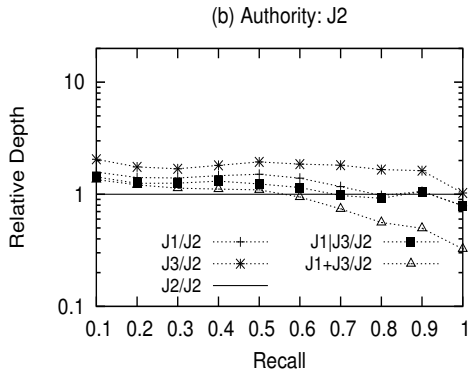


Figure 3.5: Recall depth plots for the TREC-4 experiments, using (a) JI, (b) J2, and (c) J3 as the authority.





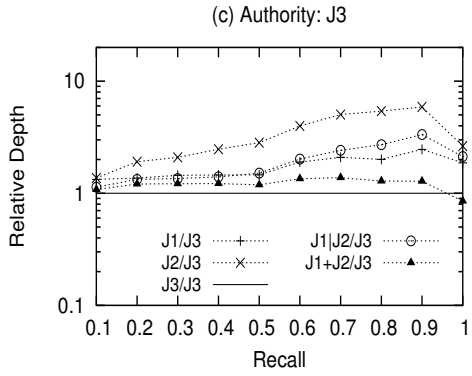


Figure 3.6: Relative recall depth plots for the TREC-4 experiments, using (a) J1, (b) J2, and (c) J3, as the authority.

recall have an inverse relationship [137], which means that J2, judging a higher prevalence pool, may have been more conservative than J1, which having a lower prevalence set of documents, was more liberal. This hypothesis is investigated from a more standard high-recall setting in Chapter 6 where we do find such evidence. Moreover, the raw numbers of documents judged relevant on average would lend credence to such a claim: J1 judged 118.1 documents relevant on average while J2 judged 73.2 documents relevant. Note that this should merely be taken as a plausible explanation of the observed phenomenon, not as a definitive answer. It is unlikely, given the lack of record keeping with respect to the supplementary judging process, that a definitive answer could be given.

Improving classifier effectiveness by diversifying training relevance appears to have had some success (Figure 3.5, Figure 3.6, and Table 3.3) with the merged surrogates generally achieving similar effectiveness to that of the better-performing constituent surrogate. Occasionally, the performance of a merged surrogate exceeds both, but we would caution that drawing conclusions from a weak trend, as observed here, may spell misfortune in the future. It is worth keeping in mind that if this were actually implemented, we might see additional performance increases, due to reduced fatigue on assessors, but this was not something we could easily incorporate into these experiments, due to the pre-existing nature of the assessments. Additional experiments that test relevance assessment using this strategy would need to be conducted to verify such a hypothesis.

More so than expected, the union-trained classifiers materially improved upon either constituent surrogate—so much so that the difference is significant with $p<0.01$. The union surrogates appear to save the authority more assessing effort than the doubled training cost signifies. Such a result indicates that a liberal assessing policy may, in general, be more suitable for training classifiers to identify relevant material than a more conservative approach. However, it is uncertain whether this is the result purely of a more liberal interpretation or of the constituent surrogates' "making up" for the mistakes of the other. The TREC-6 experiments (described in the following section) control for this more explicitly by using a single assessor and flipping the relevance of borderline documents.

3.3 Liberal Assessment

This section explores our experiments testing the hypothesis that using a more liberal conception of relevance for training will yield a superior classifier (higher recall with less effort) than will a more conservative when evaluated with respect to the authoritative assessor. We also tested whether this more liberal conception of relevance produces a superior classifier even when the conservative conception is taken to be the authority. This

Authority	Training		
	J1	J2	J3
J1	0.082% (0.058 - 0.115)	0.542%‡ (0.254 - 1.156)	0.284%‡ (0.139 - 0.584)
J2	0.103% (0.061 - 0.174)	0.087% (0.051 - 0.149)	0.161% (0.083 - 0.312)
J3	0.146%† (0.077 - 0.278)	0.359%‡ (0.162 - 0.797)	0.066% (0.044 - 0.101)
	J1|J2	J1|J3	J2|J3
J1	-	-	0.321%‡ (0.162 - 0.636)
J2	-	0.092% (0.059 - 0.143)	-
J3	0.182%† (0.096 - 0.346)	-	-
	J1+J2	J1+J3	J2+J3
J1	-	-	0.094%† (0.054 - 0.164)
J2	-	0.062%† (0.043 - 0.091)	-
J3	0.094% (0.054 - 0.164)	-	-

Table 3.3: 75% recall depth values for the TREC-4 experiments, with 95% confidence intervals. Significance is determined by comparing surrogate-trained classifiers to the authority-trained classifier with a paired t-test. († denotes $p < 0.05$; ‡ denotes $p < 0.0001$.)

second outcome tested the idea of "garbage in, garbage out" by intentionally introducing false positives into the training set of the liberal assessor.

During the course of participating in the TREC-6 ad hoc task [150], the University of Waterloo rendered their own set of manual judgments through a process of interactive search and judging [38]. In addition to the standard relevance categories of relevant and not relevant, the University of Waterloo used a third category, "iffy," to denote documents which they believed were of borderline relevance. In essence, these "iffy" documents are those for which Waterloo assessors were uncertain of the true relevance.

In a previous study by Voorhees [152], the "iffy" category was treated as not relevant for the sake of simplicity. The results of the TREC-4 have provided evidence that a more liberal conception of relevance, when used for training, would result in a superior classifier than a more conservative conception. We can simulate these two scenarios by varying how the "iffy" category is treated from a training perspective (i.e., treating the category as relevant or not relevant). When the "iffy" category is treated as not relevant, we believe that this reflectd a more "conservative" approach since it permits a greater number of false negatives. Such a conservative approach may stem from a belief in the "garbage in, garbage out" policy, where the goal is to limit false positives.

On the other hand, when we treat "iffy" as relevant, we may be approximating a more

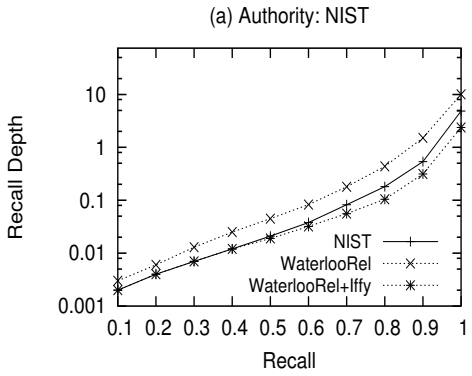
"liberal" approach, since we are allowing the possibility of false positives. This approach might be seen as trying to leverage the features of the "iffy" documents that make them appear potentially relevant to train a better classifier (i.e., to have more evidence/weight for relevant [looking] document features). We labelled the "conservative" approach WaterlooRel and the "liberal" approach WaterlooRel+Iffy. Our baseline set of assessments was simply the official NIST assessments. Recall that for the experiments in this section, we used only documents judged by NIST and Waterloo (plus the random sample), and documents outside this intersection were uniformly treated as not relevant.

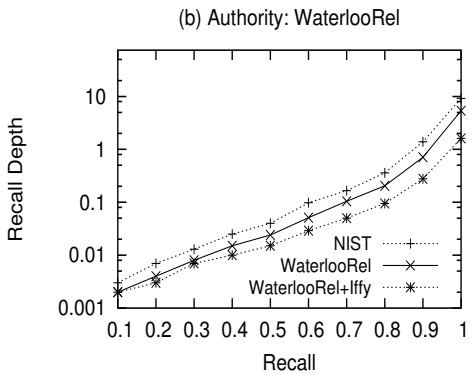
Following our experimental protocol, we used the official NIST assessments to compare the effect of training a classifier on each of: the official NIST assessments; the WaterlooRel assessments; and, the WaterlooRel+Iffy assessments. This single condition explores our initial hypothesis regarding liberal versus conservative interpretations of relevance. For continuity with our other experiments, we also swapped the roles of authority among the two sets of Waterloo assessments. While the two Waterloo assessment sets are not independent of each other, evaluating with respect to each other may provide additional insight into the "garbage in, garbage out" phenomenon and what effect false positives can have on training.

3.3.1 Results

When comparing liberal versus conservative conceptions of relevance for training, Figures 3.7a and 3.3a clearly show that our formulation of a "liberal" surrogate is superior to the "conservative" surrogate. WaterlooRel+Iffy is able to substantially improve upon WaterlooRel with respect to NIST and is competitive with NIST itself. In particular, our results show that the difference between NIST and WaterlooRel+Iffy is statistically indistinguishable at 75% recall depth (Table 3.4). What we can conclude from this is that the liberally trained classifier is equivalent to the authority in this case, and is far superior to the conservatively trained classifier, at all recall levels.

When NIST was exchanged as the authority with either Waterloo surrogate, we observed results which were consistent with the results of Section 3.2. Interestingly, the conservative Waterloo surrogate was easier for NIST to predict than the liberal surrogate. It is not immediately apparent why this may be the case. One might assume that it results from a phenomenon similar to what was observed for J1 and J2 (i.e., J1 trained more documents that looked relevant as relevant). But the fact that NIST and WaterlooRel judged on average approximately the same number of documents to be relevant (53.98 and 56.24, respectively) would seem to contradict such a hypothesis. It may be the case that there are





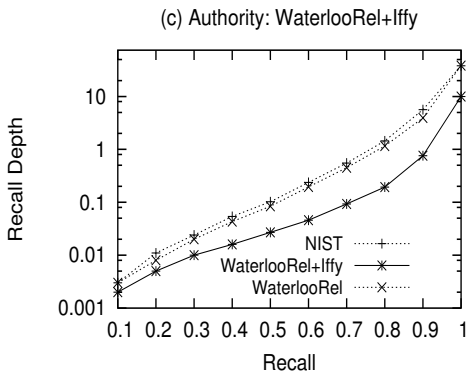
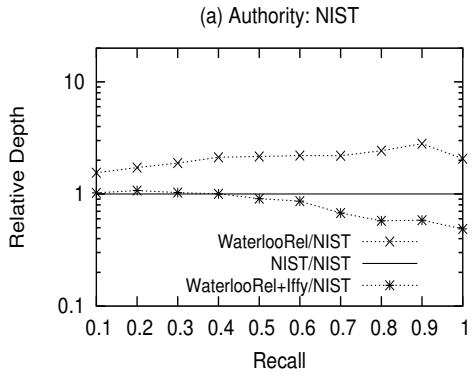
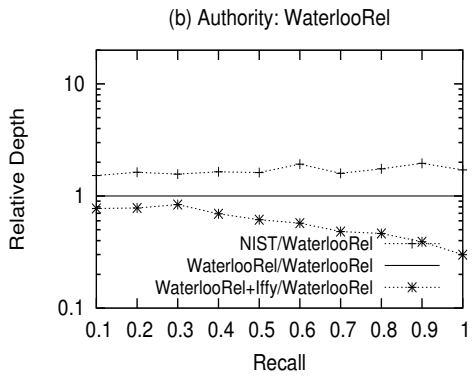


Figure 3.7: Recall depth plots for the TREC-6 experiments, using classifiers trained by each surrogate, and evaluated by each authority.





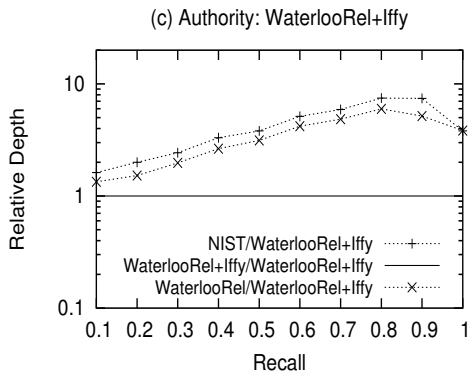


Figure 3.8: Relative recall depth plots for TREC-6 experiments, using classifiers trained by each surrogate, and evaluated by each authority.

important differences in *what* documents each surrogate judged to be relevant in addition to the "liberalness" of the judgments being rendered. However, it is less surprising that NIST, which is less liberal than WaterlooRel+Iffy (82.6 relevant documents on average), would train an inferior classifier to the liberal surrogate under such reasoning.

Furthermore, while Table 3.4 does not report a significant difference between the NIST-classifier and the WaterlooRel-classifier when WaterlooRel is the authority, Figures 3.8 does show material differences in their performance, indicating that there were differences between the two assessors in their conceptions of relevance.

It was not necessarily anticipated that WaterlooRel+Iffy would be substantially and significantly superior to WaterlooRel. This result indicates that it may be preferable to have a more liberal conception of relevance when training a classifier regardless of whether or not the authority is more conservative. This result would strongly contradict the idea of "garbage in, garbage out" and may suggest that all "garbage in" is not equivalent. It is clear that more false positives are not necessarily bad from a training perspective.

Authority \ Training	NIST	WaterlooRel	WaterlooRel+Iffy
NIST	0.110% (0.065 - 0.185)	0.261%† (0.142 - 0.481)	0.072% (0.049 - 0.105)
WaterlooRel	0.244% (0.130 - 0.458)	0.152% (0.094 - 0.246)	0.068† (0.049 - 0.094)
WaterlooRel+Iffy	0.882%‡ (0.515 - 1.511)	0.699‡ (0.451 - 1.084)	0.129% (0.094 - 0.177)

Table 3.4: 75% recall depth values for the TREC-6 experiments for Waterloo- and NIST-trained classifiers, evaluated using NIST assessments, with 95% confidence intervals. Significance is shown relative to the NIST-trained classifier and is determined by a paired t-test. († denotes $p < 0.05$; ‡ denotes $p < 0.0001$.)

3.4 TREC 2009 Legal Track

In this section, we describe our experiments to extrapolate our previous results to a more "realistic" high-recall setting and domain. The context of these experiments is the interactive task at the TREC 2009 Legal track [76], which simulated a high-recall retrieval task that permitted human-in-the-loop algorithms and interaction with the authoritative assessor. We have previously described this task in Chapter 2, and do not reproduce a thorough description or summary of the task here. We highlight that for each topic, the judging pool was a stratified sample of the document collection (partially based upon participant

runs), rather than the standard depth pooling used historically at TREC. This sample was initially assessed by a group of volunteer assessors (law students and contract attorneys) under the direction of a TREC-designated *Topic Authority* (a volunteer senior lawyer). As part of the task, these initial assessments were provided to participants to facilitate an appeals process to refine and ensure the quality of the initial assessments. These appeals were handled by the Topic Authority who was the final arbiter of relevance for a topic. After the appeals process, these final assessments were used as the basis for final evaluation at the track.

The University of Waterloo conducted a manual, human-in-the-loop approach for the interactive task [52] that combined interactive search and judging with active learning. Resulting from their participation, a second independent set of assessments was generated for 4 topics (201, 202, 203, and 207). Which, on these topics and post appeals process, was the best performing submission with respect to precision and recall.

The TREC judging pool, due to its nature as a stratified sample, contained a large random sample of the document collection. As a result, a relatively small fraction (17.7%) of the official judging pool was judged by Waterloo; the remainder was excluded by Waterloo's retrieval algorithm as these documents were believed to be not relevant. To maintain consistency with the task, we choose to limit the training set to the judging pool, which meant that the remaining 82% of the pool would be deemed as not relevant by Waterloo. An alternate approach may, instead, be to supplement the Waterloo assessments with the initial assessments. This meant that for any document not judged by Waterloo, we used whatever the volunteer initial assessor judged that document to be.

Accordingly, we can create 4 sets of surrogates for training: the "Waterloo" surrogate, which deems excluded documents to be "not relevant"; the "Initial" surrogate, which uses the initial assessments generated for the judging pool; the "Waterloo w/ Initial" surrogate, which deems excluded documents to be of the same relevance as the "Initial" surrogate; and, the "Final" surrogate, which uses the official judging pool (i.e., the assessments after the appeals process was concluded). The Waterloo w/ Initial surrogate is then similar to the merged surrogates of Section 3.2 without balancing the effort of the constituent surrogates.

Using these 4 sets of surrogates, we trained 4 classifiers and evaluated them while treating the Final surrogate as the authority. We did not swap the authoritative assessor in these experiments since this task attempted to model a real-world scenario where a senior lawyer's assessment would be the final assessment (i.e., the Waterloo assessor and Topic Authority are not necessarily interchangeable or equivalent in this task model). Given that Waterloo relevance assessments exist for 4 topics, we did not compute average curves due

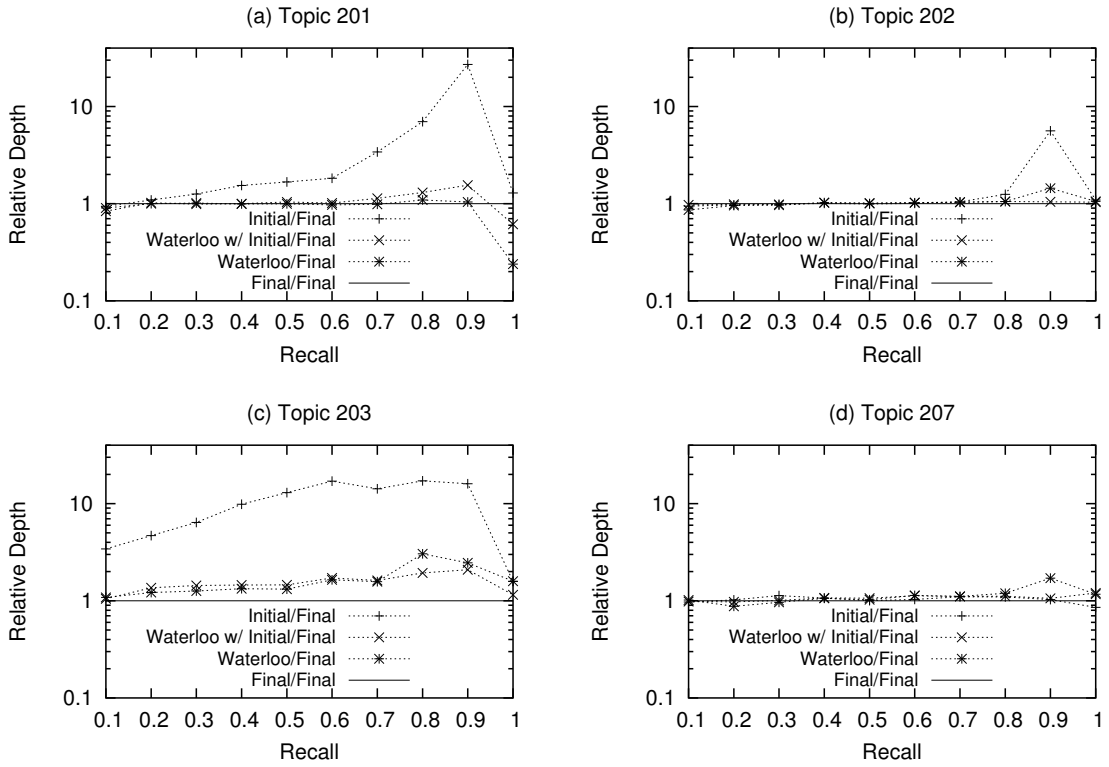


Figure 3.9: Relative recall depth plots for the TREC 2009 Legal experiments, using classifiers trained by each surrogate, and evaluated by the final assessments.

to a lack of utility in doing so. To maintain consistency with the previous experiments we calculated recall solely with respect to presence in the judging pool (i.e., we did not use the inclusion probabilities associated with stratified sampling to calculate recall), which differs from the official evaluation which made use of the inclusion probabilities. By not extrapolating recall to the entire collection, the recall values reported here do not accurately reflect each classifier's "true" recall. Previous work by William Webber [157] would suggest that, by not estimating collection-wide relevance, we have provided a more fair comparison, as no classifier is overly penalized for mistakes in assessing.

Topic	Initial	Waterloo	Waterloo w/ Initial	Final
201	1.056%	0.214%	0.254%	0.215%
202	1.005%	0.977%	0.993%	0.936%
203	6.542%	0.955%	0.816%	0.456%
207	1.314%	1.401%	1.324%	1.236%

Table 3.5: 75% recall depth values for the TREC 2009 Legal experiments, using classifiers trained by Waterloo and by Initial assessments, and evaluated using the Final assessments.

	Judging Pool		Full Corpus	
Topic	Precision	Recall	Precision	Recall
201	0.40	0.70	0.05	0.76
202	0.93	0.93	0.27	0.80
203	0.47	0.25	0.13	0.25
207	0.98	0.88	0.89	0.79

Table 3.6: Recall and precision of Initial assessments in the TREC 2009 Legal track judging pool versus the full corpus.

3.4.1 Results

Figure 3.9 reports relative depth for the 4 topics with respect to the Final surrogate. Recall depth plots are omitted for brevity and for the lack of additional information presented in such plots. While surrogate-trained classifiers appear to be inferior to the authority at high recall, we see material difference only for the Initial surrogate and topics 201 and 203. Due to stratified sampling, the Initial surrogate appears to have much higher precision and recall in the judging pool than in the full corpus, as depicted in Table 3.6. Such differences may account for the inconsistency of the Initial-trained classifiers across topics and likely stems from the use of stratified sampling to create the judging pool [76]. The Initial surrogate appears to have been more accurate for topics 202 and 207, which may have resulted in the better performance overall for these topics. Its performance may indicate that there were more meaningful mistakes in the labelling that may have contributed to this deficit. Similar conclusions can be drawn from the 75% recall depth values in Table 3.5.

When restricted to the judging pool for training, the Waterloo surrogate appears to identify approximately the same number of relevant documents as does the Final surrogate. As such, a firm conclusion as to whether or not the Waterloo surrogate would be liberal or not cannot be drawn. From 4 topics it is difficult to draw a solid conclusion as to the causes of the Initial-trained classifier's inconsistent behaviour. Whether or not these results are

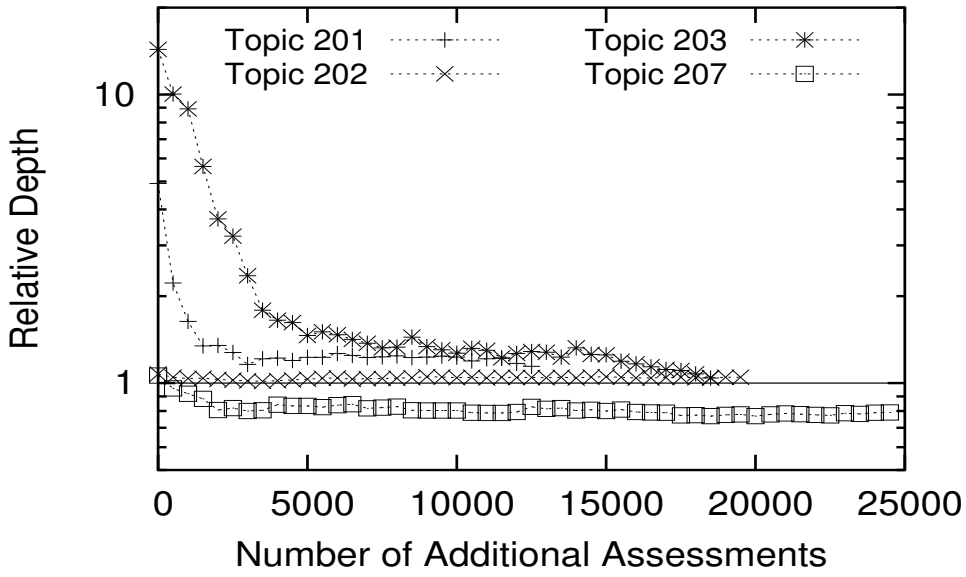


Figure 3.10: Per-topic 75% relative recall depth plots for the retrospective TREC 2009 Legal experiment, using classifiers trained on initial assessments, progressively augmented by Waterloo assessments, and evaluated using final assessments.

due to the legal setting being more sensitive to incorrect judgments than the previous ad hoc scenarios is uncertain. Further investigation is required to draw firm conclusions.

Both Waterloo surrogates appear to train equivalent classifiers regardless of the use of the Initial assessments. However, inclusion of the Initial assessments does appear to provide some additional utility for 3 of the 4 topics. In all cases, the combination of Waterloo and Initial surrogates is as good as or better than the Initial surrogate alone. A more accurate picture may come from viewing the Waterloo w/ Initial assessments as the Waterloo assessments being added to the Initial assessments (i.e., diversifying the Initial assessments). That is, if we replace less than 20% of the Initial assessments with Waterloo assessments, we get improved performance across the board. This provides further evidence that diversifying relevance assessments by merging surrogates may be a useful assessing strategy.

3.4.2 Interactive Training

The experiments on the 2009 Legal track data, described above, still follow a passive supervised learning set-up. To more accurately model the human-in-the-loop setting of the interactive task, we retrospectively conducted an initial study to further investigate the applicability of our results. In this supplemental exercise, we incrementally added batches of 500 Waterloo assessments, including documents outside the judging pool, to all of the Initial (pre-appeals) assessments and tracked improvement in relative depth. Ideally, we would have liked to add these batches of Waterloo assessments in the order that they were made, however, such logs have been lost to the sands of time, so batches were randomly selected (without replacement) from the Waterloo set.

Figure 3.10 shows 75% relative recall depth as a function of the additional number of Waterloo assessments. Given the previously reported results, the dramatic increases in performance with respect to topics 201 and 203 is not all that surprising but is good confirmation that diversification can help. A similar, but less dramatic, situation occurs for topic 207, which starts with a competitive classifier to the Final-trained classifier and then quickly outperforms it. There does not appear to be any additional benefit for topic 202, and perhaps, there is some small degradation in effectiveness, but such a result is not generalizable, in and of itself. Regardless of the topic, there does appear to be a point where inclusion of further Waterloo assessments hits a plateau in 75% recall depth but this point does not appear to have any obvious consistency between topics. The result of this supplementary experiment appears to indicate that controlled, incremental diversification of training assessments has the ability to improve classifier effectiveness.

3.5 Discussion

3.5.1 Whose Authority?

The experimental results presented in this chapter illustrate that it clearly matters who assesses relevance. This is particularly important when the assessments used to train a system are also those used to evaluate the output of the system. If we wish to convey meaningful information with effectiveness measures, we need to consider whom (i.e., what authority) these measures are computed with respect to.

Such an issue arose during one of the first cases in which a legal court ruled in favor of machine learning use in the electronic discovery process. The producing party's brief [8] asserted:

> Given that recall for linear review averages only 59.3%, [the responding party] proposes an acceptable recall criterion of 75%. In other words, predictive coding will conclude once the sampling program establishes that at least 75% of the relevant documents have been retrieved from the [responding party's electronically stored information] and are available to [them] for discovery purposes.

The linear review average was derived from an analysis of the 2009 Legal track's results by Grossman and Cormack [69], where the precision and recall of the initial assessments were evaluated with respect to the final authoritative assessments (i.e., after the appeals process). Part of these results are reported in Table 3.6. The 75% recall stopping criterion was determined by the producing party's own assessors (i.e., lawyers and paralegals), but it is likely that these reviewers were also involved in training the retrieval system used in the discovery process. Our results indicate that had an independent reviewer (e.g., from the requesting party, or a domain expert provided by the court) evaluated recall for the production, the recall value calculated could have been substantively lower.

In IR evaluation, it often necessary to deem one assessor's opinion to be "authoritative" in order to sidestep known uncertainties in the definition of relevance for assessment, which directly affects the computation of recall [162]. While it is a certainty that some assessors are more skilful and knowledgeable than others, even expert assessors have the potential to disagree on non-trivial numbers of assessments [15, 152]. The Voorhees study [152] suggests that an assessment effort by one expert assessor, when evaluated by a different expert, would be unlikely to achieve better than 65% recall and 65% precision. The assessors used in our TREC-4 and TREC-6 experiments were those used by Voorhees in her study, so we observed comparable levels of recall: with J1 (the official assessor) as authority, J2 and J3 achieved recall and precision of (52.9%, 80.8%) and (63.1%, 78.1%), respectively; when assessed by NIST, our Waterloo surrogates achieved recall and precision of (62.8%, 65.2%) and (86.6%, 50.0%), which correspond to the conservative and liberal surrogates, respectively.

In keeping with the Cranfield paradigm, it is only reasonable that we assume that any of the surrogates used in this chapter would have judged the entire corpus. While this is not strictly accurate, it is standard practice within the IR community and encompasses existent best practice for system evaluation. Our results clearly show that a classifier trained using surrogate assessments would achieve higher recall with less effort (i.e., a small fraction of the corpus) than the surrogate (and perhaps, the authority) alone. Consider our "worst" surrogate, J2, this surrogate's classifier achieved 55% recall at a depth of 0.1% of the corpus, on average, and achieved 75% recall at a depth of 0.542% of the corpus. Results such as these do not support the claim that the use of machine learning will "amplify" or

"magnify" inconsistencies and disagreements between the surrogate and the authoritative assessor. If we are to take anything away from these results, it is that machine learning has the potential to greatly expedite the assessing process.

This prolonged discussion should not be construed as an appeal to do away with authority. It is a meaningful endeavour to attempt to maximize recall while reducing effort with respect to an authority, which may be a mandated arbiter of truth or someone acting as a proxy for an unavailable authority (e.g., a judge or regulator). Using an authority appears to be the most expedient and consistent manner in which we can measure a system's effectiveness. It should be the case that if a system achieves high recall with minimal effort according to one independent assessor, it should perform comparably with respect to another independent assessor. If this is not the case, then we must seriously question the empirical results produced by such a system and examine how effective it is in reality.

3.5.2 Improving Surrogate Assessment

Our experimental results have consistently shown that using a surrogate instead of the authority for training can substantially impact the effort required to achieve high recall. When this difference between surrogate and authority is large, it is not attributable to chance ($p<0.05$ after Bonferoni correction). While we have seen several instances where the surrogate can produce a classifier that is competitive with, or better than, the authority, we cannot say that this will always be the case. There are differences in assessors, and such differences may at times be beneficial, while at other times they can be detrimental.

From the experiments in Section 3.2, we have seen that randomly dividing assessments between two surrogates achieves performance similar to the better of the constituent assessors with the potential to reduce mental fatigue and/or cognitive load on the surrogates. These results were replicated when we combined the Waterloo and Initial surrogates in our TREC 2009 Legal track experiments. Whether such observations extend to real scenarios requires further investigation and remains an open question.

On the other hand, taking the union of two sets of surrogate assessments can result in substantial improvements over both surrogates and approaches the effectiveness of the authority. We construed such an action as an attempt to increase the liberality of the surrogate assessments. When we explicitly labelled marginally relevant documents as relevant in Section 3.3, we saw that there was significant and substantial improvement over treating those marginally relevant documents as not relevant. This experiment explicitly tested whether intentionally taking a liberal interpretation of relevance was preferable to

a conservative one when training a classifier. Moreover, taking this liberal strategy materially improved performance over the authority in this particular experimental context. This is a singular result and should not be generalized to other contexts, where it may not be applicable (i.e., those different from our experimental setting). However, the result does indicate that it is possible to train a classifier that is as effective as one trained by the authority.

Furthermore, we have seen that training with this more liberal approach yields performance far superior to that of the conservative approach regardless of which assessor is taken to be the authority. Of particular importance is the fact that the classifier trained by the liberal surrogate was significantly and substantially better able to predict the conservative surrogate than was the conservatively-trained classifier. If such a result does not cast doubt upon the idea of "garbage in, garbage out," it is doubtful that any result would do so. We note that Cheng et al. [33], in experiments dealing with multiple review passes, found that the lower-quality (i.e., higher false positive rate) first-pass reviewer produced a better classifier than the higher-quality second-pass reviewer, which accords with the results of this chapter.

3.5.3 Limitations

The experiments we have conducted do come with several limitations that should be addressed before further investigating the applicability of the observed results. First and foremost, these experiments have only studied the case of passive supervised learning, where we use a fixed set to train the machine learning algorithm to rank the document collection. The resultant ranked list is then reviewed in rank order until high recall is achieved. Cormack and Grossman's Simple Passive Learning (SPL) [40] is a specialization of passive learning for electronic discovery that uses random sampling to generate the training set, and our results are most amenable to such an approach. The state of art is perhaps better reflected in interactive, active learning approaches that make use of human-in-the-loop assessment [120, 40, 41, 97]. Any guidance that our results may have for informing active learning strategies has not been investigated, and the research community would benefit from additional insight.

While our experimental approach is reflective of SPL, we have still made use of "convenience samples" of available collections and judgments which is a far cry from the random sampling employed in SPL. The judging pools were either derived using depth pooling (a strategy aiming to identify enough relevant documents to compare systems) or stratified sampling (to more accurately estimate system performance). The values presented in Table

3.6 clearly show that there is a substantial difference between our convenience sample and the entire collection.

We have seen that over the various combinations of surrogates and authorities, some combinations fare poorly and others perform beyond initial expectations. Additional experiments should be conducted, with more assessors, to gain a more nuanced and developed understanding of the underlying causal factors for these differences. While there is the potential to artificially create additional surrogates based upon existing data,[7] more controlled experimentation through user studies should be conducted to provide the necessary insight. There have been some studies to this effect [165, 132, 91], but not all dimensions have been explored. In particular, we should seek out additional confirmation (or refutation) of the observation that increasing diversity or liberality of training data improves ranking quality when compared to that of a single (conservative) surrogate.

3.5.4 Extensions

For all three sets of experiments presented in this chapter, we made use of (substantially) fewer assessments for training than were rendered by some of the assessors. From a practical standpoint, this amounts to a great deal of wasted effort but was necessary to ensure equitable comparison between these different assessors. Suppose we had not removed the additional documents judged by Waterloo and NIST in the experiments in Section 3.3. Then when training the classifiers and evaluating them it would not be clear whether or not the effects observed were due to disagreements on documents shared between the two sets or due to documents that were specific to a particular set. For example, NIST judged many more documents (approximately 5 times more, the majority of which were not relevant) than Waterloo—this discrepancy could unduly confound the results in ways that may not be easily managed.

Similarly, in Section 3.2, we could have followed the lead of Voorhees and supplemented the secondary assessment sets with the documents judged by the official assessor. However, this would confound the experiment, since it would resemble our merged surrogates rather than an independent assessor. Alternatively, we could have treated documents judged by the official assessor but unjudged by the secondary assessors to be not relevant for the secondary assessors, as was done for the Waterloo assessments in Section 3.4. However, this approach was only valid in the Legal track experiments, because we could reasonably

[7] With the TREC-4 surrogates, we might recast J2 and J3 such that for each topic J2 has the set of assessments with fewer relevant judgments than J3 has, making J2 as conservative as possible and J3 as liberal as possible.

assume that if Waterloo did not assess a document as relevant, and thus submit it, Waterloo would have considered it not relevant. Such is almost certainly not the case for the secondary NIST assessors from the Voorhees study [152], who likely would not have found all documents unjudged by them to be not relevant.

Accordingly, we believe that the decision to confine training documents, except the randomly added documents, to those judged by all assessors or to the pooled documents in Section 3.4 was the most reasonable course of action given the desire for a fair and equitable comparison on even footing. In Chapter 7, we discuss some potential strategies that utilize machine learning to align overlapping sets of assessments that are potential avenues of future inquiry.

Chapter 4

TREC 2015 Total Recall Track

Total recall was a term first coined by Zobel, Moffat, and Park in an essay for the SIGIR forum [170] to describe the situation where a user would be dissatisfied with any recall level less than 100%. In this spirit[1], the Total Recall track, first offered at TREC 2015, was designed with the goal of investigating the implementation and evaluation of high-recall retrieval systems. In particular, the track was concerned with methods that took advantage of iterative human feedback to achieve high-recall with as little human effort as possible. The formal task definition is as follows:

> Given a topic description (like those used for ad hoc and Web tasks), 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. ([122])

Such a formalization facilitated the creation of an assessment service that would provide documents, topics, and relevance assessments to track participants in an automated fashion through a Web API. One can imagine this task as being a variant of the adaptive filtering task found in previous TREC tracks [117] and a continuation of the ideas behind several TREC Legal tasks [76, 46, 71]. An important difference being that the Total Recall track has focused on minimizing the assessing effort required to achieve high recall, which was not always of chief importance in previous endeavours.

[1] And partially, for the tongue-in-cheek references to the Arnold Schwarzenegger film.

This evaluation service was primarily contained in a Web server (Section 4.3.1) that facilitated run creation, corpus distribution, the aforementioned document assessment process, and some basic online evaluation. Included in this platform was a baseline model implementation (the "BMI," Section 4.3.3), distributed as a VirtualBox virtual machine (VM), that participants could modify and augment as they saw fit.

The development of this service was the result of experience with and an examination of past attempts at high-recall retrieval TREC tasks (e.g., Spam track [49], Legal track[18], the HARD track [12], and Filtering track [117], which revealed several issues that must be managed for high-recall tasks to be successful and which we believe the track has largely addressed. The following were identified as several key issues in ensuring equitable and fair evaluation of high-recall retrieval methodologies:

- It has been observed, that more interaction with the topic authority in Legal Track tasks may correlate to better final performance [75, 109, 149].
- Different versions of data sets exist with varying amounts of formatting, processing, and cleaning: for example, the multiple versions of the Enron corpus and the subsequent discordance between them and their associated relevance assessments [2].
- Use of real data must ensure that any personal information is not made available to participants. Even cleansed corpora have been known to leak damaging material (e.g. social security numbers) [93].
- Simulation and evaluation of *interactive* high-recall retrieval requires *complete* relevance assessments which require a great deal of care and effort to create and for which traditional depth pooling does not suffice.[3]
- Such tasks often have a high threshold to participation; not only must participants process the corpora but they must also create working systems to get started.

With respect to these issues, this chapter describes the data collections used in the 2015 iteration of the track (Section 4.2), the Web service itself (Section 4.3.1), the conceptualized design of participant systems (Section 4.3.3), and the submissions to track and the results produced by these submissions (Section 4.5.2). As they become pertinent, design choices

[2] As far as the thesis author is aware, there has not been a systematic or successful attempt to align the relevance assessments for the various versions of the Enron corpus.

[3] If one seeks only to evaluate systems, complete relevance assessments may not be necessary, as stratified sampling can be used to estimate performance. Complete relevance assessments are necessary when one is simulating human involvement, where each document *should* have a corresponding judgment.

influenced by these issues will be discussed and the logic behind them explained. We conclude with a discussion of the 2015 track iteration and a retrospective look at how well these issues were handled and potential improvements for the future (Section 4.6).

4.1 Task Description

We begin by providing a more grounded description of the task at the 2015 Total Recall track. Participants implemented automatic or semi-automatic methods(e.g., systems made use of human intervention aside from the Web service) methods to identify as many relevant documents as possible, for a given topic and document collection, with as little simulated human effort as possible (i.e., judgments rendered by the service). To facilitate this a Baseline Model Implementation (BMI) was provided to participants as a fully automatic baseline which they were free to implement and modify. An overview of potential participant system implementations and the BMI are provided later in the chapter (Section 4.3.3). The BMI was meant to provide the simplest means of participation: all a participant had to do was run the BMI with their name on it, analogous to rubber duck races for charity. Essentially, the participants got a working and competitive system for free that they could spend as much or as little time optimizing and improving as they wished.

Document collections, topics, and pre-generated relevance assessments were supplied to participants via the aforementioned Web service (described in detail in Section 4.3.1). A system would begin by downloading the collection and theb processing each topic for that collection, identifying documents from the collection that the system believed to be relevant to the topic and submitting them in batches (whose size depended on the participant's system) to the service for automated assessment. The service maintained a log of each document received and immediately produced an authoritative (i.e., gold standard) assessment for each document received in a particular batch. Accordingly, the track made use of collections that had been pre-labelled (as relevant or not) for each (document, topic) pair, and the service merely returned this existent judgment.

By taking this controlled approach to high-recall retrieval experimentation, we hope to mitigate any bias that might be perceived because of "too much time" spent with a topic authority (gold standard assessor) or the ability to ask the right questions. The focus is then on the algorithms used and less so on the human interaction between topic authorities and participants. This should not be construed to suggest that such interaction is bad but that we merely sought to eliminate any confounds that different approaches to such interaction might have on system evaluation. Such an approach also facilitates experimentation with more topics over more test collections, since there is no requirement to provide access to a

high-quality topic authority (that is willing to volunteer their valuable time), a requirement that has previously limited how many topics and how much interaction could be performed [76, 46]. Keeping in mind the results of Chapter 3, we used a single authoritative conception of relevance for purely pragmatic reasons. That is, getting one set of complete assessments is often very hard and getting a second would be next to impossible. While having training and evaluation assessments would be the best-case scenario, it is something that was not possible for the 2015 iteration but may become feasible as better solutions to the high-recall retrieval problem are found.

The track offered two modes of participation to interested parties: an `At-Home` mode where interaction occurred over the Web; and a `Sandbox` mode where interaction occurred locally on a single machine with no access to any thing outside of that single machine. Participation in the `Sandbox` mode required participants to submit a working virtual machine (VM) that would be then run, without Internet access, on test collections that were unknown to the participants. Figure 4.1 provides a high-level conceptual depiction of how the `At-Home` and `Sandbox` modes differ.

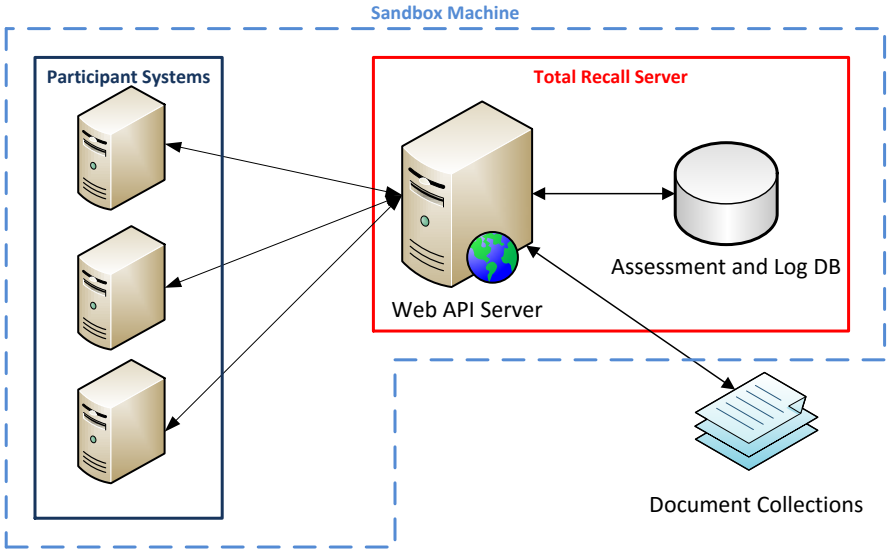


Figure 4.1: A high-level look at how the various components interact in live and sandbox environments. Note that the dashed blue line denotes the Total Recall server and participant VMs running on the same machine.

Regardless of configuration, the data collections (the types and formats described in Section 4.2) are largely "plug-and-play," meaning that new data collections can, with little effort, be added to the platform. The result is that new experiments can be run on partic-

ipant systems without necessitating participant action and without requiring participant resources. Furthermore, the `Sandbox` mode of participation allows collections to be used that would otherwise be "too hot" to distribute or would require onerous, time-consuming, and imperfect anonymization.

4.2 Test Collections

For our purposes, we consider a test collection to be a set of documents along with a set of qrels, which are tuples that map documents and topics (information needs) to relevance assessments. Test collections are transmitted to participant clients through the Total Recall service first by transmitting the set of documents and then by having systems request relevance assessments in a document-at-a-time manner. To aid in system development, each test collection was curated so that a single file contained a single document. This is contrary to many past TREC collections, where a file might contain multiple documents, and was performed to simplify document parsing. We additionally translated each document into a plain text rendering from whatever native format it was originally stored in (PST, WARC, etc.) to ease the burden of system design, as participants would not have to worry about processing multiple file types.

For the purposes of system development, we used three publicly available test collections: 20-Newsgroups;[4] Reuters-21758;[5] and, a variant of the Enron email collection[6] created by the University of Waterloo in the course of participation in the 2009 Legal track. Inclusion of the first two data sets was to facilitate rapid development and testing of participant systems, as both contain approximately 20,000 documents. Neither was meant to be representative of a valid test collection but merely there to help participants gain confidence that their system was working. Moreover, we offered samples of these first two corpora as a mechanism to ensure participants understood and could correctly interact with the Web service, without devoting the extensive execution time of a larger corpus. The Enron collection was our attempt to provide a representative collection, as it had previously been used in the TREC 2009 Legal track [76], that would more accurately indicate the effectiveness of participant systems. For all these collections, we provided an online mechanism to provide recall, effort, precision, and F1 scores for checking system performance.

[4] `http://qwone.com/~jason/20Newsgroups/`

[5] `http://www.daviddlewis.com/resources/testcollections/reuters21578/`

[6] `http://plg.uwaterloo.ca/~gvcormac/treclegal09/`

At the beginning of July, the `At-Home` phase of participation began and three additional corpora were released to participants (after signing the appropriate usage agreements). These collections were the official test collections, so as a means of preventing any potential meta-learning by the systems, participants did not receive any explicit system performance feedback. We acknowledge that some participants may have had internal quality assurance processes that would have helped determine their own performance.

Not all corpora could be distributed by the service due to usage agreements or the risk of divulging sensitive data to the public, so, generally, such restricted corpora was distributed only in the form of `Sandbox` participation (discussed in Section 4.3.2). This helps to ameliorate issues with imperfect anonymization and the inability to give all interested parties access to the data. We do not guarantee to have prevented all forms of information leakage; however, we have taken steps towards more privacy-aware experimentation and structured the release of the collections in this way to aid in preserving the privacy of all parties when private collections are used. The following two subsections discuss the `At-Home` and `Sandbox` test collections, including the assessment process, the topics, and other corpus statistics.

4.2.1 `At-Home` Collections

The `At-Home` collections (`athome1`, `athome2`, `athome3`) are email, Web forum, and Web news collections. The `athome1` collection was used solely by the TREC 2015 Total Recall track, while the second and third collections were partially created in conjunction with the TREC 2015 Dynamic Domain (DD) track. Due to miscommunication and different task purposes, the final joint `athome2` and `athome3` collections were not suitable for a high-recall retrieval task, since they had very sparse assessments across many topics. Accordingly, these collections were augmented with additional re-assessments conducted by the track coordinators.

The assessments for `athome1` and re-assessments for `athome2` and `athome3` were conducted by a track coordinator and based upon the active learning approach of Cormack and Mojdeh [52], which consists of an initial phase of interactive searching and judging followed by interactive learning. This process continued until the coordinator felt that they had identified as many relevant documents as possible. A more rigorous approach, including a quality assurance phase, would have been ideal, but, due to time constraints and the performance of the aforementioned method in previous TREC tracks, it was believed such performance would be sufficient.

Topic	R	% of Corpus
School and Preschool Funding	4542	1.561
Judicial Selection	5836	2.006
Capital Punishment	1624	0.558
Manatee Protection	5725	1.967
New Medical Schools	227	0.078
Affirmative Action	3635	1.249
Terri Schiavo	17135	5.888
Tort Reform	2375	0.816
Manatee County	2375	0.816
Scarlet Letter Law	506	0.174
Average	4398	1.511

Table 4.1: The 10 topics assessed and used as part of the `athome1` collection and released to participants.

`athome1` - Jeb Bush Collection

This collection contains the 290,999 *redacted* emails of Jeb Bush[7] during his eight-year tenure as governor of Florida. The files were originally distributed as Microsoft Outlook PST files, which were rendered into single plain-text files that maintained basic email information, including the sender, recipient(s), subject, body, and date. A custom tool was written [8] to process the original PST files and output the desired text. Each individual email was rendered into a single file for two reasons: Outlook PST files do not attach globally unique identifiers to emails, making it hard to disambiguate them; and, to facilitate ease of processing on behalf of participant systems. While keeping metadata, such as email threading, would have been potentially beneficial to participants, it was not believed that such metadata was crucial to perform the task, so additional coordinator time was not taken to attempt to preserve such metadata.

Topics, depicted in Table 4.1, were created by choosing 10 issues that have been associated with the governorship of Jeb Bush and were believed to be sufficiently important and/or interesting by the track coordinators.

[7] Available from `http://jebemails.com/email/search`.
[8] By coordinator Gordon V. Cormack.

Topic	R	% of Corpus
paying for amazon book reviews	265	0.057
CAPTCHA Services	661	0.142
Facebook Accounts	589	0.127
Surely bitcoins can be used	2299	0.494
paypal accounts	252	0.054
Using TOR for anonymous browsing on the internet	1256	0.270
Rootkits	182	0.039
Web Scraping	9517	2.046
article spinner spinning	4805	1.033
Offshore Host Sites	179	0.038
Average	2000.5	0.430

Table 4.2: The 10 topics assessed and used as part of the `athome2` collection and released to participants.

`athome2` - Illicit Goods Collection

This is the first collection derived from the same TREC 2015 Dynamic Domain track collection. It consists of 465,147 Web forum threads (i.e., lists of related posts) that were collected from Blackhat World (`http://www.blackhatworld.com`) and Hack Forum (`http://hackforums.net`). Crawls of these two forums were were originally provided in a version of the Concise Binary Object Representation (CBOR) format,[9] which necessitated the use of a simple tool to process each CBOR object and extract each forum thread, after which is was ran through the `lynx` command-line tool to render the forum thread with all HTML and Javascript removed. This rendered forum thread was stored as a single document for retrieval. This choice was made because individual posts were relatively small, and both Dynamic Domain and Total Recall coordinators thought that using the entire thread would make more sense since all the posts in a thread are likely related to each other.

Since the original DD topics were too sparse, the Total Recall coordinators selected 10 topics that seemed reasonably interesting and re-judged each of the selected topics while maintaining any official NIST assessments. Table 4.2 describes the 10 topics and their relevance counts.

[9]CBOR is inspired by the popular JSON format with binary encodings. For more information see `http://cbor.io/`.

Topic	R	% of Corpus
pickton murders	255	0.028
pacific gateway	113	0.013
traffic enforcement cameras	2094	0.232
rooster turkey chicken nuisance	26	0.003
occupy vancouver	629	0.070
rob mckenna gubernatorial candidate	66	0.007
rob ford cut the waist	76	0.008
kingston mills lock murders	1111	0.123
fracking	2036	0.225612
paul and cathy lee martin	23	0.003
Average	642.9	0.071

Table 4.3: The 10 topics assessed and used as part of the `athome3` collection and released to participants.

`athome3` - Local Politics Collection

The second collection derived from the Dynamic Domain track consisted 902,434 Web news articles that were crawled from various new sources in the northwestern United States and southwestern Canada. These articles form a subset of the TREC 2014 KBA streamcorpus [64], which forms a crawl of the Web revolving around social media (e.g., blogs) and news media from October 2011 to May 2013. Similar to the `athome2` collection, these documents were contained in a binary format, called Thrift[10], so the raw contents of each article was selected from the formatted data and rendered, using `lynx` again, into a plain-text format with HTML elements removed.

The Local Politics collection also had 10 topics selected for additional re-assessment by Total Recall track coordinators, which are outlined in Table 4.3.

4.2.2 `Sandbox` Collections

Sandbox collections are meant to facilitate experimentation on sensitive and private data (i.e., data that has personal or restricted information) as well as to provide black box tests to system creators. The 2015 track provided two sandbox collections that encapsulated two different tasks; information governance of former governor of Virginia Tim Kaine's

[10]See `https://thrift.apache.org/` for the exact specification.

Topic	R	% of Corpus
public record	131698	32.765
open record	166118	41.328
restricted record	14341	3.568
Virginia Tech shooting (hold)	20083	5.000
Average	83060	20.664

Table 4.4: The 4 topics assessed and used as part of the kaine collection.

email, and finding medical records with a particular ICD-9 code.[11] As in Section 4.2.1, we present descriptions of how each sandbox collection was collected and used for the TREC 2015 iteration of the track.

`Kaine` - Tim Kaine Collection

This collection comprises a 401,953 email subset of the 1.3 million emails located at the Library of Virgina that were collected[12] during Tim Kaine's eight-year tenure as governor of Virginia. This subset was used because it had previously been labelled by the Library of Virginia archivist with respect to the following 4 categories: public record, open record", restricted record, and Virginia Tech shooting (hold). Each of these categories formed a topic for the `Kaine` test collection using only those labels generated by the Library archivist. In addition, these emails were stored in Microsoft Outlook PST format, using the PST formatting tool that was used with the `athome1` collection was also used with this collection. In this case the tool was made to be semi-autonomous, as no coordinators could interact with the data due to its sensitive nature. Further, the staff at the Library of Virginia did not have the resources to monitor the software at all times. Accordingly, some concessions were made with respect to limiting the size of attachments that would be produced by the tool (i.e., the first 1MB of each attachment was appended to the corresponding email). Otherwise, the tool would occasionally stall due to the limitations of the machine running the software. Table 4.4 provides the usual description of topic prevalence.

[11] ICD refers to the International Statistical Classification of Diseases and Related Health Problems. The ICD-9 codes are a high-level classification of possible diseases. A list of the ICD-9 codes is available from Wikipedia: `https://en.wikipedia.org/wiki/List_of_ICD-9_codes`.

[12] See `http://www.virginiamemory.com/collections/kaine/under-the-hood` for more details on the collection.

MIMIC - MIMIC II Clinical Collection

The MIMIC II Clinical collection[13] consists of 31,538 patient Intensive Care Unit visit records that were anonymized and time-shifted. Only the textual components of each patient record were used as a document in the MIMIC collection, which included one or more nurses' notes, radiology reports, and discharge summaries. Each record was also tagged with various ICD-9 codes and the 19 top-level codes were used as topics for the MIMIC test collection. For the sake of brevity, we do not reproduce the 19 top-level codes and merely report the average relevance of 7794.37 documents or 24.70% of the corpus. We note that documents in this collection could have more than one associated ICD-9 top-level code, but that should not affect per-topic results as they are treated independently. While this data set had the potential to be an At-Home collection due to the anonymization that had already taken place, it was decided that to ensure the safety of the individuals in the records, the collection should not be distributed.

4.3 Service Architecture

This section describes the overall architecture of the Web service, including a description of the Web API, how it was sandboxed, and the envisioned design of participant systems, which includes a description of the Baseline Model Implementation that was provided to "jump-start" system development.

4.3.1 Server

The Total Recall server operates as a Web service that participant systems interact with over HTTP using a REST(ful) API. There are three main types of interactions between systems and the service:

- Request for information on a corpus and a link to download the corpus. Such details include the type of corpus (e.g., email, newswire) and the language of the corpus (English only for 2015).
- Request for information on a topic to process, which includes the corresponding corpus and a description of the information need.

[13] Available at https://physionet.org/mimic2/.

- Request for relevance assessment of documents with respect to a particular topic.

Secondary interactions can occur, including the following: starting and finalizing a run; result generation for developmental data collections; error log checking; and, "calling one's shot," which is briefly described in the evaluation section (4.4). Figure 4.2 shows the general of sequence of interactions that would occur between all the components of the Total Recall architecture.

Figure 4.3 provides the general workflow that a client system might be expected to perform. For brevity, we omit the actual API that was used to interact with the service and instead direct interested parties to the API documentation.[14] Note that all requests and responses to and from the server were encoded in the JSON format, which has become a standard format for passing data between a Web server and client software.

In terms of implementation, the entirety of the Total Recall Web service was written in Node.js,[15] which is a Javascript runtime using an event-driven model for developing sever-side Web applications (e.g., REST(ful) APIs). Node.js is designed to be efficient and lightweight and is popular in modern Web development. It would have been entirely reasonable to choose any other framework or environment (e.g., Django, Flask) or even to write the server from the ground up. Taking all things into account, we decided that Node.js provided an interesting opportunity to explore the state of the art in current Web development processes while also ensuring that the service would generally be reliable and efficient. Where necessary (e.g., in storing judgments and a log of participant requests), the server relied on a default installation of MySQL. While other database software would have sufficed, MySQL is available on all major operating systems and is regularly maintained. In addition, the coordinator responsible for the implementation of the service (i.e., the thesis author) was familiar with the administration of a MySQL server. Accordingly, using MySQL was a choice made out of convenience and not necessarily one meant to optimize for any particular feature. However, as will be discussed in Section 4.6, a default installation was perhaps a poor choice and resulted in a small bottleneck near the end of the experimental period.

All code for the Total Recall track is available for public download at `http://repo.trec-total-recall.com`.

[14] `quaid.uwaterloo.ca:33333/#/api`

[15] See `https://nodejs.org/en/` for more information.

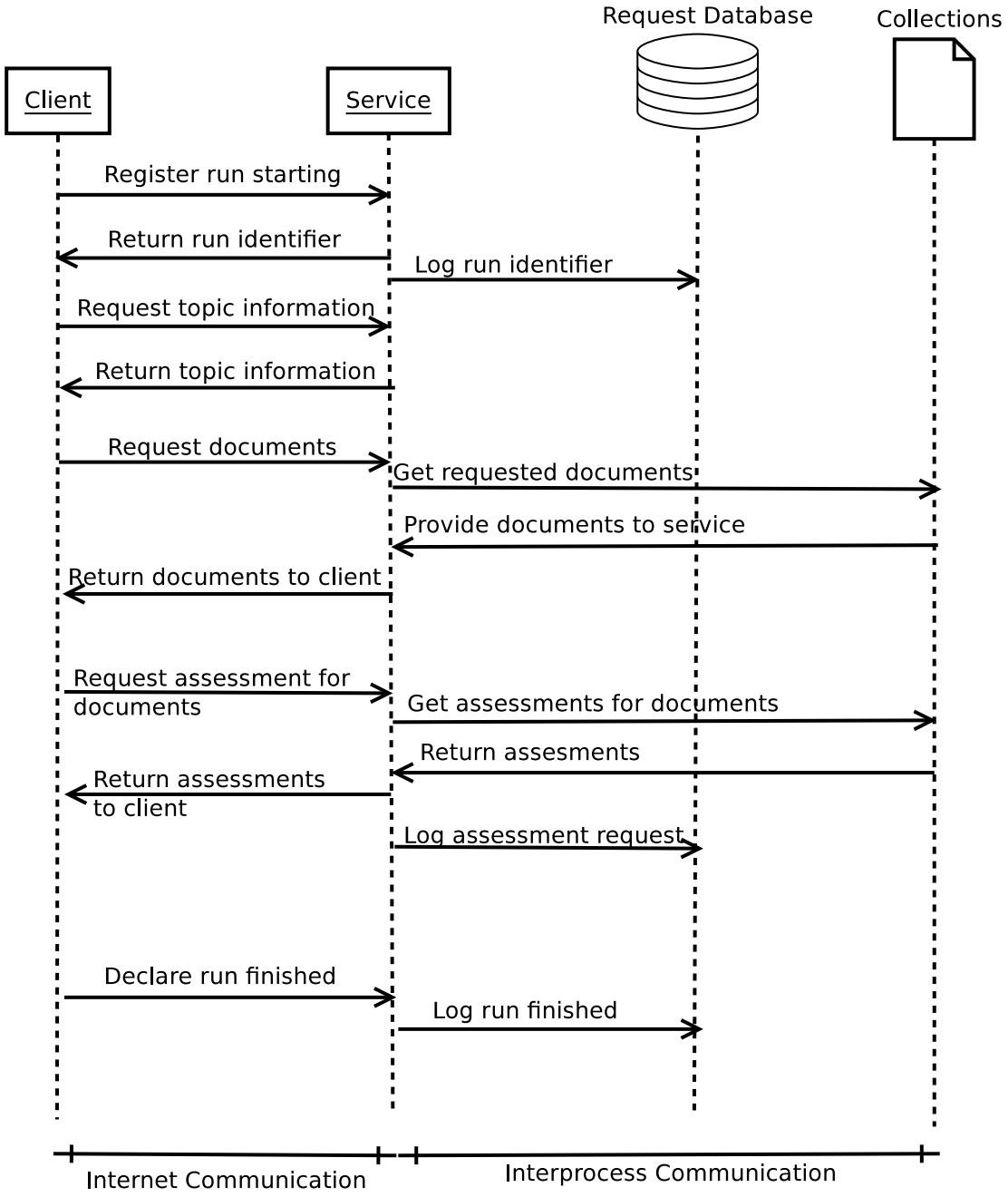


Figure 4.2: The general sequence of interactions between a client, the service, and the underlying database and data collections.

4.3.2 Sandboxing the Server

To facilitate the use of data collections that require complicated transmission protocols (e.g., participants would have to receive the collection out-of-band) or are "too hot to handle" (e.g., raw versions of confidential email), we provided a `Sandbox` mode of participation. In this mode, participants submitted their systems as a VM (more details in Section 4.3.3) rather than running them on their own hardware. These submitted clients ran on the same machine as an instance of the service, as outlined in Figure 4.1.

Data collections do not have to necessarily be stored directly on the sandbox machine; instead data providers are able to provide the data collections via USB flashdrive or external hard drive. Once the data collection is loaded into the service (using some relatively simple shell scripts), the sandboxed participant systems are run against these collections much in the same way that they would be in the live scenario. However, the participant's VM is prohibited from accessing the Internet in any way so as to prevent data leakage from the sandbox. This restriction is enforced by limiting access of the clients' machines to the service and (optionally) by air-walling the sandbox machine (i.e., never connecting it to the Internet once the data collections are present). Air-walling was performed for the `Kaine` collection due to the sensitive nature of the documents and associated labellings (i.e., some documents cannot be released to the public).

By enforcing these restrictions on the sandbox machine, we attempted to protect the data from unintended transmission and personal and private data from unintended dissemination. Furthermore, we could have limit the output of the sandbox server to be only summary evaluation measures and statistics, which are discussed in Section 4.4, once all participant systems had been run. The goal would have been to prevent accidental distribution of private data that may be contained in qrels and document identifiers. However, given the nature of the collections themselves and the formatting that was applied, it was deemed unnecessary to undertake this level of obfuscation.

4.3.3 Envisioned Participant Systems

The envisioned workflow for a client system is depicted in Figure 4.3, and is implemented by the the Baseline Model Implementation (Section 4.3.3). Participants can construct several different systems to interact with the Total Recall service. Clients for live experiments can be developed in several different fashions:

1. A purely automatic program (or set of programs) that performs the Total Recall task.

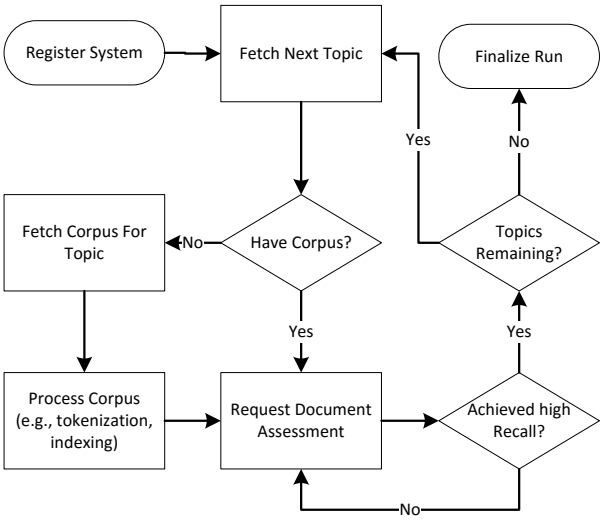


Figure 4.3: The envisioned workflow of a Total Recall participant system.

2. A customized virtual machine that runs the above.

3. A "sidecar" (i.e., a directory containing additional scripts). supplied to the BMI's virtual machine).

A sidecar submission is effectively an encapsulated version of (1) that is interchangeable with any other sidecar. That is, all the required components come with the sidecar and the BMI's VM does not need to be modified.

Additionally, the Total Recall task can be performed manually or through some combination of manual and automatic approaches. Accordingly, we allowed participants to submit a single manual run, which was to be performed before developing any automatic systems for submission. Participants could submit semi-automatic (or manual in standard TREC parlance) runs using either the Web API (as used by automatic systems) or a Web-based interface. The Web-based interface was quite simple and was just a pretty interface for the API. The intent was to provide just enough of an interface to allow participants to perform the task, without providing any functionality that would not be available to the automatic-only participants. Out of the 3 manual teams, 2 used the Web interface and 1 directly communicated with the Web service via the API. Figure 4.4 is a screenshot of the manual run interface that was used by those 2 groups.

For sandbox submissions, participants were required to submit systems that fall under the full VM or sidecar categories. This was done with the hope that either would work

out of the box, which was mostly true for TREC 2015. Previous experience with running participant systems in a sandbox for the TREC Spam tracks [49, 36, 37] led us to believe that allowing the submission of arbitrary code would lead to too much hassle. Indeed, the Music Information Retrieval Exchange (MIREX) has struggled with similar issues in their algorithms-to-data model from having allowed arbitrary implementation languages to be submitted [59, 60]. From this, we determined that the hassle required of the coordinators to debug and fix up any code that was not compatible with the sandbox environment (e.g., installing libraries, having necessary compiler versions) would be too high. By using a VM, we hoped to limit the necessity of such tasks. We were successful, as only one system require major coordinator intervention (out of 11 different systems).

A final benefit to requiring submissions of VMs or sidecars is that it allows commercial vendors of high-recall retrieval software to submit their products without the source code (which was required in previous tracks that used sandboxing). Instead, vendors can submit a fully compiled version of their software that runs on the VM and without exposing their intellectual property to track coordinators, data providers, or any other parties that might come into contact with the software. Extra steps can also be taken, such as encrypting the VM and/or software so that outside parties cannot access the software at all (outside of the normal interaction between client and server). However, to the author's knowledge no vendor software was submitted for `Sandbox` participation in TREC 2015.

Baseline Model Implementation (BMI)

The primary baseline is an augmented version of the Continuous Active Learning (CAL) method originally presented by Cormack and Grossman [40], which is called AutoTAR [41]. Unlike the original version of CAL, AutoTAR uses exponentially increasing batch sizes, and unlike in the experiments of the AutoTAR paper, we used a tuned version of `sofia-ml`[16] based upon suggestions from the package's author. The precise options supplied to `sofia-ml` were as follows: `--learner_type logreg-pegasos --loop_type roc`, `--lambda 0.001 --iterations 200000`, and `--dimensionality 1100000`. Furthermore, tf-idf features derived from alphabetic-only words were used rather than the overlapping byte 4-grams used in the original AutoTAR. This change was performed to reflect the widespread use of such features as well as a noticeable performance increase during early experiments. The exact specifics of AutoTAR and CAL were previously described in Chapter 2, so they are not discussed further.

This baseline was implemented, by coordinator Gordon V. Cormack, as part of the

[16] `https://code.google.com/archive/p/sofia-ml/`

Figure 4.4: Screenshot of the manual run toolkit.

virtual machine with sidecar approach, discussed previously, where the algorithm used is designed as a sidecar for a simple Debian VM. This was done to provide a model implementation of what such a sidecar might look like and a working implementation that participants would be free to modify and use in conducting experiments. The BMI was implemented using a combination of C++ and bash scripts (along with associated command-line tools). The canonical implementation is available for public download at `http://plg.uwaterloo.ca/~gvcormac/trecvm/`.

Cormack and Grossman have shown that AutoTAR generally outperforms a simple CAL implementation. Accordingly, our intent was to provide a reasonable baseline that could "fast-track" participants to a working system, without requiring them to worry about API programming. We hoped that the BMI would provide ample opportunity for participants to improve upon its results or be inspired by the technique and devise original algorithms of their own.

4.4 Evaluation and Metrics

At TREC 2015, the Total Recall track used a combination of set and rank-based metrics for the purposes of evaluating systems. The gain curve, which measures recall as a function of effort, was the primary evaluation metric as it clearly shows how much or little effort is required to achieve high-recall by a set of systems. Additionally, precision-recall curves, which have been a mainstay of TREC evaluation, were used as a secondary form of ranked evaluation. Precision-recall curves, in contrast to gain curves, more clearly provide an illustration of how much superfluous (non-relevant) material is introduced as recall increases. By providing both types of curves, it was hoped that a full picture of system performance could be derived by participants, coordinators, and other interested parties.

For those runs that chose to "call their shot," recall, precision, and F1 were computed at each such point, as these were typical values of interest in the literature, past iterations of high-recall task (i.e., the Legal track), and the domain of electronic discovery. The intended goal was to provide a more tangible, "real-world" scenario that acknowledges the fact that the review must stop at some point (typically much sooner than reviewing the whole corpus). Additional metrics could have been computed at these points; however, it was unclear how concerted an effort would be made by participants, nor was it clear what other metrics might be useful to participants.

Finally, the track introduced a new metric: recall @ aR+b effort,[17] where R is the

[17]This may be more aptly called X @ aR+b effort where X is any suitable metric (e.g., precision, recall,

number of relevant documents for a topic. This metric is meant to measure recall when effort is proportional to the number of relevant documents for a particular topic. One may consider it a post hoc evaluation inspired by the "reasonableness and proportionality" aspects of the 26th rule of Federal Rules of Civil Procedure [111], where achieving high recall should be tempered by the needs of the case. Recall @ aR+b focuses on achieving high recall while not expending exorbitant effort relative to to the number of relevant documents. A further modification to such a metric may be to model more carefully the importance of certain documents (e.g., smoking guns) and use that importance in the calculation of R, and thus, of recall. Such modelling is not employed in the 2015 track and is left for further research.

A more intuitive perspective on aR+b would be thus: the a variable represents the amount of allowable non-relevant material (e.g., $a = 2$ would indicate that we accept one non-relevant document for every relevant document); the b variable, on the other hand, represents a fixed overhead that may be necessary to "learn" a topic at the outset of the review process (e.g., to to mitigate the cold-start problem).

The intended goal of the recall $aR+b$ effort measure is to provide a summary measure that is a balance between the information presented in gain curves and that preented in precision-recall curves and can also conduct meaningful statistical analyses that may otherwise be hard to render with such curves. Without any existing guidance on what is acceptable effort for a particular task, all combinations of $a = [1, 2, 4]$ and $b = [0, 100, 1000]$ were computed for the track as a means of exploring the evaluation space. It is worth noting that when $a = 1$ and $b = 0$, the metric is equivalent to R-Precision, the precision at R documents in ranked list, which is otherwise known as the recall-precision break-even point (since they are equal at this position).

Note that in the course of submission, manual (i.e., semi-automatic) participants were asked but not required to log and submit the number of documents that were reviewed by a human outside of the online service. The numbers reported in this thesis do not account for any adjustments, since it was not a required component and there was doubt among coordinators about the accuracy of some of the submissions.

4.5 TREC 2015 Results

In this section, we provide an overview of the participating systems as reported in the official proceedings as a basis for discussion later. We then present the results of the first

F1)

iteration of the track and provide some discussion based upon those results.

4.5.1 Systems Descriptions

Many of the submissions to the track modified the BMI in some way, so for those runs, we report only the differences from the BMI. The following are descriptions of the participant systems in no particular order and given by their TREC group identifier. As far as we are aware, all but one of the teams that participated submitted a final report to the proceedings. The exception is the NINJA group, who, to maintain confidentiality, did not release a track report.

catres [114]

The catres group conducted a semi-automatic submission for the `athome1` collection only. Their process made use of proprietary continuous active learning technology that was seeded by manually identified documents. A team of three searchers spent a single hour training the technology on a particular topic. This hour of work included the searchers familiarizing themselves with the topic, issuing queries to find relevant material, and identifying relevant material when found. After this initial hour, the CAL system interacted solely with the Web service with no further human intervention. Due to a late start, the catres group also undertook various heuristics to expedite the running of their software, which included feature set reduction and incremental batch size increases for training with CAL.

CCRi [56]

The CCRi team created a single system that made use of dynamic neural networks to complete the task. Initially, CCRi produced tf-idf feature vectors which were then condensed into a smaller feature space using the skip-gram word-embedding model of `word2vec` [107]. A similar process was used to convert each topic into a similar condensed feature space. Documents were then initially selected based upon highest-scoring cosine similarity for relevance feedback (i.e., submission to the server). Using these documents, a new neural network was trained to expand the embedding space and the nearest-neighbour search was repeated on the expanded embedding. Batch size (i.e., the k-nearest neighbors) was updated as a power of $10^{\#\text{ relevant documents found}}$ until a maximum batch of 2000 documents was reached.

eDiscoveryTeam [99]

eDiscoveryTeam conducted a semi-automatic approach for all three `At-Home` collections. Similar to the catres group, eDiscoveryTeam consisted of a group of searchers who conducted searches to seed a CAL technology. Unlike the catres group, eDiscoveryTeam used a variety of searching methods, ncluding keyword search, uncertainty sampling, random sampling, relevance sampling, and so on. However like catres, searchers spent some amount of time to familiarize themselves with a particular topic for which they had insufficient prior knowledge. In total, the team reports that they spent 360 hours reviewing and analyzing over 16 million documents or, on average, approximately 46,000 documents per hour.

TUW [103]

The TUW team submitted a series of variants of the BMI to `athome1` collection and the `Sandbox` phase of the track. In particular, they examined the effect of including and removing stopwords from documents, the inclusion of a BM25-based terming weighting that uses a collection-specific `b` parameter for the tf-idf weighting, and the fusion of the default BMI classifier with 5 other classifiers in the `sofia-ml` package. This fusion was conducted across the previous two features (i.e., stopwords or not and the modified weighting). Accordingly, the team submitted 6 separate runs to the track.

UvA.ILPS [148]

UvA.ILPS submitted two systems that were variants of the AutoTAR algorithm[18] to both the `At-Home` and `Sandbox` phases of the track. Their two approaches shared the same core in AutoTAR, with key differences being that they would adjust batch size based upon relevant documents identified and would stop when batch size was 0 and 1% of the corpus was reviewed. The difference between the two systems is that one used logistic regression (as per BMI) and the other made use of a random forest classifier.

WaterlooClarke [168]

The WaterlooClarke group attempted to improve upon the BMI in several ways. They begin by using clustering to increase diversity of seed documents and then embed these

[18] Variants of AutoTAR rather than BMI, since they entirety was reimplemented in Python with different packages. Determined through personal discussion with the lead participant.

clusterings into a weighted graph until at most 50 documents were judged. Combining these judged documents and the synthetic document of BMI, they ran 5 logistic regression classifiers[19] and selected the documents to be judged from the fusion of their result lists. If there exists sufficient relevant material in these topic documents, they selected additional documents from this list. Otherwise, query expansion was conducted based upon the judged documents and then used with BM25 to find additional documents. The search results were then fused with the classifiers, and the remaining documents were selected from that list. The batch size for the next iteration was increased similarly to the BMI.

Note that the WaterlooClarke group consisted of graduate students located at the University of Waterloo (home institution for three of the coordinators at the time), but they were afforded only the same access to the collections as other groups.

WaterlooCormack [43]

The WaterlooCormack submission was identical to the BMI with the exception of two different "shot calling" criteria. The first attempted to find the "knee" in the BMI's gain curve, using the marginal precision drop between iterations as a threshold for when this occurs. The second used a simple calling criterion when effort exceeds $N = \alpha r + 2399$ effort for various values of α, and r is the number of documents assessed relevant.

Note that WaterlooCormack consisted of two track coordinators (Gordon V. Cormack and Maura R. Grossman) who had knowledge of the `At-Home` collections but had no knowledge of the `Sandbox` collections when the stopping criteria were selected. The BMI (developed by the same coordinators) was frozen prior to any data set development.

Webis [72]

Webis submitted two systems to the entirety of the `At-Home` and `Sandbox` phases. One system was a baseline approach that began by combining the results of the top 16 results of a BM25 search and of an SVM trained on a combination of pseudo-relevant (top BM25 hits) and pseudo-not-relevant documents (randomly sampled). After this batch was judged, a CAL approach was taken such that the next batch size increased if the precision of the current batch exceeded 0.5 or decreased if it was less than 0.4. The process terminated when the batch size reached zero. The modification of the baseline approach uses phrase extraction from identified similar relevant documents and to form a new query that is used

[19] Trained with different randomly sampled pseudo-not-relevant documents.

to find new pseudo-relevant documents with which to train the SVM classifier in each iteration.

WHU_IRGroup [166]

WHU_IRGroup submitted a single run to the `athome1` collection. Their methodology was to conduct iterative query expansion and reformulation over identified relevant documents. The iterative approach used documents identified as relevant to select distinct terms that represent relevant material.

4.5.2 Results

Gain and Precision-Recall Curves

Figures 4.5 through 4.9 depict average gain curves and average interpolated precision-recall curves for the best runs submitted by each group.[20] Use of the best-performing run is to ensure that the plots are comprehensible and aesthetically pleasing to read. Furthermore, for many of the runs from the same group, the differences between runs were generally minute, but this was not always the case, as discussed in the following section.

If we consider the `athome1` results, we can observe that the better-performing systems are achieving 80-90% recall, on average, at around 10,000 documents reviewed or just over twice the average R, which appears to be reasonable. Similar trends appear to hold for the `athome2` and `athome3` collections, where this point happens sooner due to the lower average prevalence (see Tables 4.1 through 4.4 for the average prevalence of all corpora). That is, for `athome2` and `athome3`, the best-performing systems have already hit the "high-recall plateau," where the systems must look for those remaining few relevant documents, by 10,000 documents.

The `Sandbox` collections, at first glance, appear to do substantively worse than the `At-Homes` but once we consider that the average R for `MIMIC` and `Kaine` is many times higher than for any `At-Home` collection, these results seem much better. For example, most runs on the `MIMIC` collection achieved over 85% average recall for 15,000 documents reviewed (about twice the average R), which comparatively appears to be better than the `At-Home` results. However, this is quite likely due to the much higher average prevalence of the `MIMIC` collection when compared to any of the `At-Home` collections. The `Kaine`

[20] The best runs were determined by sorting based upon R-Precision, i.e., R-Recall.

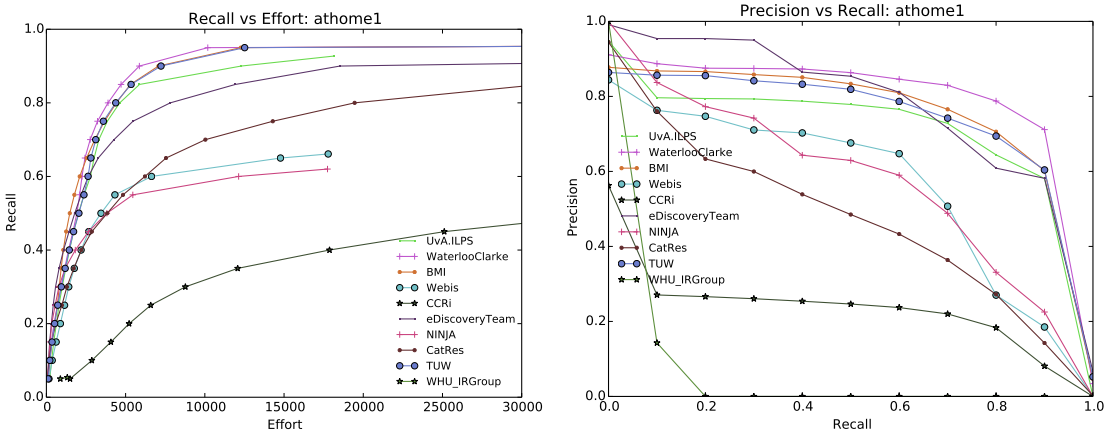


Figure 4.5: Average gain and precision-recall curves for the `athome1` collection.

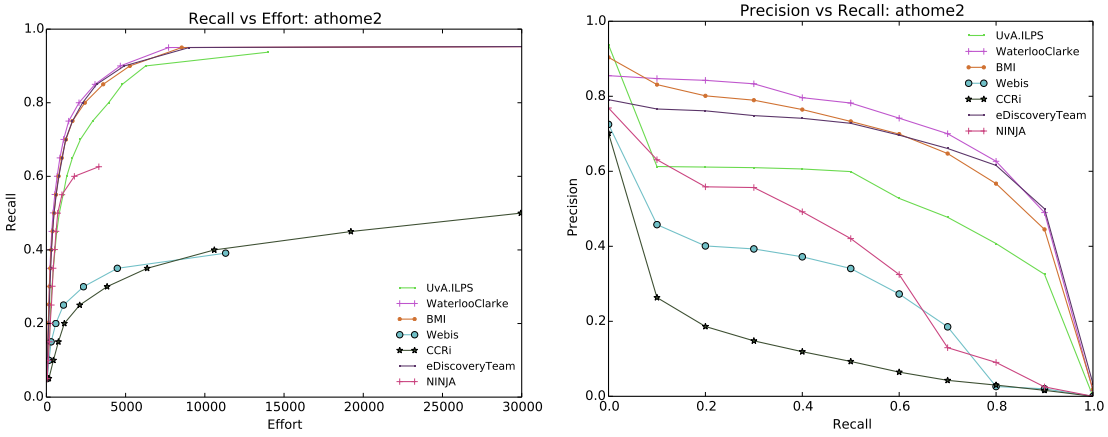


Figure 4.6: Average gain and precision-recall curves for the `athome2` collection.

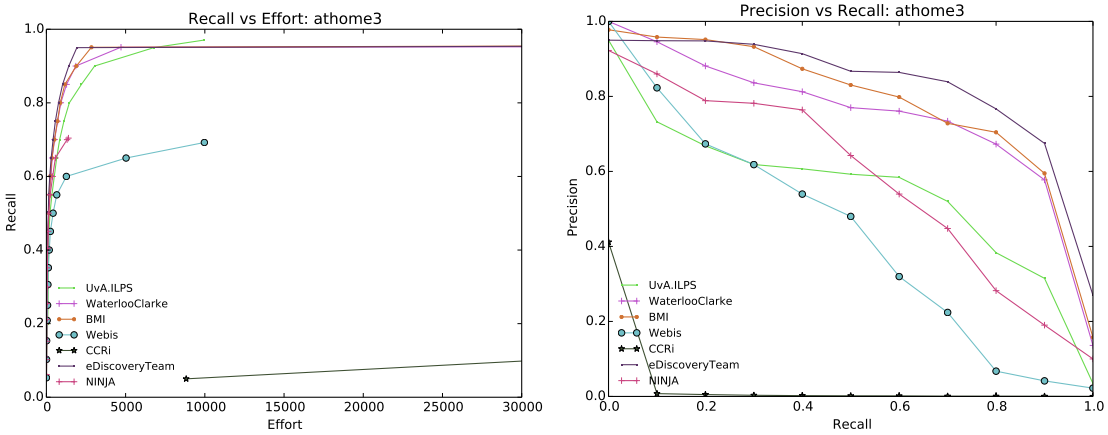


Figure 4.7: Average gain and precision-recall curves for the `athome3` collection.

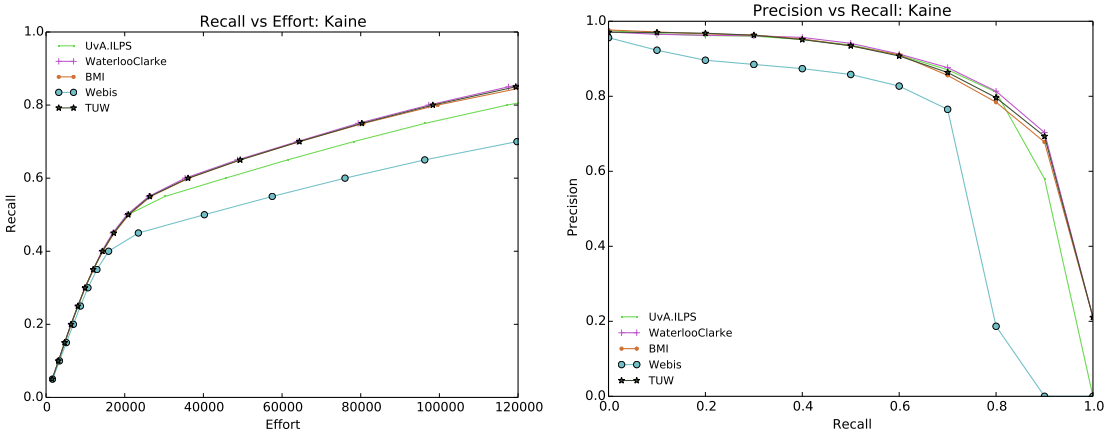


Figure 4.8: Average gain and precision-recall curves for the `Kaine` collection.

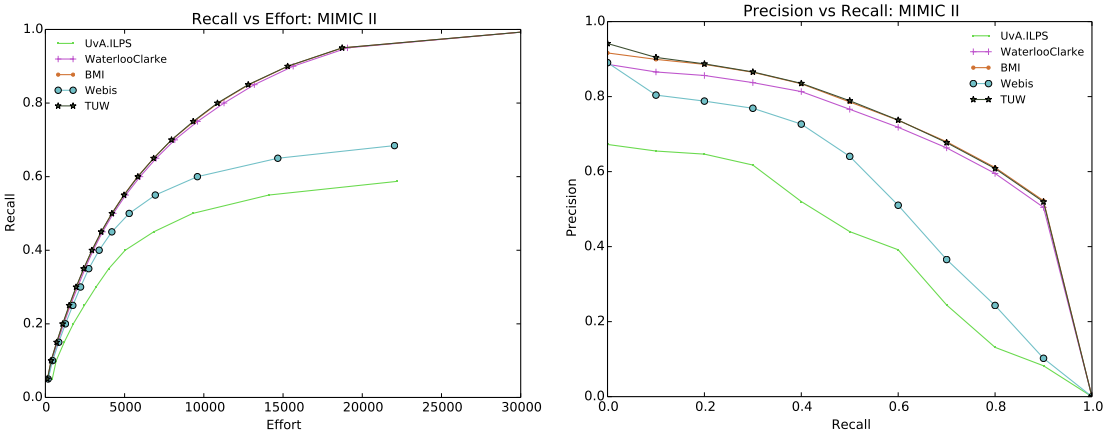


Figure 4.9: Average gain and precision-recall curves for the `MIMIC` collection.

collection is a much more complex scenario as there were only 4 topics: 2 of them had Rs exceeding 100,000 documents and the other two topics were around 20,000 relevant documents apiece. Accordingly, the average gain curve is positively influenced early on by these "smaller" topics but negatively influenced by the more prevalent topics later on because that not much progress can be made quickly (i.e., the dramatic knee in the curve). The `Kaine` case highlights a potential disadvantage of the average gain curve: the curve is easily influenced by the prevalence of the underlying topics. This issue is discussed further in Chapter 5 and a potential improvement is suggested. For posterity the per-topic gain curves for the best-performing runs are presented in Appendix A but results from per-topic are hard to generalize due to the vast amounts of information presented.

Turning our attention to the precision-recall curves, we see much more differentiation between the submissions; however, as with the gain curves, there does not appear to be a consistently clear "winner." For the `At-Home` collections, many systems appear to achieve in excess of 80% recall while maintaining what would generally be considered good precision ($> 60\%$). The `Kaine` precision-recall curve may lend further credence to the idea that an average gain curve may not be suitable a suitable: all but one system is able to achieve 80% recall with $> 80\%$ precision, which one might assume not to be the case if one examined only the gain curves.

Much of the behaviour of `Kaine` is due to the dichotomy of the topics: we have two very high-prevalence and two lower-prevalence topics (but still relatively high in comparison to the `At-Home` collections), which means that the apparent knee in the gain curve for `Kaine` is occurs only when the two smaller topics are mined out and the larger topics are being

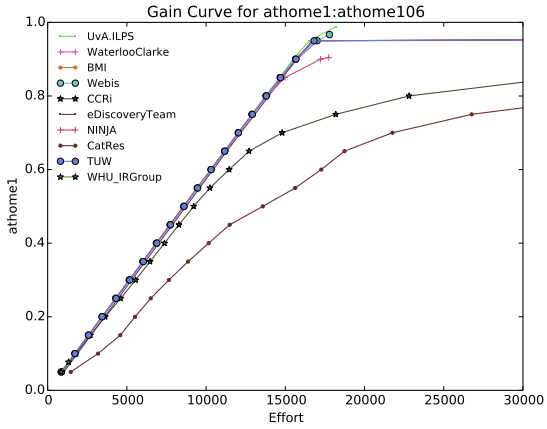
finished.

Gain curves and precision-recall curves both provide useful insight into the behaviour of high-recall retrieval systems, so it is not clear that either measure alone is sufficient. However, we do see, surprisingly, much more variation among systems when examining precision-recall curves than when comparing systems against the corresponding gain curves. Much of this may have to do with the averaging process for gain curves, where averages topics at fixed levels of effort that are intolerant of per-topic differences. Accordingly, the average at 10,000 documents' worth of effort may mean that some topics are nearly finished and others have barely started for a particular system—whether this is due to high prevalence or topic difficulty is moot in this case. Thus, the resulting average gain curve could be biased to either under- or over-estimate the performance of a system, depending on one's point of view.

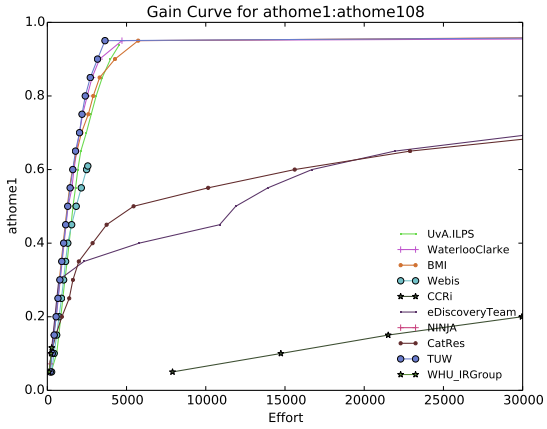
For example, Figure 4.10 takes two apparently "easy" topics from the `athome1` and `athome3` collections, subfigures (a) and (c), and compares them against two seemingly "harder" topics from the same collections, subifgures (b) and (d). The average gain curves for these collections should account for these large variations in performance, but it is not necessarily apparent that they do. Consider the the curve for Figure 4.10(a), the curve is almost ideal until just after 17,000 documents, where most systems have their knee form. This performance is almost entirely due to the fact that the vast majority of the 17,135 emails are automatically generated emails exhorting the former governor to intervene in the case of Terri Schiavo. It is only a very small portion that consist of substantively different content from those generated emails. On the other hand, the initial curves for most systems in Figure 4.10(b) are pretty good but not ideal, probably close to a slope of 0.5 for most systems, and then all systems, if they finish, have an extremely long tail. How an average gain curve can accomodate this different behaviour is not clear. The example topics from the `athome3` collection, subfigures (c) and (d) in Figure 4.10, offer an even more drastic example of divergent behaviour across topics, even when they have similar prevalence.

Ideally, an effective system should perform well with respect to both gain curves and precision-recall curves, while a meaningful measure would capture both the absolute effort required and effort relative to the number of relevant documents found (i.e., precision) for particular levels of recall. It is not clear that the gain curve accomplishes this and precision-recall curves provide no quantification of the effort involved. In the following section, we discuss the recall @ aR+b measure, which attempts to capture these two aspects and the results of the submitted systems on this measure.

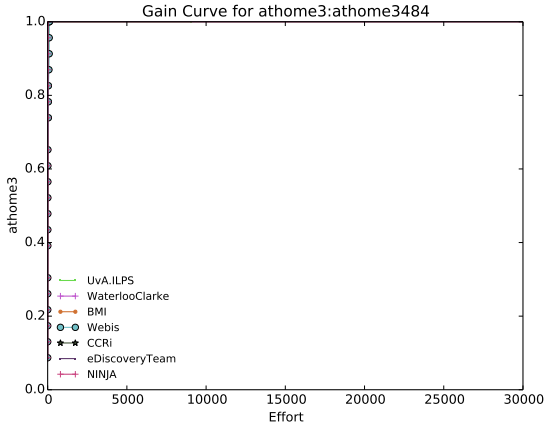
“Easier” Topics “Harder” Topics



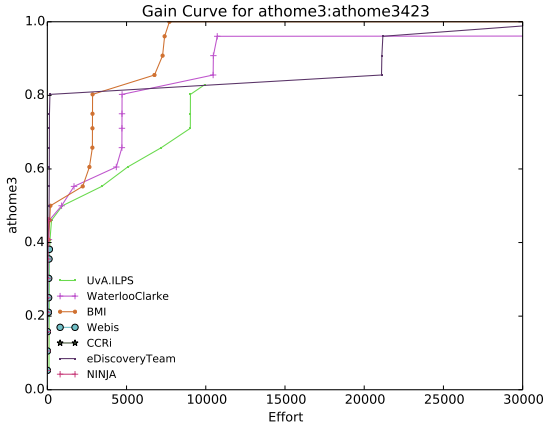
(a) Topic: Terri Schiavo (R: 17,135)



(b) Topic: Manatee County (R: 2,375)



(c) Topic: Paul and Cathy Martin (R: 23)



(d) Topic: Rob Ford Cut the Waist (R: 76)

Figure 4.10: Comparison of gain curves of two seemingly “easier” topics from the `athome1` and `athome3` collections with two seemingly “harder” topics from the same collections.

Recall aR+b Effort

Tables 4.5,4.6,4.7,4.8, and 4.9 report the average recall @ aR+b the values of a and b described in Section 4.4. As the WaterlooCormack runs differ only in when shots were called and by random chance, we report only one and label it BMI due to its baseline nature. Furthermore, we report all submissions rather than just the best performing.

Much like in the previous section, there we do not encounter a consistently superior system across collections or levels of effort. However, it does become apparent that some strategies that in their current implementation are inferior to others. By looking at those groups that submitted multiple variants on the same system, we can readily notice that such changes either resulted in little to no change (in the case of TUW) or had the potential to dramatically reduce effectiveness (in the cases of UvA.ILPS and Webis).

These results continue to show that semi-automatic systems (i.e., with a real human involved) are not necessarily superior to fully automatic systems. The eDiscoveryTeam went to great levels of human effort to achieve the performance they did and were only competitive, but not consistently superior to, the BMI and many of the participant BMI variants. This raises an interesting question: could BMI performance be improved by adding a live human into the loop to assist in the process? And if so, would the effort required be less than what was used by the eDiscoveryTeam? Such a question is very interesting but is not explored further in this thesis.

From these results and the curves, we can easily see that while the BMI may, on occasion, achieve lower recall at low effort when compared to other systems, but it easily regains lost ground and is very competitive at higher levels of recall, which is the area we are most interested in. Indeed, it appears that the BMI remains to be beat and will continue to pose a substantial challenge for parties interested in high-recall retrieval in the coming years. However, such a challenge may dissuade additional participation because the problem is either "solved" or the BMI provides too much competition and attempting to beat it seems like a fruitless task.

Facet-Based Evaluation

Based upon the results discussed above, we thought that looking at facet-based measures may illuminate more meaningful differences between systems than does binary relevance. For example, consider topic athome3423, which is plotted in Figure 4.10(d): several systems appear to have step function-like recall, which may be indicative of plateaus where a system attempts to identify the next relevant facet. Accordingly, a post hoc facetization of the `athome1` collection was undertaken after the TREC 2015 conference. Facets

Run	1R+0	1R+100	1R+1000	2R+0	2R+100	2R+1000	4R+0	4R+100	4R+1000
catres-attemptone	0.4958	0.5050	0.5686	0.6536	0.6571	0.6910	0.7457	0.7484	0.7712
CCRi	0.2329	0.2362	0.2597	0.3087	0.3101	0.3197	0.3639	0.3643	0.3721
eDiscoveryTeam	0.7832	0.7924	0.8194	0.8424	0.8434	0.8528	0.8619	0.8625	0.8731
NINJA	0.5768	0.5997	0.6198	0.6187	0.6187	0.6201	0.6201	0.6201	0.6201
TUW-1NB	0.7097	0.7404	0.8155	0.8412	0.8432	0.8611	0.8695	0.8706	0.9548
TUW-1SB	0.7135	0.7373	0.8139	0.8383	0.8417	0.8587	0.8689	0.8733	0.9664
TUW-1ST	0.7151	0.7349	0.8984	0.8862	0.9050	0.9522	0.9664	0.9665	0.9724
TUW-6NB	0.7168	0.7408	0.8170	0.8413	0.8440	0.8597	0.8708	0.8719	0.9613
TUW-6SB	0.6304	0.6538	0.7337	0.7530	0.7616	0.7778	0.7888	0.7907	0.8739
TUW-6ST	0.7100	0.7272	0.8923	0.8661	0.8880	0.9500	0.9637	0.9649	0.9718
UvA.ILPS-BASELINE2	0.4520	0.4672	0.5952	0.6691	0.6779	0.7215	0.7266	0.7270	0.7349
UvA.ILPS-BASELINE	0.7208	0.7397	0.8366	0.8462	0.8495	0.8608	0.8582	0.8588	0.8625
WaterlooClarke-UWPAH1	0.7616	0.7842	0.8497	0.8759	0.8793	0.8867	0.8916	0.8971	0.9878
WaterlooClarke-UWPAH2	0.7613	0.7810	0.8477	0.8682	0.8742	0.9000	0.9097	0.9218	0.9872
BMI	0.7075	0.7358	0.9038	0.9006	0.9172	0.9561	0.9677	0.9701	0.9744
Webis-Baseline	0.6277	0.6353	0.6589	0.6611	0.6611	0.6611	0.6611	0.6611	0.6611
Webis-Keyphrase	0.5599	0.5682	0.5881	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898
WHU_IRGroup-iterative_expansion	0.0537	0.0537	0.0537	0.0537	0.0537	0.0537	0.0537	0.0537	0.0537

Table 4.5: Average recall @ aR+b effort for the `athome1` collection.

Run	1R+0	1R+100	1R+1000	2R+0	2R+100	2R+1000	4R+0	4R+100	4R+1000
CCRi	0.1513	0.1674	0.2611	0.2164	0.2300	0.3004	0.3016	0.3076	0.3494
eDiscoveryTeam	0.6466	0.7319	0.8933	0.8586	0.8913	0.9508	0.9422	0.9484	0.9689
NINJA	0.4139	0.4716	0.6261	0.5347	0.5764	0.6261	0.6255	0.6261	0.6261
UvA.ILPS-BASELINE2	0.2603	0.2790	0.4684	0.4294	0.4502	0.6173	0.6152	0.6301	0.7330
UvA.ILPS-BASELINE	0.5014	0.5882	0.7989	0.7347	0.7651	0.8540	0.8370	0.8442	0.8729
WaterlooClarke-UWPAH1	0.7003	0.7733	0.9151	0.8961	0.9151	0.9595	0.9596	0.9642	0.9780
BMI	0.6699	0.7351	0.8830	0.8559	0.8768	0.9385	0.9409	0.9469	0.9705
Webis-Baseline	0.3281	0.3556	0.3767	0.3767	0.3772	0.3813	0.3816	0.3834	0.3913
Webis-Keyphrase	0.3228	0.3315	0.3533	0.3522	0.3534	0.3577	0.3581	0.3586	0.3642

Table 4.6: Average recall @ aR+b effort for the `athome2` collection.

Run	1R+0	1R+100	1R+1000	2R+0	2R+100	2R+1000	4R+0	4R+100	4R+1000
CCRi	0.0095	0.0119	0.0228	0.0156	0.0166	0.0263	0.0237	0.0238	0.0315
eDiscoveryTeam	0.8225	0.9235	0.9620	0.9341	0.9566	0.9634	0.9579	0.9631	0.9658
NINJA	0.6020	0.7042	0.7042	0.6696	0.7042	0.7042	0.7042	0.7042	0.7042
UvA.ILPS-BASELINE2	0.2704	0.3239	0.5221	0.3936	0.4290	0.6403	0.5500	0.5675	0.7143
UvA.ILPS-BASELINE	0.4686	0.7199	0.8802	0.7212	0.8334	0.9187	0.8021	0.8717	0.9263
WaterlooClarke-UWPAH1	0.7357	0.8514	0.9269	0.8732	0.8920	0.9316	0.9017	0.9017	0.9340
BMI	0.7759	0.8432	0.9353	0.8770	0.9037	0.9418	0.9141	0.9217	0.9480
Webis-Baseline	0.4320	0.5579	0.6186	0.5131	0.5867	0.6195	0.5707	0.6129	0.6696
Webis-Keyphrase	0.4118	0.4441	0.4560	0.4345	0.4522	0.4581	0.4570	0.4585	0.5129

Table 4.7: Average recall @ aR+b effort for the `athome3` collection.

Run	1R+0	1R+100	1R+1000	2R+0	2R+100	2R+1000	4R+0	4R+100	4R+1000
TUW-1NB	0.7872	0.7880	0.7990	0.9624	0.9627	0.9645	0.9900	0.9900	0.9902
TUW-6NB	0.7987	0.7998	0.8092	0.9657	0.9659	0.9669	0.9903	0.9903	0.9905
TUW-6SB	0.7980	0.7990	0.8078	0.9668	0.9669	0.9679	0.9907	0.9907	0.9908
TUW-1SB	0.7893	0.7905	0.8030	0.9630	0.9631	0.9643	0.9904	0.9904	0.9906
TUW-1ST	0.8007	0.8018	0.8119	0.9675	0.9677	0.9688	0.9914	0.9914	0.9916
TUW-6ST	0.8035	0.8047	0.8165	0.9692	0.9693	0.9701	0.9916	0.9916	0.9918
UvA.ILPS-baseline	0.8116	0.8131	0.8240	0.9213	0.9213	0.9213	0.9213	0.9213	0.9213
WaterlooClarke-UWPAH	0.8115	0.8125	0.8239	0.9723	0.9724	0.9735	0.9927	0.9927	0.9928
BMI	0.7954	0.7966	0.8078	0.9684	0.9686	0.9694	0.9915	0.9915	0.9916
Webis-baseline	0.7362	0.7377	0.7479	0.7814	0.7814	0.7814	0.7814	0.7814	0.7814
Webis-keyphrase	0.5565	0.5579	0.5679	0.6034	0.6034	0.6034	0.6034	0.6034	0.6034

Table 4.8: Average recall @ aR+b effort for the `Kaine` collection.

Run	1R+0	1R+100	1R+1000	2R+0	2R+100	2R+1000	4R+0	4R+100	4R+1000
TUW-1NB	0.6884	0.7025	0.7485	0.8943	0.8961	0.9119	0.9730	0.9735	0.9776
TUW-1SB	0.6907	0.7041	0.7489	0.8943	0.8961	0.9111	0.9725	0.9730	0.9766
TUW-1ST	0.6945	0.7057	0.7516	0.8957	0.8986	0.9126	0.9739	0.9743	0.9772
TUW-6NB	0.6806	0.6950	0.7408	0.8891	0.8913	0.9072	0.9708	0.9712	0.9751
TUW-6SB	0.6848	0.6966	0.7427	0.8896	0.8914	0.9065	0.9711	0.9714	0.9752
TUW-6ST	0.6886	0.7000	0.7437	0.8913	0.8935	0.9074	0.9723	0.9728	0.9760
UvA.ILPS-baseline2	0.5156	0.5182	0.5695	0.5545	0.5650	0.5872	0.5872	0.5872	0.5872
UvA.ILPS-baseline	0.4529	0.4779	0.4894	0.4899	0.4899	0.4899	0.4899	0.4899	0.4899
WaterlooClarke-UWPAH	0.6841	0.6946	0.7422	0.8894	0.8921	0.9080	0.9705	0.9713	0.9745
BMI	0.6971	0.7072	0.7529	0.8965	0.8983	0.9132	0.9735	0.9739	0.9768
Webis-baseline	0.6241	0.6365	0.6620	0.6803	0.6812	0.6844	0.6838	0.6844	0.6844
Webis-keyphrase	0.5131	0.5260	0.5385	0.5588	0.5588	0.5588	0.5588	0.5588	0.5588

Table 4.9: Average recall aR+b effort for the `MIMIC` collection.

were created for the athome1 collection for each document by using the combination of its sender (e.g., "Homer Simpson") and its recipients (e.g., "Marge Simpson" and "Ned Flanders") as a single facet label for that document, which textually might be represented as From:HS-To:MS-NF. Since these facets are topic agnostic, they do not correspond to the traditional IR conception of a subtopic. Of all possible facet labels, 94 had at least one relevant document for any of the 10 topics used in the athome1 collection.

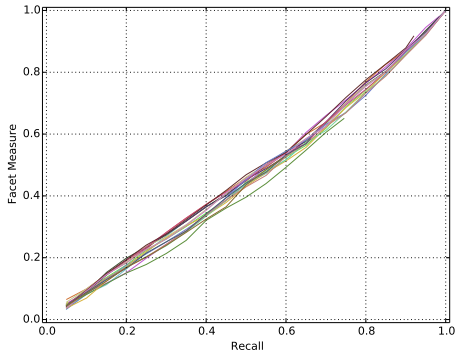
Due to limited resources and uncertainty about the quality of these facets, no further facetization took place for the athome2 and athome3 collections (nor was it clear what type of facets should be used for these collections). The Kaine collection had only 4 topics, so the utility of creating facets for this collection is unclear due to the lack of inferential power. MIMIC may look at first glance to be promising, because the 19 ICD-9 codes can be naturally broken down into existent subcodes. The small size and the extremely uniform performance of runs on MIMIC, as evidenced in Appendix A, led us to believe the utility of such facets would be low.

Using the athome1 facets, we explored various facet-based evaluation measures, including: mean facet recall, which is the average facet recall over all topics; a relaxed variant of the CubeTest [102];[21] subtopic recall [167], where facets are taken as subtopics; and the 10th and 15.8th[22] percentiles of facet recall. Figure 4.11 compares binary recall (i.e., relevant or not) to the various facet-based measures for all of the Total Recall 2015 runs, averaged across the athome1 topics. By and large, what we can observe is that none of these measures differs all that dramatically from binary recall. The 10th and 15.8th percentiles appear the most promising, due to the fact that they appear to correlate the least well with binary recall. We compare percentiles and mean facet recall at two levels of effort, R and $2R + 1000$, in Figure 4.12 for all submitted systems. We use mean facet recall as a point of comparison due to its very similar behaviour to binary recall. It is immediately apparent that there is not much difference between the measures. While there are some minute differences, there are no sweeping changes that would lead us to believe one system is optimizing for facets, or at least, not for the facets we've created. Indeed, Cormack and Grossman [42] have shown that the greedy approach employed by continuous active learning, and the BMI, are able to achieve high facet recall and high overall recall simultaneously which is what these results have also shown. Similarly, Wang and Zhu have shown [156] that when ranking documents by decreasing likelihood of relevance, the top 20 or so documents cover most subtopics.

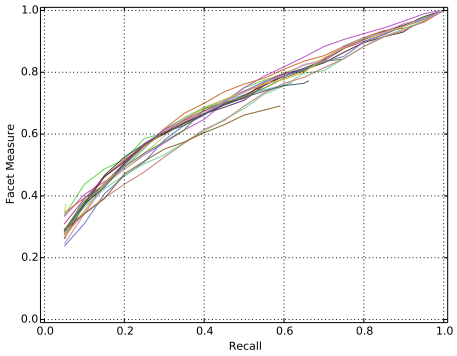
At this point it is not clear whether the lack of information in this facet evaluation

[21]Note that when used for high-recall purposes, the CubeTest effectively devolves into a simple gain function where per-facet gain decreases as more examples of that facet are found.

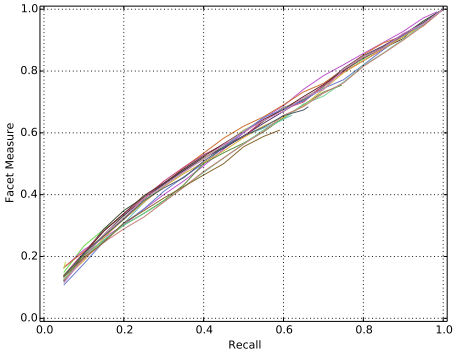
[22]This corresponds to 1 standard deviation.



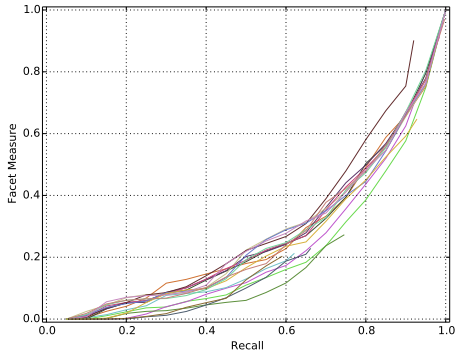
(a) Mean Facet Recall



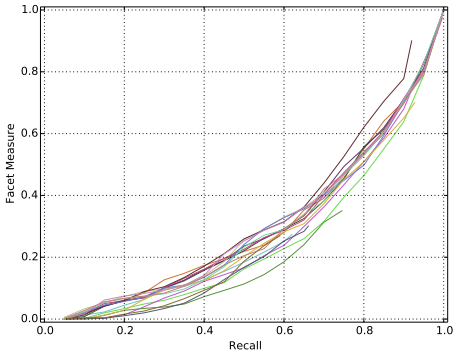
(b) Subtopic Recall



(c) CubeTest

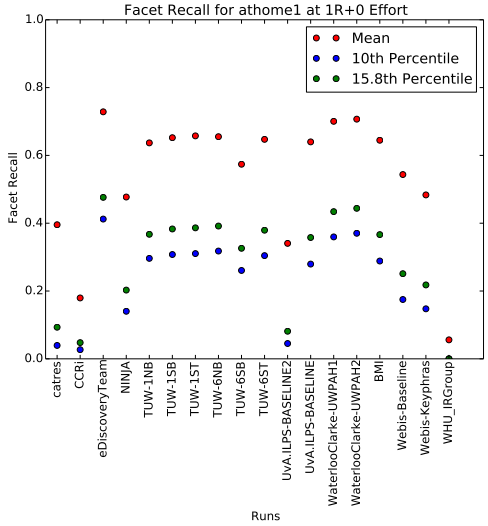


(d) 10th PCTL

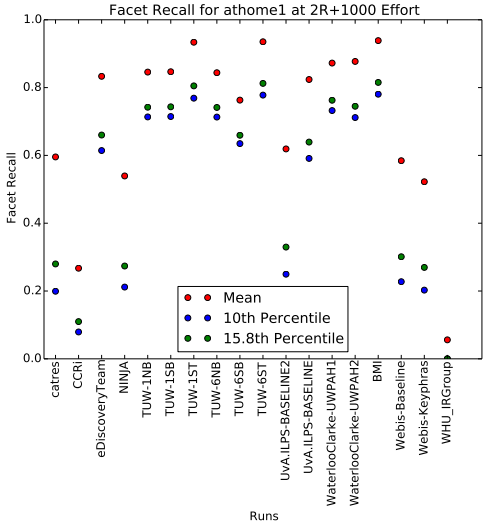


(e) 15.8th PCTL

Figure 4.11: Comparison of binary recall to various measures of facet recall for the `athome1` collection over the submitted runs to Total Recall 2015.



(a) R effort



(b) $2R + 1000$ effort

Figure 4.12: Comparison of mean facet recall, the 10th percentile and 15.8th percentile of facet recall at two levels of effort: R and $2R + 1$.

derives from the measures' being poorly suited to high recall or the facets themselves' not being particularly good. Indeed, we might naturally expect emails involving Jeb Bush as sender or recipient to be retrieved first, but this does not appear to be the case. This perhaps follows logically from the fact that such documents were not awarded any higher importance or relevance, which would mean that a system would not score more "goals" for returning such documents first. This lack of incentive may be what caused the depth-first retrieval.

As part of the Total Recall 2016 assessment effort, facets were created and used by assessors to group related emails for an additional 34 topics in the `athome1` collection. Three levels of relevance were also used: relevant and important, relevant and not important, and not relevant. But these relevance grades were reduced to the same binary relevance judgments given in the 2015 iteration so that participants did not have to substantively modify their systems to accommodate that change. Whether such in situ facetization and graded relevance will reflect meaningfully different results from those presented here will have to be investigated. Indeed, the use of binary rather than ternary relevance assessments may mean that the results using these new facets are not substantially different.

"Calling Your Shot"

We have omitted a discussion of the "call your shot" results, primarily due to limit of spaces, but also because "calling one's shot" was an optional subtask in the 2015 iteration of the track and as such, not all teams called a shot or called them consistently. Accordingly, it is not apparent how useful such an analysis would be in the context of this thesis given that only F1, Recall, and Precision were reported for those teams that participated in the subtask. Interested parties are directed to the official track overview [120] for a more detailed discussion of the "call your shot" results. It is worth noting that Total Recall 2016 has made the subtask mandatory and work is ongoing in the development of meaningful metrics for such evaluation, including those based upon the proposed Recall-Loss and Effort-Loss of Cormack and Grossman [44].

4.6 Discussion

As with the development and use of any long-running software, issues were likely to arise throughout the course of the track and did. The biggest issue was a bottleneck in document assessment requests near the submission deadline. This was primarily due to the vanilla installation of MySQL, which became bogged down when many queries were being run. In hindsight, this issue is relatively easy to solve by configuring MySQL to handle queries more effectively (e.g., increasing resources allocated to query processing) and by adding connection pools to the Total Recall server so that multiple requests are not held up by a single connection to the database.

Surprisingly, use of a REST(ful) API did not appear to cause participants too much trouble, in spite of our fears during development of the architecture that it might. However, problems may have been mitigated by providing tools such as the BMI and the manual participation interface. Accordingly, adoption of similar APIs for other TREC tracks may facilitate investigation of increasingly more "real-world" situations. For example, the Real-Time Summarization track[23] at TREC 2016 is using a similar API [123] to enforce real-time submission of temporally relevant results, and the Live QA track [9] has required participants to implement their own limited API for question answering.

Another design decision to consider for future iterations is the use of full virtual machines. There are many merits to using virtual machines, primarily that the entire system is self-contained, so security issues are more manageable (as access to the underlying hardware is controlled). However, full virtualization has definite performance setbacks. For

[23]Track details can be found at: `http://trecrts.github.io`

one of the Sandbox collections, we could only have one participant system running at a time due to the high overhead of both the participant's system and the virtualization. An alternate solution may be to use a more lightweight system, such as Docker, which makes use of software containers. In short, software containers package an application together with its dependencies for running on an arbitrary (Linux) server. The easiest way to envision software containers is as somewhere between virtual machines and auto-building tools (e.g., make, Maven). Accordingly, less emulation overhead occurs with containers but at the expense of greater exposure to the underlying system. Whether or not Docker provides a suitable replacement going forward is an avenue that should be investigated from both security and efficiency perspectives.

Our current Total Recall architecture has been designed to limit and mitigate the potential for any directly malicious attempts to use the Internet to transmit sensitive material outside of the architecture (e.g., a system downloads a collection and transmits it to a malicious party). We have also noted earlier, in Section 4.3.2, that by limiting what is returned to participants (e.g., only summary results and not entire ranked lists with assessments), we can also limit potential information leakage. In spite of this, a concerted malicious user could likely construct a system solely to identify the presence of desired information (e.g., whether or not the Kaine emails contain discussion of the 2012 Total Recall film) using only these summary measures. Designing evaluation metrics that prevent this type of information leakage was outside the scope of the 2015 iteration of the Total Recall track and, correspondingly, this thesis. Such metrics would be of great use to furthering experiments on private data and is worthy of further investigation.

Aside from purely architectural concerns, there exist other potential issues with the 2015 Total Recall track. In particular, relevance assessments were garnered by a single assessor with no quality assurance process by a historical predecessor of the BMI. While this did not appear to have any substantial impact on system performance, it did result in at least some known false positives (documents discussing "Manatee Springs State Park") for a single topic ("Manatee County") due to the assessor's lack of knowledge about the geography of the state of Florida. It is worth noting that this is approximately a false positive rate of 1%, which, while not ideal, does not seem unreasonable. This belief is further confirmed when the Sandbox collections are considered, as they were judged independently of the track itself and, thus can be viewed as an unbiased sample. Given that the BMI performed as well as or better on the Sandbox than it did on the At-Home collections, it is likely the case that there was no undue influence of the assessment protocol on the results of the At-Home experiments. That being said, the Sandbox collections should not be considered paragons of evaluative virtue, as they too had incorrect assessments.

As part of a follow-up with the Kaine collection custodians at the Library of Virginia,

the track coordinators turned over a selection of documents for which the BMI disagreed with the gold standard assessor, where disagreement was determined by the BMI returning a high relevance score (i.e., returned as part of an early batch) for the document. After a re-review of these 187 discordant documents by the original annotator: 66 documents could have had more than one category applied, 27 documents did not have a label change; and, 94 had a label change. This should not be taken as a statistical result, merely as an indicator that there were flaws in the `Sandbox` collection and that perhaps no test collection will ever be "perfect." Due to these issues, the 2016 Total Recall track is adding in a quality-assurance phase and is making facet annotation part of the core assessment task. In addition, independent NIST assessors will be judging topics that are selected by the track coordinators to help create less bias in the topics and relevance assessments. The goal of these additional measures is to ensure that the collections used in the 2016 iteration of the track are of as high quality as possible.

The previous chapter would indicate that we might see different results (i.e., decreased recall for the same effort) if the participant systems were evaluated against a secondary set of assessments. Such an experiment would be possible if such an additional set existed. It might be possible to use a manual submission as the basis of such a set. The relevant/not relevant cutoff would then be the called shots of one of the runs. However, it is not clear that such a cutoff corresponds to the manual assessor's conception of relevance but may well correspond to the assessor's interpretation of the server's conception of relevance. Such a cutoff is then targeting a specific authority rather than the hypothetical "reasonable" authority that was advocated for in the previous chapter. Accordingly, we do not report such an experiment herein and leave it for future research.

While we have not seen a strategy that wins 100% with respect to any measure, the 2015 Total Recall track has piloted an investigation into how to improve the state of the art with respect to high-recall retrieval systems as well as to their evaluation. The "one true" metric was not discovered, but we have observed how different measures can be used to fully flesh out system performance. Accordingly, it may not be the case that a single such summary measure is needed, but rather that a set of metrics may be required to truly grasp the totality of high-recall effectiveness.

Chapter 5

An Exploration of Effectiveness Measures for High-Recall Retrieval

The gain curve has become a popular method of evaluating high-recall retrieval systems because it tells readers, at a glance, that if they spend X effort, they can reasonably expect to achieve Y recall. On a per-topic basis, such a formulation is reasonable; however, averaging across topics becomes potentially problematic due to per-topic variance dramatically affecting the shape of the curve. Figure 5.1 depicts this case for two topics from the Reuters Collection Volume 1 (RCV1) and two baseline systems described later in this chapter (Section 5.2). The nature of the problem is as follows: for some topics, an arbitrary effort cutoff can correspond to some topics barely being completed and, at the same time, to other topics' having been completed earlier. Table 5.1 shows a hypothetical cutoff where the average recall for all systems is 75% but also shows the effect of potential per-topic variation. Some of the issues with gain curves were discussed previously in Chapter 4; namely, in regards to the `Kaine` collection where a similar situation to Table 5.1 occurred.

This chapter presents a case study of the pros and cons of gain curves when compared to other evaluation metrics, with a focus on measurement consistency and the ability to distinguish systems. These metrics are described in Section 5.1 and are used throughout the remainder of the chapter.

Our case study begins with a baseline analysis of these measures on two existing test collections: RCV1 and the TREC-6 ad hoc test collection (Sections 5.2.2 and 5.2.3 respectively). These two collections allow us to examine the effects that different test collections and metrics can have when comparing two baseline systems (random sampling and relevance sampling; see Section 5.2.1 for more details) while providing enough topics to

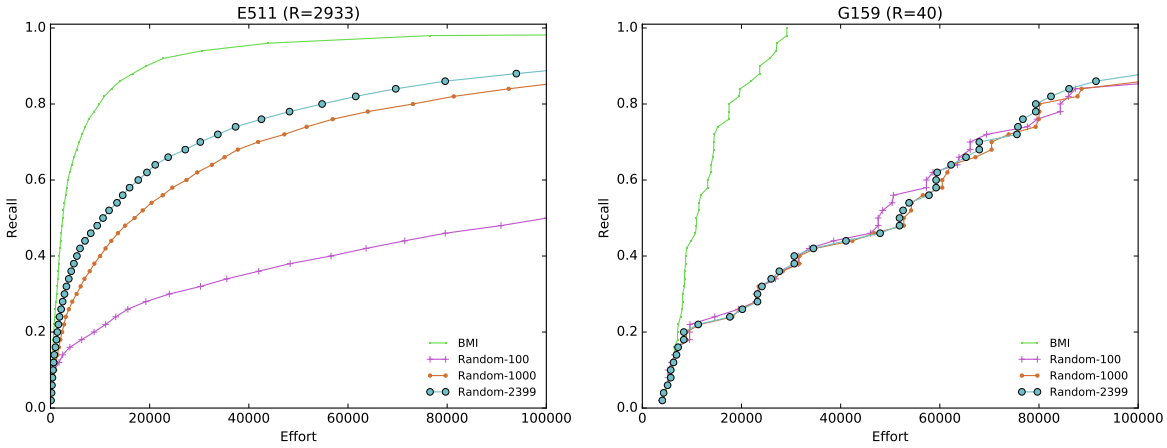


Figure 5.1: Comparison of gain curve variation for two topics from the RCV1 baseline experiments (Section 5.2).

facilitate bootstrap generation of confidence intervals for several of the metrics used.

In Section 5.1, based upon our baseline experiments, we compare a selected set of metrics with the actual systems submitted to the 2015 Total Recall track (described in Chapter 4) and determine the "real-world" applicability of our results. Due to the relative paucity of topics for any single Total Recall collection, we are not able to generate meaningful bootstrap results, though such results may be possible with subsequent Total Recall collections. We conclude with a discussion of the results and potential avenues of further investigation (Section 5.4).

5.1 Evaluation Measures

This section provides brief descriptions of the evaluation measures we use in this chapter, summarizing previously discussed measures and introducing new ones. We break these measures into curve-based and traditional IR summary measures to provide more contextually appropriate descriptions. We also introduce the idea of using root-mean-square error (RMSE) as a means of measuring the consistency of systems when their performance is plotted on a curve with a desired target (i.e., 100% recall).

5.1.1 Curves

In these studies, we focused on curves which attempt to model the trade-offs between effort and recall. While measures like the receiver operating characteristic (ROC) curve have been used to evaluate high-recall retrieval tasks, there is no explicit effort component involved,[1] so we omitted it from our analysis. Accordingly, our baseline curve for comparison was the gain curve.

The most straightforward alternate to the gain curve is to measure the number of relevant documents retrieved for a given amount of effort. Such a formulation tells us explicitly how many relevant documents we might expect to find (on average) for a given level of effort. The result is similar to the gain curve but gives a concrete number in terms of documents retrieved rather than some amount of recall achieved. Note that such a curve has a variable y-axis, since it is dependent on the topic's R. This means that the averaged curve has a y-axis maximum of the average R across the collection, which may make cross-collection comparison less desirable than when using standard gain curves (which all end at 100% recall).

An alternate measurement of effort would be the amount of effort expended *relative* to R (i.e., $\frac{effort}{R}$), which we simply call relative effort. Note that this measurement effectively solves for a in the formula effort $= aR + b$ when $b = 0$ from Chapter 4. While we could formulate relative effort with the fixed overhead parameter, b, we omitted this for simplicity and to avoid inclusion of many more plots. Such exploration is left for future work.

The relative gain curve simply plots the recall achieved for a given amount of relative effort, with the intuition that, across topics, recall should be more consistent at similar levels of relative effort for "good" systems. An average relative gain curve is then simply computed by averaging recall across topics for all relative effort values. This results in computing recall at comparable points across topics (e.g., 0.1R, 1R, 10R), rather than at arbitrary levels of effort that do not take into account per-topic prevalence.

One of the desirable properties of relative effort is that an ideal system would achieve 100% recall at a relative effort of 1. Contrast this to a gain curve, where there is no corresponding point other than identifying where R is on the curve. Furthermore, when computing the average gain curve of an ideal system, an average of 100% recall would only be observed once the effort achieved the maximum R in the test collection. Of course, in the average relative gain curve, this still happens at exactly a relative effort of 1.

[1] Thus, potentially making it harder to reason about from a pure cost perspective, i.e., how much more recall would X effort get?

These three curves all tell users for a specified level of effort, how much relevant material they can consume. When taken as an average across topics, the curves tell users the expected amount of relevant material they will consume by applying the same amount of effort on each topic. It would be ideal if, in expectation, rendering the same amount of effort across topics yielded similar levels of relevant material produced. Per-topic prevalence, however, can greatly skew the evaluation as one topic may have not had enough documents reviewed to achieve particular levels of recall. Accordingly, the use of relative effort seeks to ameliorate that issue by measuring recall at equivalent points of effort expended by an assessor.

Our final alternative curve is the recall depth curve from Chapter 3, which measures effort as a function of recall (i.e., for X recall, how much effort do we need to expend?). This is effectively the other side of the coin of the gain curve but may have more interesting results when averaged, since it models the situation in which a user is targeting a specific recall level and is asking for the amount of effort they have to expend. This expenditure can either be per topic or in expectation across topic. In terms of Cormack and Grossman's work on reliability [44], users requires a 100% reliable system for their particular recall cutoff and wish to know how much assessing effort it might cost to achieve that level of recall and reliability.

For simplicity and ease of comparison with the other curves, average recall depth curves were calculated with the arithmetic rather than geometric mean, differing in this way from the computation in Chapter 3. However, we still calculate the percentage of the corpus reviewed rather than the raw effort to maintain a reasonable scale on the y-axis.

While these are not all the possible formulations of possible curves, we believed they were sufficient to provide an exploration of the evaluation space without being overly verbose or repetitive. In particular, we could have reformulated recall depth curves to measure relative effort rather than absolute depth of review.

Measuring Consistency

Table 5.1 depicts 4 hypothetical systems, all achieving 75% recall on average. System A is perfectly consistent and Systems B, C, and D, depict various potential inconsistent systems. If we examine the per-topic averages, we can see that some systems are more (in)consistent than others. On the other hand, averaged summary measures provide a convenient way to determine how well, or how poorly, a system is performing at glance. They can also be misleading. Average recall is convenient because it reveals how well a

System	T1	T2	T3	T4	RMSE
A	75	75	75	75	25.0
B	50	50	100	100	35.4
C	75	75	70	80	25.2
D	60	80	80	80	26.5

Table 5.1: Four hypothetical systems achieving an average of 75% recall for four hypothetical topics at some level of (relative) effort. The per-topic recall, as a percentage, and the corresponding root-mean-square error (RMSE), as a percentage, is provided.

system is retrieving documents on average, but, as Table 5.1 depicts, the same average can be achieved from vastly different per-topic levels of recall.

Root-mean-square error (RMSE) is a method to measure the accuracy of a set of measurements. Typically, RMSE is used to measure differences between sample and population values, but we can adapt it for use in evaluating high-recall retrieval systems. The target recall of any high-recall system is 100%, so for a topic we can model the recall left behind, the loss, by simply subtracting the achieved recall from 100% (or, equivalently, from 1 when not dealing with percentages). This gives the equation $loss = 1 - recall$, which when averaged over topics yield the formula $avg(loss) = 1 - avg(recall)$. This formulation is not interesting, since it is simply the complement of a computation we already know. However, as suggested by Cormack and Grossman [44], we might more realistically model this as a quadratic loss function, $quadratic\ loss = (1 - recall)^2$, in the vein of Taguchi quality loss [143]. In doing so, we are effectively measuring the magnitude of error in a system's recall across topics (i.e., taking the mean of quadratic loss would produce the mean-square error, which can have its square root taken to scale the value correctly). Thus, we can use RMSE as an indicator of system consistency. It is worth noting that if, instead of a target of 1, we set the target to be the mean recall then MSE would be calculating the variance of the system's recall (and RMSE would be its standard deviation). However, we avoid edthe use of standard deviation in this chapter as we wish to view the distribution without making any normality assumptions.

Revisiting Table 5.1, we can see that RMSE, at a glance, provides us the similar information to that which can be gleaned from examining per-topic recall does. In particular, System A is perfectly consistent, followed by systems C, D, and B, which one might generally consider a good ranking of systems. Furthermore, as RMSE is a single number, we can plot it as a function of (relative) effort. In doing so, we can then easily compare the consistency of systems using these resultant curves.

Note that in the case of recall depth, we could set the target value of RMSE to 0 (i.e.,

we want 100% recall with 0% of the result list reviewed), but that is impractical and not even remotely realistic. An alternative would be to use the number of relevant documents for each topic in the RMSE calculation, but the meaningfulness of such a calculation is not entirely apparent. Accordingly, we left the investigation of RMSE's applicability to recall depth as an avenue of future work. A similar argument can be made when measuring the number of relevant documents retrieved.

It is worth noting that RMSE was previously suggested as a means of evaluation in the 2010 Legal track's learning task.[2] However, the final track overview [46] does not discuss the measure nor why it was not used.

Furthermore, one might reasonably see connections between this idea of measuring the consistency of recall achieved and Wang and Zhu's work [156] of adapting economic portfolio theory [106] to information retrieval. In their work, they experimented with balancing the trade-off expected relevance of a ranked list and its variance (e.g., diversity of the ranking). The end goal was to maximize the mean relevance and minimize the variance with respect to a particular evaluation metric. Due to Wang and Zhu's focus on optimizing ad hoc retrieval runs, their work is not directly applicable to what we are investigating with RMSE, which is to measure the consistency of the high recall achieved across topics. Accordingly, their approachmay be more directly applicable (out of the box) to approaches like continuous active learning. However, it may be possible to adapt their model to more accurately measure system consistency across topic, which could be an interesting avenue of future investigation.

5.1.2 Summary Measures

The curve-based measures described above present an overall picture of a system's performance but are not easily consumable nor are they easy to incorporate into statistical tests. High-recall retrieval evaluation has made use of a variety of measures in an attempt to determine system effectiveness.

Two of the most commonly used are: the area under the ROC curve (AUC)[3], which effectively measures the likelihood that a relevant document will appear before a non-relevant document in a particular ranking; and, hypothetical F1, which reports the F1 score that corresponds to the optimal cutoff in the ranking, with the assumption that the

[2] As detailed in the task guidelines `http://plg.uwaterloo.ca/~gvcormac/legal10/legal10.pdf`.

[3] While we do not look the ROC curves, AUC provides a convenient summary measure with an intuitive meaning.

optimal cutoff was "known" prior to evaluation. Both measures have been used extensively in spam filtering [49, 36, 37] and eDiscovery tasks [46, 71, 163].

Mean Average Precision (MAP) [27, Chapter 2], while not strictly a high-recall-oriented measure, does evaluate over the entirety of a ranked list. Note that MAP is a discrete integral that approximates the area under the precision-recall curve. Though mathematically, it is simply the average of average precision. Average precision is just the expectation over possible user stopping points corresponding to relevant documents (i.e., the average over precision@i where i is the rank of the ith relevant document). Both averages assume uniform distributions. It suffices to say that while MAP has been effective for IR evaluation and comparison, it lacks an intuitive meaning [27, Chapter 2] similar to the case of AUC.

Recall @ $aR+b$ effort (defined in Chapter 4) can also function as a summary measure, so we included it in our analysis, using the same values as in Chapter 4. We also set b to 100 and 1000 with a = 0 as a means of examining more traditional evaluation contexts (i.e., early precision). Furthermore, we extend it to compute Precision and F1 $aR+b$ effort as additional possible measures. For each of these cutoffs, we also compute Drucker et al.'s coverage ratio [61], which produces, when effort is less than R, precision and, when effort is greater than R, recall. Such a measure tries to balance the inevitable trade-off between recall and precision as the traversal depth of a ranked list grows.

Cormack and Grossman have formulated high-recall retrieval evaluation in terms of recall-loss and effort-loss [44], which are formulated as:

$$\text{loss}_\text{r} = (1 - \text{recall})^2$$

$$\text{loss}_\text{e} = \left(\frac{b}{|C|}\right)^2 \left(\frac{\text{effort}}{R+b}\right)^2$$

The b parameter is analogous to the fixed overhead parameter of $aR+b$ effort, and C is the collection. In correspondence with Cormack, it was revealed that the b parameter was set to 1,000, so we used that in our experiments for consistency with prior work. Furthermore, they compute the average of recall-loss and effort-loss as $loss_{re} = 0.5loss_r + 0.5loss_e$. These proposed loss measures can be computed hypothetically (i.e., as hypothetical F1) or at given levels of effort (i.e., as in recall $aR+b$). However, loss_e is independent of system performance. so it is not, by itself, a meaningful measure for comparing systems. Accordingly, we omitted reporting the loss_e as it does not add anything to our analysis. Similarly, the hypothetical loss_r is always 0 for a complete ranking of a document collection, so we do not report this measure.

5.2 Baseline Experiments

In this section, we describe the results of running our baseline systems (relevance sampling and random sampling) over the RCV1 and TREC-6 test collections and of then comparing them with the previously discussed evaluation measures (Section 5.1). These two methods were chosen as they have previously been shown to be substantially different [40], and, more importantly, relevance sampling was shown to be superior to random sampling. Accordingly, a good measure should be one that can distinguish such approaches.

5.2.1 Baseline Systems

For our implementation of relevance sampling, we simply used the Total Recall track's Baseline Model Implementation (BMI), suitably augmented to not require the use of a virtual machine. Recall that the BMI is simply an optimized version of Cormack and Grossman's AutoTAR protocol [41] which extends their previous continuous active learning protocol [40] by using exponentially increasing batch sizes. Furthermore, the initial seed document is a pseudo-relevant document created from the topic statement and combined with a fixed number of randomly sampled documents treated presumptively as not relevant for the purposes of training. Other than removing the necessity of a virtual machine, we used BMI "out of the box."

Our random sampling implementation is a modification of the BMI such that for all topics we used a fixed random sample of size K and trained on this random sample (with the topic statement as a pseudo-relevant document) and perform one round of classification. Once this was done we used the scores generated by `sofia-ml` (the BMI's machine learner) to rank the collection and generate our final result list. We used two formulations of random sampling: optimistic random sampling, where the original random sample is allowed to be ranked with the rest of the corpus; and, simple passive learning (SPL), which follows Cormack and Grossman and places the random sample at the beginning of the ranking. Effectively, the optimistic case provided an upper bound on performance while SPL provided a more realistic approach. For the purposes of reporting results, optimistic random sampling was simply titled "random" and SPL, "SPL". Three values of K (100, 1000, 2399) were selected to allow us to examine the effect that sample size would have on these comparisons. Note that the value of K=2399 was selected due to its use in the literature as a "meaningful" random sample size [43].

5.2.2 RCV1

The RCV1 collection is comprised of 800,000 newswire articles that have been manually labelled for 103 hierarchical topic codes. Accordingly, there are 4 top-level topic codes which do not have any parents and of which all 99 other topic codes are children. For our purposes, a topic statement is the topic code's description concatenated with the description code of any of its ancestors in the hierarchy. For example, the top level code MCAT would have the topic statement "ALL Securities and Commodities Trading and Markets," and its descendant M143 would have the topic statement "trading in all commodities trading in all energy products ALL Securities and Commodities Trading and Markets."

RCV1 has a large variation in the numbers of relevant documents per topic, ranging from 5 to 381,327 documents, with an average of 25309.5 relevant documents per topic. Accordingly, the average prevalence of RCV1 is much higher than in most of the collections used in this thesis but provides a convenient testbed since, it is completely assessed.

Figure 5.2 plots the various curves for the baseline systems averaged over all 103 topics. As we might expect, BMI outperforms both random sampling approaches though the degree to which it is superior changes depending on the curve. In terms of recall depth and relevant documents retrieved, BMI appears to "win" by a slimmer margin than in the gain curves. As is expected, optimistic random sampling performs as well as or better than SPL for the same sample size. However, it is interesting to note that the relative gain curve is the only one of the 4 curves to consistently show SPL(2399) performing worse than Random(2399). The other 3 curves appear to show that SPL(2399) eventually catches up to Random(2399) in terms of performance. As we'll see, this will become a running trend for the remaining experiments and analyses.

If we compare the RMSE of the gain and relative gain curves (Figure 5.3), we again see the expected result that BMI is superior (has a smaller curve) than random sampling, regardless of the curve. But when we compare the two curves, we still see disagreement about the performance of SPL(2399); in fact, it now appears to strongly be no better than SPL(1000), which does not entirely disagree with the earlier results. It would appear that measuring recall as a function of relative effort picked up on behaviour that may be overshadowed in a gain curve using absolute effort, due to the masking effect of some topics being (nearly) complete while others are not. It is likely the case that for the low-prevalence topics, the addition of the 2399 random documents greatly skewed how much relative effort was required in those cases. This comparison has highlighted that reviewing 100,000 documents for every topic will produce over 80% recall on average, but spending the same amount of relative effort (e.g., reviewing 2 non-relevant documents for every relevant) across topics can highlight substantial differences in per-topic performance.

a	b	Recall	Precision	F1	$loss_r$	$loss_{re}$	Coverage
0	100	0.01749 (0.06313)	0.1879 (0.2925)	0.02047 (0.06395)	-0.02626 (0.08892)	-0.01313 (0.04446)	0.1933 (0.2956)
0	1000	0.1183 (0.1736)	0.319 (0.2237)	0.1322 (0.1473)	-0.1613 (0.2071)	-0.08064 (0.1035)	0.3437 (0.2408)
1	0	0.3489 (0.139)	0.3489 (0.139)	0.3489 (0.139)	-0.3551 (0.1881)	-0.1775 (0.09404)	0.3489 (0.139)
1	100	0.367 (0.1375)	0.3458 (0.1282)	0.3533 (0.1325)	-0.3678 (0.192)	-0.1839 (0.09601)	0.367 (0.1375)
1	1000	0.3851 (0.1623)	0.2991 (0.1163)	0.3285 (0.1235)	-0.3564 (0.202)	-0.1782 (0.101)	0.3851 (0.1623)
2	0	0.3831 (0.1794)	0.1915 (0.0897)	0.2554 (0.1196)	-0.3056 (0.2278)	-0.1528 (0.1139)	0.3831 (0.1794)
2	100	0.3873 (0.1789)	0.1864 (0.08504)	0.2508 (0.1151)	-0.305 (0.228)	-0.1525 (0.114)	0.3873 (0.1789)
2	1000	0.383 (0.1863)	0.1624 (0.07253)	0.2258 (0.1012)	-0.2862 (0.2217)	-0.1431 (0.1108)	0.383 (0.1863)
4	0	0.3335 (0.2164)	0.08339 (0.05409)	0.1334 (0.08654)	-0.2252 (0.2347)	-0.1126 (0.1174)	0.3335 (0.2164)
4	100	0.3348 (0.2176)	0.08139 (0.05236)	0.1307 (0.08424)	-0.2228 (0.2327)	-0.1114 (0.1164)	0.3348 (0.2176)
4	1000	0.3251 (0.2187)	0.07263 (0.04565)	0.1183 (0.07485)	-0.2081 (0.2245)	-0.1041 (0.1123)	0.3251 (0.2187)

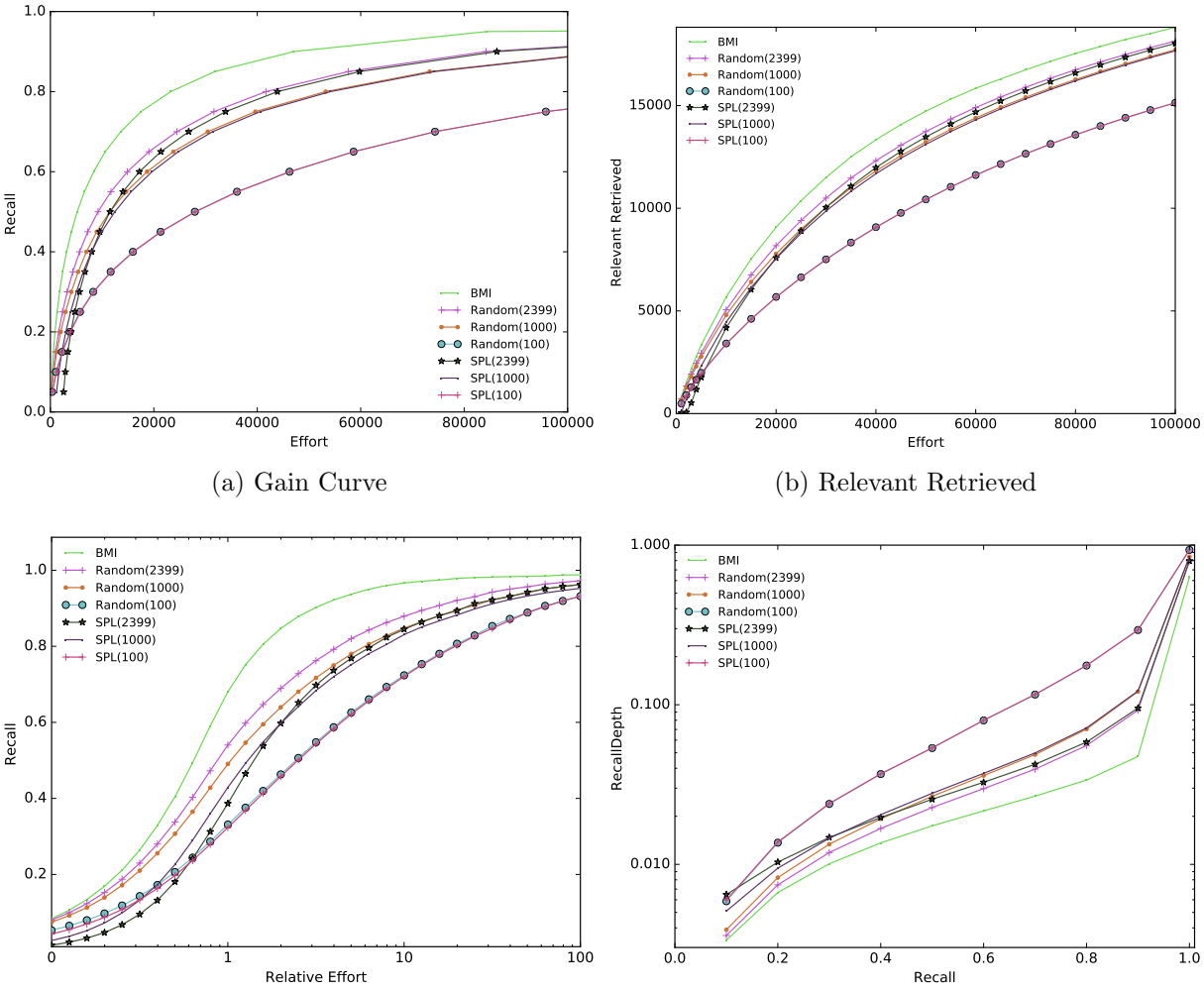
Table 5.2: Random(100) difference from BMI results for $aR + b$ evaluation, mean, and standard deviation from 101 bootstrap samples.

a	b	Recall	Precision	F1	$loss_r$	$loss_{re}$	Coverage
0	100	0.009615 (0.07263)	-0.02509 (0.2258)	0.008215 (0.0736)	-0.01322 (0.09908)	-0.006612 (0.04954)	-0.01967 (0.2331)
0	1000	0.08233 (0.1515)	0.1346 (0.1788)	0.08255 (0.1235)	-0.101 (0.1714)	-0.05049 (0.08569)	0.1556 (0.2089)
1	0	0.1891 (0.1281)	0.1891 (0.1281)	0.1891 (0.1281)	-0.1794 (0.1678)	-0.08971 (0.08389)	0.1891 (0.1281)
1	100	0.2058 (0.1332)	0.1885 (0.1114)	0.1943 (0.1192)	-0.1916 (0.1724)	-0.0958 (0.08622)	0.2058 (0.1332)
1	1000	0.2139 (0.1525)	0.1525 (0.08139)	0.1723 (0.09748)	-0.1742 (0.1684)	-0.08712 (0.08419)	0.2139 (0.1525)
2	0	0.207 (0.1645)	0.1035 (0.08224)	0.138 (0.1096)	-0.1487 (0.1895)	-0.07437 (0.09475)	0.207 (0.1645)
2	100	0.2107 (0.1676)	0.09936 (0.07568)	0.1342 (0.1034)	-0.148 (0.1881)	-0.07402 (0.09405)	0.2107 (0.1676)
2	1000	0.2038 (0.1696)	0.08131 (0.05655)	0.1147 (0.08214)	-0.1304 (0.1756)	-0.06521 (0.08779)	0.2038 (0.1696)
4	0	0.1726 (0.1874)	0.04316 (0.04684)	0.06905 (0.07495)	-0.1031 (0.1839)	-0.05157 (0.09194)	0.1726 (0.1874)
4	100	0.1736 (0.1895)	0.04144 (0.04436)	0.06667 (0.07159)	-0.1007 (0.1806)	-0.05034 (0.09028)	0.1736 (0.1895)
4	1000	0.1644 (0.1843)	0.03496 (0.03556)	0.05736 (0.05902)	-0.0886 (0.1665)	-0.0443 (0.08325)	0.1644 (0.1843)

Table 5.3: Random(1000) difference from BMI results for $aR + b$ evaluation, mean, and standard deviation from 101 bootstrap samples.

a	b	Recall	Precision	F1	$loss_r$	$loss_{re}$	Coverage
0	100	0.005904 (0.07341)	-0.07148 (0.2106)	0.002132 (0.07372)	-0.006903 (0.09791)	-0.003452 (0.04895)	-0.066 (0.2153)
0	1000	0.06471 (0.1317)	0.08826 (0.1382)	0.06095 (0.09672)	-0.07417 (0.1348)	-0.03709 (0.0674)	0.107 (0.1734)
1	0	0.14 (0.1088)	0.14 (0.1088)	0.14 (0.1088)	-0.1283 (0.1376)	-0.06414 (0.06878)	0.14 (0.1088)
1	100	0.1561 (0.1186)	0.1409 (0.09357)	0.1457 (0.1019)	-0.1408 (0.1463)	-0.07042 (0.07313)	0.1561(0.1186)
1	1000	0.1618 (0.1311)	0.1122 (0.06482)	0.1277 (0.07859)	-0.1252 (0.1345)	-0.06259 (0.06723)	0.162 (0.1311)
2	0	0.158 (0.1411)	0.07901 (0.07055)	0.1053 (0.09407)	-0.1095 (0.1586)	-0.05473 (0.07932)	0.158 (0.1411)
2	100	0.1623 (0.1456)	0.07563 (0.06345)	0.1023 (0.08723)	-0.1095 (0.1563)	-0.05475 (0.07817)	0.1623(0.1456)
2	1000	0.1546 (0.1423)	0.06038 (0.04516)	0.08551 (0.06605)	-0.09306 (0.1381)	-0.04653 (0.06903)	0.155 (0.1423)
4	0	0.1303 (0.1538)	0.03259 (0.03844)	0.05214 (0.06151)	-0.07515 (0.146)	-0.03758 (0.07302)	0.130 (0.1538)
4	100	0.1311 (0.1568)	0.03097 (0.03581)	0.04988 (0.05792)	-0.07285 (0.1422)	-0.03642 (0.07108)	0.131 (0.1568)
4	1000	0.1216 (0.1491)	0.02539 (0.02755)	0.04176 (0.04597)	-0.06141 (0.1274)	-0.0307 (0.06371)	0.122 (0.1491)

Table 5.4: Random(2399) difference from BMI results for $aR + b$ evaluation, mean, and standard deviation from 101 bootstrap samples.



(a) Gain Curve

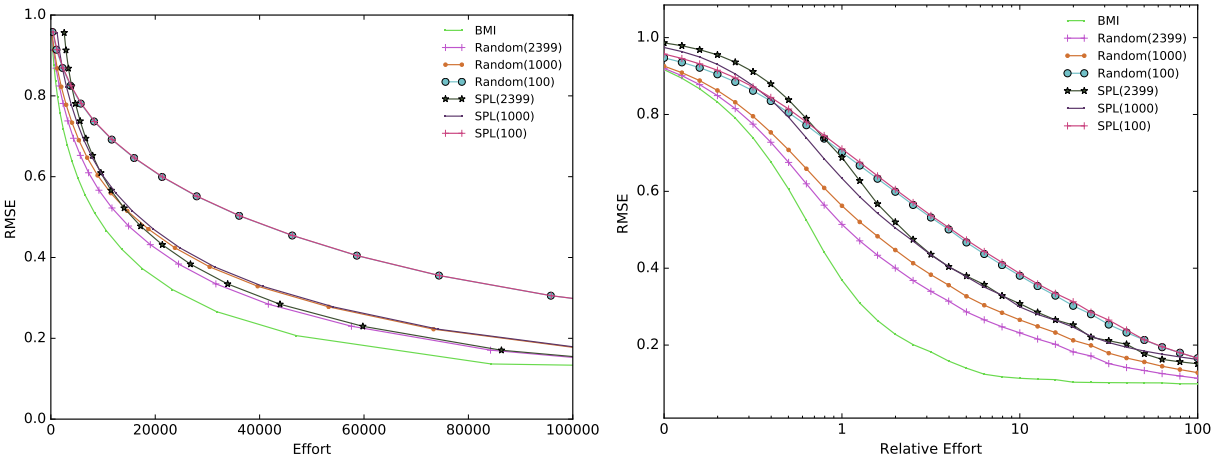
(b) Relevant Retrieved

(c) Relative Effort

(d) Recall Depth

Figure 5.2: Average curves for the RCV1 collection comparing gain, relative effort, and relevant retrieved curves.

Tables 5.2 through 5.5 compare BMI to optimistic random sampling (i.e., for each table the difference of BMI and random sampling was computed), with all three values of K, for all of the summary measures. SPL runs are omitted due to the similarity of results to optimistic random sampling and to minimize the amount of redundant information. These values were generated through 100 bootstrap samples of the RCV1 collection and by then computing measures on the original topic set as well as on each of the bootstrap samples.



(a) Gain Curve

(b) Relative Effort

Figure 5.3: RMSE for gain and relative gain curves on the RCV1 collection.

Run	AUC	MAP	Hyp-F1	Hyp-$loss_{re}$
Random(100)	0.08682 (0.06274)	0.4101 (0.157)	0.3482 (0.1288)	-0.005342 (0.0131)
Random(1000)	0.02718 (0.02792)	0.2152 (0.1525)	0.1913 (0.1223)	-0.00181 (0.003933)
Random(2399)	0.01687 (0.01943)	0.1566 (0.1262)	0.1455 (0.1027)	-0.00112 (0.002856)

Table 5.5: Mean and standard deviation of difference from BMI for ranked evaluation after 101 bootstrap samples.

Note that due to the difference calculation, BMI "wins" if the mean is greater than 0 for any non-loss metric or less than 0 for any loss metric.

The general result of the tables is that BMI is still superior to random sampling but the standard deviations are large enough to cast doubt as to whether or not such differences are statistically significant. Such a result might seem surprising but when we examine the per-topic gain curves for RCV1 (two exemplars are presented in Figure 5.1), we see highly variable performance among all systems, a lack of consistency that would contribute to the variability we see in these tables. But these per-topic curves, which are omitted for brevity, also show that BMI is at least as good as, and often much better than, random sampling, which is not apparently being captured by these measures as well as we might wish.

An astute observer may notice that in Tables 5.3 and 5.4, BMI achieves a higher recall but lower precision than optimistic random sampling when computed for a depth of

100 documents. Such an occurrence should rightfully be cause for alarm; however, after examining the per-topic results for BMI and Random(2399), we found that this result appears to be related to the use of the arithmetic mean. That is, for 28 out of the 103 topics, BMI has a higher recall, while for 67 topics it has lower recall, but when BMI "wins" it does so by a larger margin than does Random(2399). Winning on recall can potentially have a much larger impact than winning on precision, as several topics have fewer than 100 documents. BMI, alone, appears to have a wider variation in per-topic performance as well, with a recall standard deviation (from the bootstrap sample) of 0.098 while Random(2399)'s recall standard deviation is 0.076. A similar difference is found for precision. Accordingly, this particular quirk of evaluation appears to have occurred as a result of per-topic variance and the use of the arithmetic mean, which can be unduly influenced by outliers.

Based upon these observations, we posit that the somewhat unintuitive RCV1 results may be resulting from unique properties of the collection itself (and, perhaps some issues with the arithmetic mean). That is, per-topic prevalence is, in general, very high but also shows a large topic-to-topic variation, which may not make it suitable for high-recall retrieval evaluation at a variety of (arbitrary) points. Indeed, the prevalence issue may also aid random sampling, since it has a higher probability of sampling documents that are relevant (or belong to a similar topic hierarchy) and, thus, of performing more competitively in this experimental scenario than in more "real-world" corpora. This motivated our replication of this experiment in the context of the TREC-6 ad hoc test collection which has more "standard" values of per-topic prevalence.

While we could stratify the RCV1 topics to group together similar prevalence topics and average them over those groups, this approach sidesteps the issue of having a general-purpose effectiveness measure that is meant to be general purpose. Further, such stratification of topics may not be possible in all collections and may result in reduced statistical power since each prevalence group would comprise fewer topics. It may be possible to use forms of meta-analysis [50] to combine the prevalence group results to give general results.

Similarly, we have focused on the effect of per-topic variations of prevalence as an underlying cause in the unintuitive results we've seen in these experiments. There has been extensive work on the effect of topic difficulty (generally, [28, 74]), in which prevalence plays a part, on IR evaluation. Further exploring the effect of topic difficulty on high-recall retrieval evaluation is important, especially in the legal setting where some requests may be easier to meet than others (e.g., low- versus high-stakes cases), but we omitted such investigation since prevalence at least appears to account for much of the variation we've seen in this set of experiments. This is due in part to the nature of the curves we are

examining: gain curves when averaged across topics have a greater chance of reflecting differences in prevalence. Consider two hypothetical topics: one with a prevalence of 500 documents and one with a prevalence of 50,000 documents. When we examine the recall attained at 25,000 documents for both topics and effective systems, we might reasonably expect (based on the results of the previous chapter) that the first topic to be mostly retrieved while the second is, at most, half-retrieved. Accordingly, prevalence and our measurement of when to compute recall are interrelated and, thus our main focus, but we do acknowledge that undoubtedly topic difficulty also plays a role.

5.2.3 TREC-6

Based upon the interesting results of the previous section, we have replicated the same set of experiments on the TREC-6 test ad hoc collection (described in Chapters 3 and 6). The TREC-6 collection also makes use of newswire documents, but they are older, and the collection has only 50 topics with standard TREC-style topic statements. For the purposes of creating our pseudo-relevant document, we used only the topic title and no other information, following our similar experimental set-up from Chapter 6. The other point worthy of note is that the collection is not completely assessed; it is only partially assessed, using the standard TREC pooling method. Accordingly, for any document not judged by the NIST assessors, we treated such documents as not relevant, which is in line with Cranfield-style evaluation and the other experiments in this thesis.

Figure 5.4 compares BMI to random sampling with respect to all of the various curves. In contrast to what we saw in the RCV1 experiment, it is immediately apparent that BMI is vastly superior to any of the random sampling approaches, regardless of sample size. More interestingly, the curves all generally appear to agree on the performance of SPL(2399) with respect to the other random-sampling runs. This indicates that the variance in prevalence in the RCV1 topics may have had undue influence on the performance and evaluation of the runs.

The RMSE curves (Figure 5.5) for TREC-6 also illustrate the same trends as those of the other curves in Figure 5.4. Interestingly, when plotting with relative effort, SPL(2399) appears to lag behind much more than when plotting with absolute effort. In this case, such an effect is due to the low topic prevalence in the corpus, which means that those 2399 random documents delay finding relevant documents when compared to SPL(1000). Such behaviour is hidden in the gain curve, where SPL(2399) does appear to be better than SPL(1000), at all times which is likely not the case initially. These results appear to indicate that using absolute effort as a basis for evaluation may result in non-trivial behaviour being hidden.

a	b	Recall	Precision	F1	$loss_r$	$loss_{re}$	Coverage
0	100	0.467 (0.042)	0.265 (0.037)	0.265 (0.031)	-0.617 (0.047)	-0.309 (0.024)	0.527 (0.042)
0	1000	0.875 (0.025)	0.078 (0.011)	0.131 (0.017)	-0.952 (0.021)	-0.476 (0.011)	0.875 (0.025)
1	0	0.321 (0.037)	0.321 (0.037)	0.321 (0.037)	-0.485 (0.051)	-0.243 (0.025)	0.321 (0.037)
1	100	0.599 (0.039)	0.210 (0.023)	0.284 (0.028)	-0.759 (0.042)	-0.379 (0.021)	0.599 (0.039)
1	1000	0.888 (0.024)	0.068 (0.009)	0.119 (0.015)	-0.956 (0.021)	-0.478 (0.010)	0.888 (0.024)
2	0	0.471 (0.045)	0.236 (0.023)	0.314 (0.030)	-0.635 (0.053)	-0.317 (0.027)	0.471 (0.045)
2	100	0.681 (0.040)	0.166 (0.016)	0.254 (0.022)	-0.818 (0.041)	-0.409 (0.021)	0.681 (0.040)
2	1000	0.898 (0.023)	0.061 (0.008)	0.108 (0.013)	-0.959 (0.021)	-0.480 (0.010)	0.898 (0.023)
4	0	0.621 (0.045)	0.155 (0.011)	0.248 (0.018)	-0.760 (0.046)	-0.380 (0.023)	0.621 (0.045)
4	100	0.757 (0.038)	0.118 (0.009)	0.200 (0.015)	-0.863 (0.038)	-0.432 (0.019)	0.757 (0.038)
4	1000	0.911 (0.022)	0.050 (0.006)	0.093 (0.010)	-0.963 (0.020)	-0.482 (0.010)	0.911 (0.022)

Table 5.6: Random(100) difference from BMI results for $aR + b$ evaluation, mean, and standard deviation from 101 bootstrap samples on the TREC-6 ad hoc collection.

a	b	Recall	Precision	F1	$loss_r$	$loss_{re}$	Coverage
0	100	0.456 (0.044)	0.247 (0.035)	0.252 (0.030)	-0.599 (0.050)	-0.299 (0.025)	0.508 (0.043)
0	1000	0.845 (0.031)	0.072 (0.011)	0.122 (0.016)	-0.909 (0.030)	-0.455 (0.015)	0.845 (0.031)
1	0	0.307 (0.036)	0.307 (0.036)	0.307 (0.036)	-0.462 (0.050)	-0.231 (0.025)	0.307 (0.036)
1	100	0.581 (0.041)	0.199 (0.021)	0.271 (0.026)	-0.730 (0.046)	-0.365 (0.023)	0.581 (0.041)
1	1000	0.857 (0.029)	0.063 (0.009)	0.111 (0.014)	-0.912 (0.029)	-0.456 (0.015)	0.857 (0.029)
2	0	0.452 (0.045)	0.226 (0.022)	0.301 (0.030)	-0.603 (0.054)	-0.302 (0.027)	0.452 (0.045)
2	100	0.660 (0.042)	0.158 (0.015)	0.242 (0.021)	-0.784 (0.046)	-0.392 (0.023)	0.660 (0.042)
2	1000	0.866 (0.029)	0.056 (0.007)	0.101 (0.012)	-0.913 (0.030)	-0.457 (0.015)	0.866 (0.029)
4	0	0.595 (0.045)	0.149 (0.011)	0.238 (0.018)	-0.722 (0.048)	-0.361 (0.024)	0.595 (0.045)
4	100	0.730 (0.041)	0.112 (0.009)	0.191 (0.015)	-0.824 (0.044)	-0.412 (0.022)	0.730 (0.041)
4	1000	0.877 (0.028)	0.047 (0.005)	0.086 (0.009)	-0.915 (0.030)	-0.458 (0.015)	0.877 (0.028)

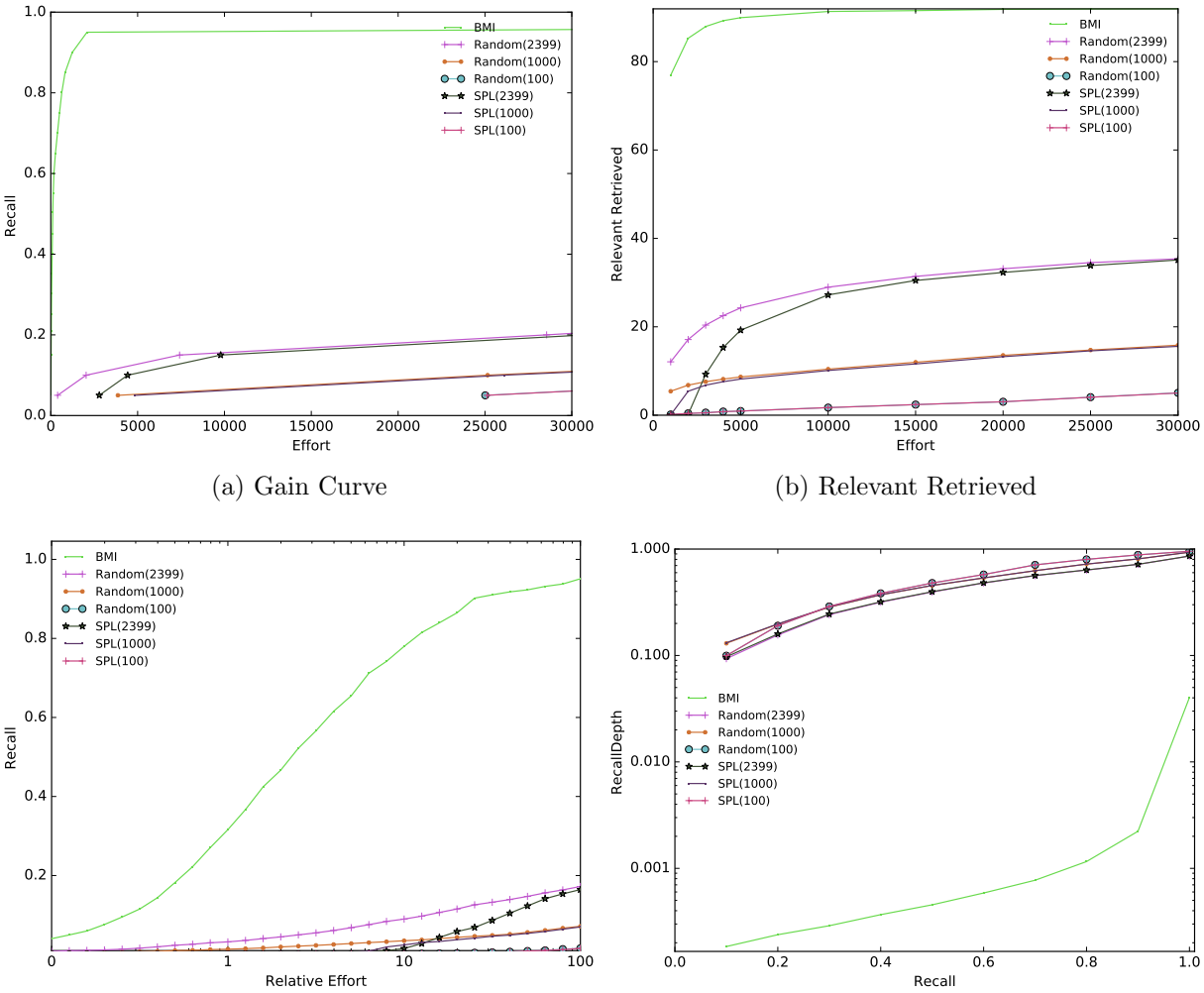
Table 5.7: Random(1000) difference from BMI results for $aR + b$ evaluation, mean, and standard deviation from 101 bootstrap samples on the TREC-6 ad hoc collection.

a	b	Recall	Precision	F1	$loss_r$	$loss_{re}$	Coverage
0	100	0.441 (0.047)	0.230 (0.035)	0.238 (0.031)	-0.572 (0.056)	-0.286 (0.028)	0.488 (0.047)
0	1000	0.803 (0.038)	0.066 (0.009)	0.111 (0.014)	-0.845 (0.041)	-0.422 (0.020)	0.803 (0.038)
1	0	0.289 (0.038)	0.289 (0.038)	0.289 (0.038)	-0.430 (0.054)	-0.215 (0.027)	0.289 (0.038)
1	100	0.559 (0.045)	0.187 (0.022)	0.255 (0.027)	-0.690 (0.054)	-0.345 (0.027)	0.559 (0.045)
1	1000	0.810 (0.037)	0.057 (0.007)	0.100 (0.012)	-0.841 (0.042)	-0.421 (0.021)	0.810 (0.037)
2	0	0.427 (0.047)	0.214 (0.024)	0.285 (0.031)	-0.560 (0.059)	-0.280 (0.030)	0.427 (0.047)
2	100	0.631 (0.047)	0.148 (0.015)	0.228 (0.022)	-0.736 (0.054)	-0.368 (0.027)	0.631 (0.047)
2	1000	0.816 (0.037)	0.051 (0.006)	0.091 (0.011)	-0.839 (0.042)	-0.420 (0.021)	0.816 (0.037)
4	0	0.561 (0.048)	0.140 (0.012)	0.224 (0.019)	-0.666 (0.056)	-0.333 (0.028)	0.561 (0.048)
4	100	0.692 (0.047)	0.104 (0.009)	0.178 (0.015)	-0.763 (0.054)	-0.381 (0.027)	0.692 (0.047)
4	1000	0.821 (0.037)	0.042 (0.005)	0.077 (0.008)	-0.835 (0.044)	-0.417 (0.022)	0.821 (0.037)

Table 5.8: Random(2399) difference from BMI results for $aR + b$ evaluation, mean, and standard deviation from 101 bootstrap samples on the TREC-6 ad hoc collection.

Run	AUC	MAP	Hyp-F1	Hyp-$loss_{re}$
Random(100)	0.495 (0.009)	0.302 (0.032)	0.404 (0.031)	-0.201 (0.007)
Random(1000)	0.466 (0.023)	0.291 (0.030)	0.389 (0.031)	-0.192 (0.011)
Random(2399)	0.411 (0.029)	0.278 (0.032)	0.367 (0.034)	-0.172 (0.014)

Table 5.9: Mean and standard deviation of difference from BMI for ranked evaluation after 101 bootstrap samples.



(a) Gain Curve

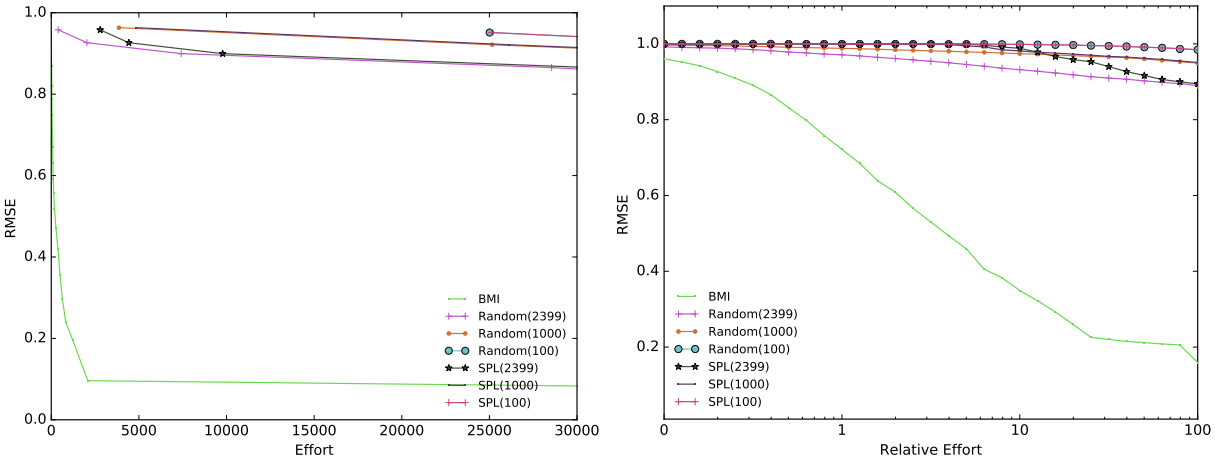
(b) Relevant Retrieved

(c) Relative Effort

(d) Recall Depth

Figure 5.4: Average curves for the TREC-6 ad hoc collection comparing gain, relative effort, and relevant retrieved curves.

Tables 5.6 through 5.9 depict the results of computing the bootstrap difference of BMI and each optimistic random sampling run. SPL results are again omitted due to similarity in performance and for brevity. We remind readers that BMI "won" for a particular measure if the mean was greater than 0 for non-loss metrics or less than 0 for any loss metric.



(a) Gain Curve

(b) Relative Effort

Figure 5.5: RMSE for gain and relative gain curves on the TREC-6 ad hoc collection.

Similar to the results of the previous section, in this experiment BMI appears to be superior according to our tabular results. Unlike in the RCV1 experiment, BMI appears to be significantly better than the random sampling strategies, regardless of sample size, due to the magnitude of the average differences and the relatively small standard deviations. In many respects, this result is not surprising, nor should it be given previous studies, but it doesn't necessarily help to answer our original line of inquiry: "Are some metrics better than others when it comes to distinguishing systems?" In the case of RCV1, no summary measure appeared capable of consistently distinguishing BMI from random sampling, but for TREC-6 every measure appeared to do so. It is likely the case that topical prevalence was a large factor in the summary measure results, which indicates that such measures may be far too dependent on prevalence to provide consistent measurement across corpora.

On the other hand, we have seen that curve-based metrics clearly show BMI to be superior to random sampling, sometimes by a large margin, and to be resilient (to some degree) to variations in prevalence across topics. However, there may be some peculiarities in these curve-based metrics which still needs to be accounted for. Overall, these results indicate that a good high-recall retrieval measure needs to take into account the totality of system performance rather than a single summary statistic. Due to these results and the relatively few topics in each test collection, we have restricted ourselves to the curve-based metrics in our empirical validation described in the following section.

5.3 Empirical Validation with Real Systems

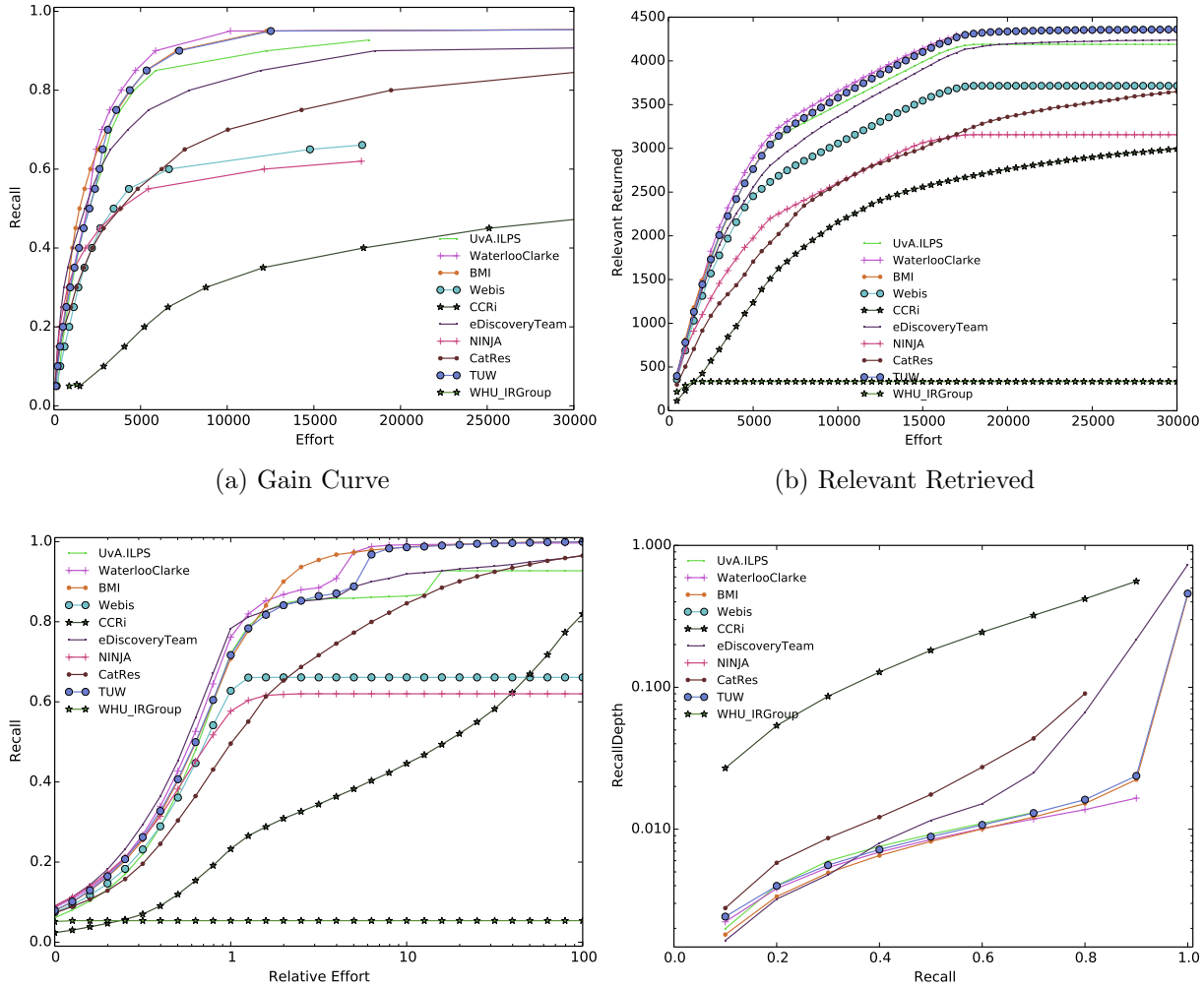
This section extends our exploration of high-recall retrieval evaluation metrics to more real-world scenarios with more realistic systems on completely assessed corpora. In particular, we examine the differences among the systems submitted to the Total Recall track and the associated test collections. These collections and systems were previously described in Chapter 4, so we have omitted further description here.

Based upon the results of our baseline analysis (Section 5.2), we focused solely on the various curve-based metrics. This is also partly motivated by the fact that the 2015 Total Recall collections had relatively few topics, which means that bootstrap sampling on such collections would be potentially influenced by any outliers, much more than would be the case with a larger number of topics. While it might be possible to bootstrap *across* collections, such a bootstrapping process has not been investigated to our knowledge, it would be worthy of its own investigation.

5.3.1 Relative versus Absolute Effort

Figures 5.6, 5.7, and 5.8 plot the four curves for each group's best run for each of the `At-Home` collections. The `Sandbox` collections will be discussed in a subsequent subsection as their behaviour motivates a different discussion.

At first glance, the gain and relevant retrieved curves tend to look similar, but there are some notable differences. In particular, if we examine the performance of the Webis and NINJA runs (and also catres for `athome1`), we can very clearly see that returning more relevant documents on average does not necessarily yield a higher average recall. For example, in Figure 5.7 Webis returns more relevant documents but CCRi is comparable to Webis in terms of the gain curve. A similar outcome can be seen when comparing Webis and catres runs. This might seem counterintuitive; how could returning more relevant documents yield lower recall? But this can occur when runs return many relevant documents on some topics topics but fewer on others, which can skew the average. This is particularly true when those successful topics are high prevalence and the others are low prevalence, which may result in equivalent recall being attained. By missing documents on low-prevalence topics, such systems can achieve lower recall on average compared to systems that may not do as well on high-prevalence topics but are more consistent overall. This results from the underlying calculation being performed; higher-prevalence topics mean that an individual document adds less recall in total than in low-prevalence topics, at least in the case of binary relevance.



(a) Gain Curve

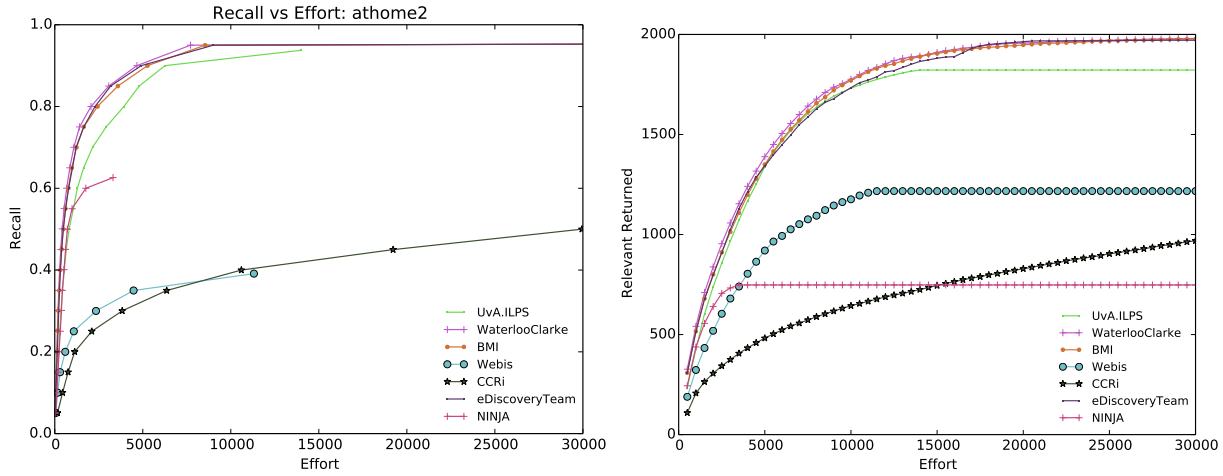
(b) Relevant Retrieved

(c) Relative Effort

(d) Recall Depth

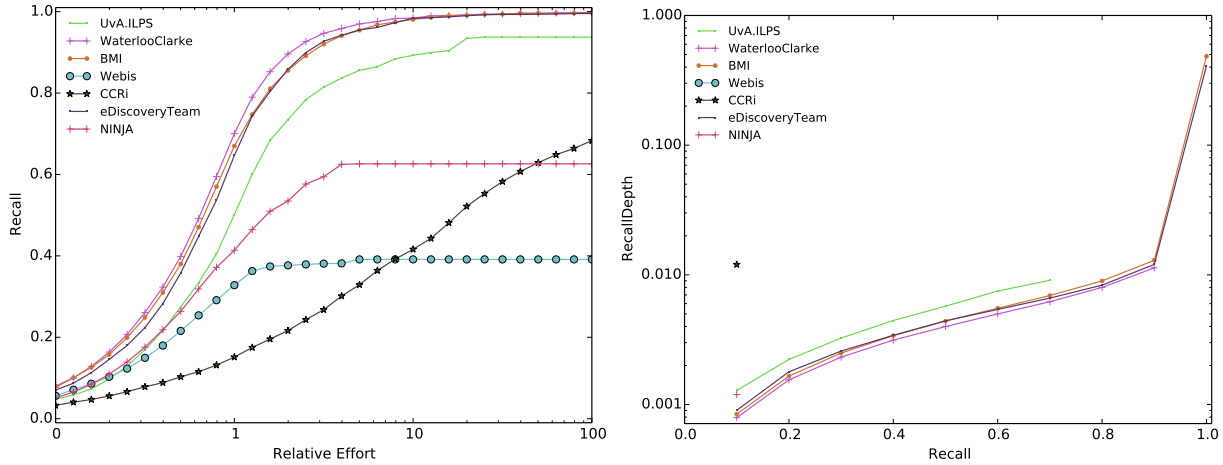
Figure 5.6: Average curves for the `athome1` collection comparing gain, relative effort, and relevant retrieved curves. Only the best run for each group is presented for readability.

One of the most interesting results is that relative gain curves appear to accommodate these differences of opinion. For example, the catres run starts out worse than either of NINJA or Webis, as in the relevant retrieved curve, but comes up and overtakes both systems as effort increases, much as in the gain curve. This may indicate that measuring system performance at comparable points is taking into account these competing factors: retrieving lots of relevant material, but doing so consistently across topics. This is similar



(a) Gain Curve

(b) Relevant Retrieved



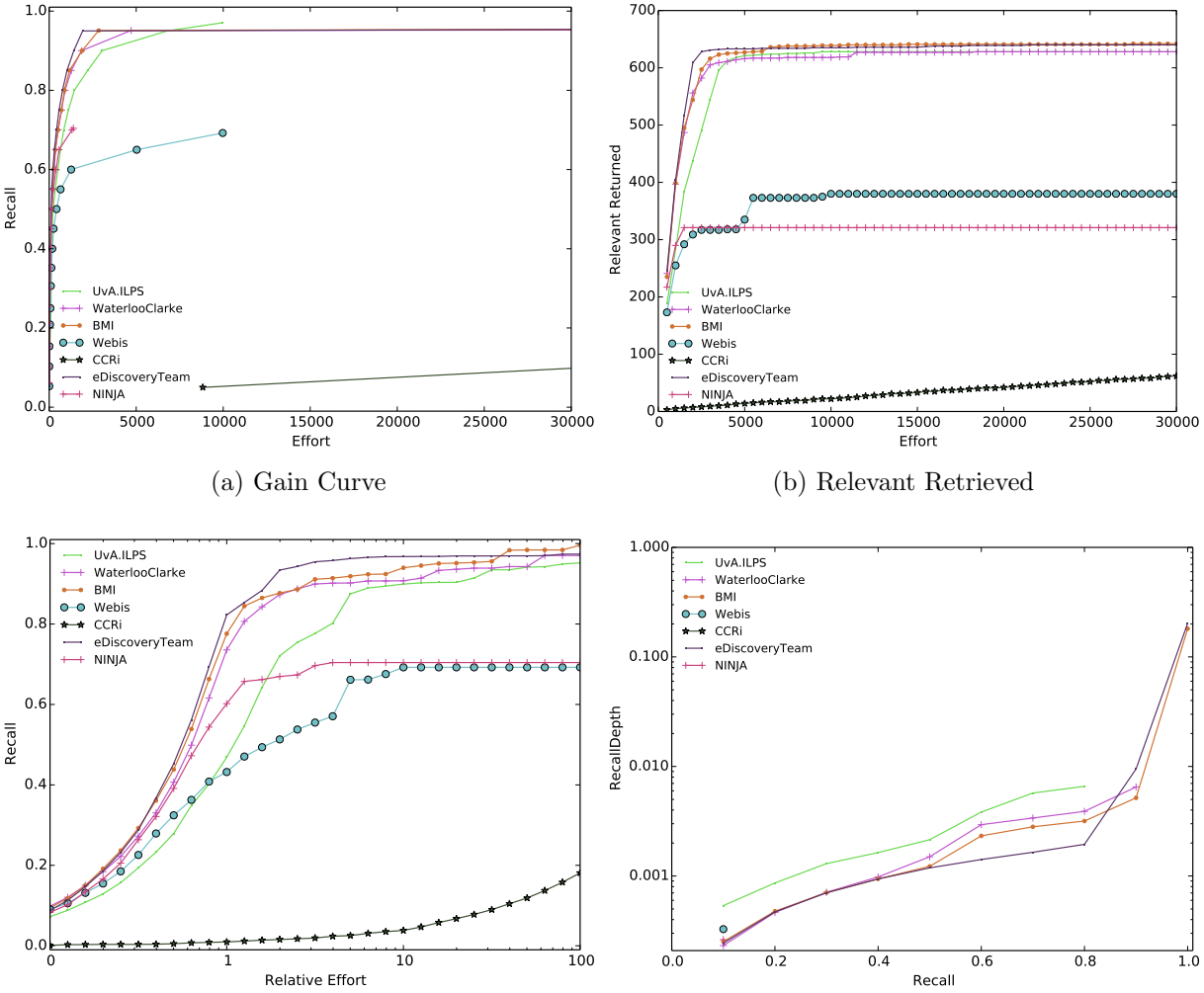
(c) Relative Effort

(d) Recall Depth

Figure 5.7: Average curves for the `athome2` collection comparing gain, relative effort, and relevant retrieved curves. Only the best run for each group is presented for readability.

to the result we observed previously in the RCV1 experiment, where the relative gain curve was able to tease out similar behavioural differences for the SPL(2399) run.

Thus far, we have focused on the less well performing systems, but in IR evaluation we are often more concerned with distinguishing the best-performing systems. Based upon the average gain curves, we might be inclined to say that there is little meaningful difference



(a) Gain Curve

(b) Relevant Retrieved

(c) Relative Effort

(d) Recall Depth

Figure 5.8: Average curves for the `athome3` collection comparing gain, relative effort, and relevant retrieved curves. Only the best run for each group is presented for readability.

between such systems. Largely speaking, the relevant retrieved curve would agree with this assessment. The relative gain curves indicate that there are differences occurring, which can be substantial. The recall depth curves tend to indicate that there is some difference at high-recall but it is not always substantial. Regardless, the performance of these systems is much more competitive than those in the baseline experiments, which was the intended

benefit of this validation study.

For athome1, the baseline system (BMI) appears to dramatically pull ahead from the other systems just after a relative effort of 1. However, the UvA.ILPS and eDiscoveryTeam systems appear to swap roles with the gain curve for athome1 (or even the relevant retrieved curve). Recall depth generally appears to agree with the relative gain curve for athome1, which may indicate that UvA.ILPS may be better than eDiscoveryTeam for this collection. As we will discuss in Section 5.3.2, these particular differences may result from differences in the consistency of systems rather than from issues with the evaluation.

The gain, relative gain, and recall depth curves all tend to agree with each other on the relative ordering of systems on athome2. Only the relevant retrieved curve appears to lump most of the systems together, which may again be an example of systems finding equivalent numbers of relevant documents on *average*. Howeveer, from the other curves it appears that this still results in different recall being achieved which is due in part to differing retrieval rates on a per-topic basis. The relative gain curve appears to be able to tease the system performance apart to a higher degree than does the gain or recall depth curve. This aspect of the relative gain curve appears to be fairly consistent across all of the experiments in this chapter. However, this may just be due to the automatic-scaling effect present in the relative gain curve rather than to any innate superiority of relative effort measurements[4].

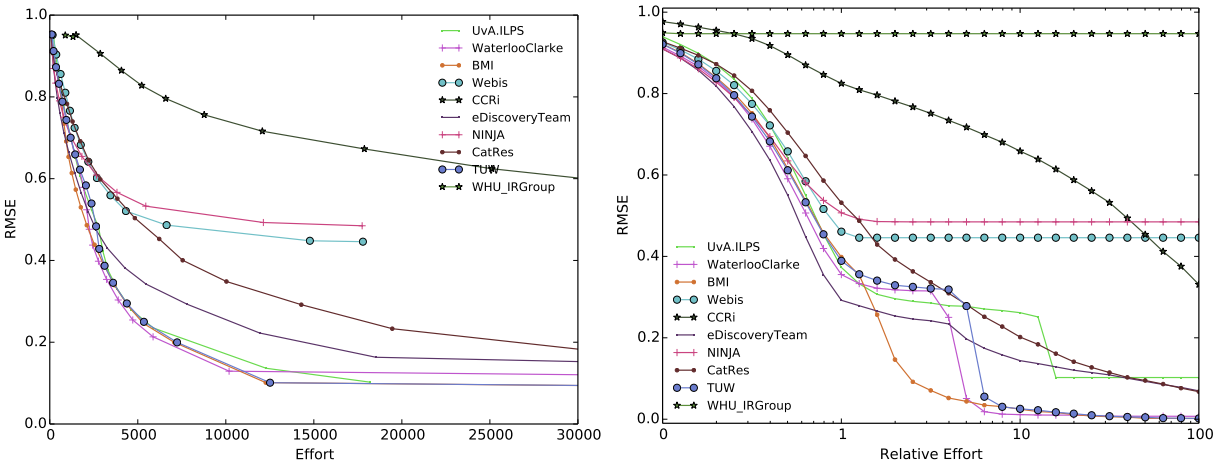
The athome3 collection has a relatively low average R of 643 documents, especially when compared to the other At-Home collections (athome1: 4,398; athome2: 2,001). In fact, 6 out of the 10 topics had less than 300 relevant documents. Accordingly, the average curves based upon absolute effort will potentially be skewed based upon those topics. This skew appears to be occurring in both the gain and the relevant retrieved curves, which provides some evidence that accommodating such issues is crucial in a good evaluation measure. Relative effort appears to make such accommodations and shows meaningful differences in system performance that were not apparent otherwise. Computing recall at comparable points (e.g., with respect to R), appears to have non-trivial impact on the apparent system behaviour. With that reasoning in mind, we might infer that this is why the recall depth and relative gain curves tend to agree more with each other than with the other curves. That is, they both attempt to measure systems at comparable points across topics (i.e., 0.5R or 50% recall). For the curves which utilize raw effort, some of these issues may stem from the fact that the scale of the plot far exceeds the prevalence and, thus, may not provide high enough resolution to distinguish systems visually. Scaling

[4]That is, had we scaled with respect to some other measure, say BMI performance, we might have seen similar nuances in behaviour.

issues will always be present for such curves if we desire comparable plots (i.e., by having similar x-axis dimensions).

We have avoided discussing the statistical significance of these results for several reasons. The first is that, as we show in Section 5.3.2, many runs are inconsistent with respect to their average recall and comparing such runs may not yield meaningful outcomes. Second, the choice of where (on a curve) to conduct statistical tests is not apparent. That is, at which value of (relative) effort should such tests be conducted? Is there more than one? If so, how do we decide which to use? Finally, conducting many post hoc significance tests would require subsequent correction for multiple hypothesis testing that would almost certainly render the results moot. Further investigation is necessary to determine best practice when it comes to significance testing with respect to the curves presented herein.

5.3.2 Recall and Consistency

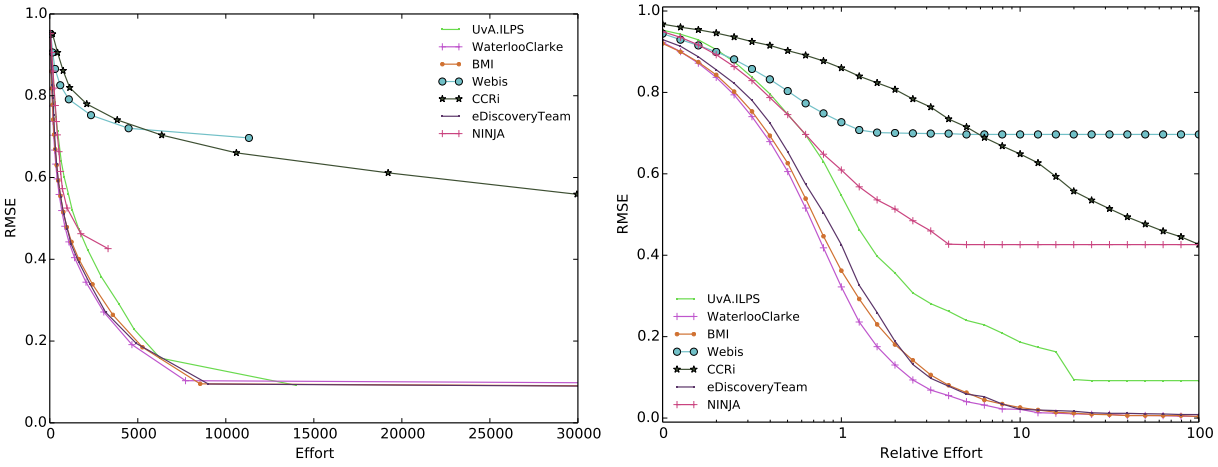


(a) Effort

(b) Relative Effort

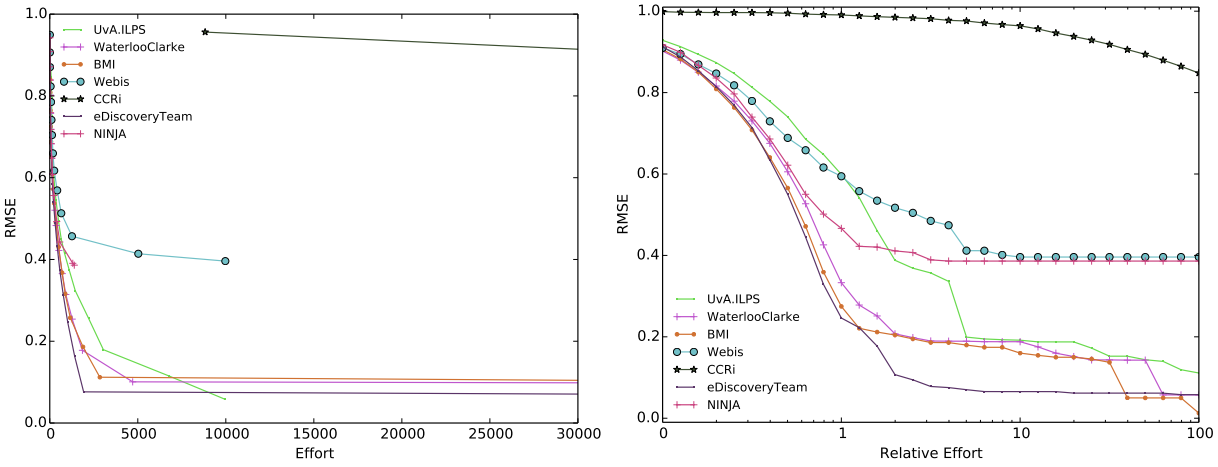
Figure 5.9: RMSE curves for the `athome1` collection.

Figures 5.9, 5.10, and 5.11 depict the RMSE curves for each of the `At-Home` collections, and correspond to Figures 5.6, 5.7, and 5.8 respectively. For `athome1`, it immediately becomes apparent that several runs were greatly inconsistent with what their average recall reported. With a relative effort of just over 1, the BMI appears to become much



(a) Effort (b) Relative Effort

Figure 5.10: RMSE curves for the `athome2` collection.



(a) Effort (b) Relative Effort

Figure 5.11: RMSE curves for the `athome3` collection.

more consistent than other systems, which roughly corresponds to its higher average recall. Interestingly, the best-performing manual system, eDiscoveryTeam, appears initially competitive but does not appear to be able to keep pace with other top-performing systems when attaining similar levels of high recall. The main benefit of RMSE is made very clear

when we compare RMSE and average recall at a relative effort of 1 for athome1. Several systems appear to be competitive with the BMI with respect to recall but are vastly inconsistent at that same level. This indicates that average recall alone may not be sufficient and that some kind of consistency measure is necessary when attempting to determine true system performance. Even among the competitive systems there appears to be a a great deal of variance that was not present in the baseline experiments.

With respect to athome3, it is still the case that several systems had precipitous drops in RMSE as relative effort increased, which is an indicator that the system performed very well on some topics but lagged behind on others prior to this drop. It is worth noting that eDiscoveryTeam appears to be more consistent for this collection than for athome1.

On the flip side, the RMSE curves for athome2 appear to generally agree with their corresponding gain curves. It appears that there is something different about athome2 from the other At-Home collections. Whether this is because of the nature of the documents (Web news and blogs versus email and message forums), the topics, or perhaps the underlying relevance assessments, is not clear.

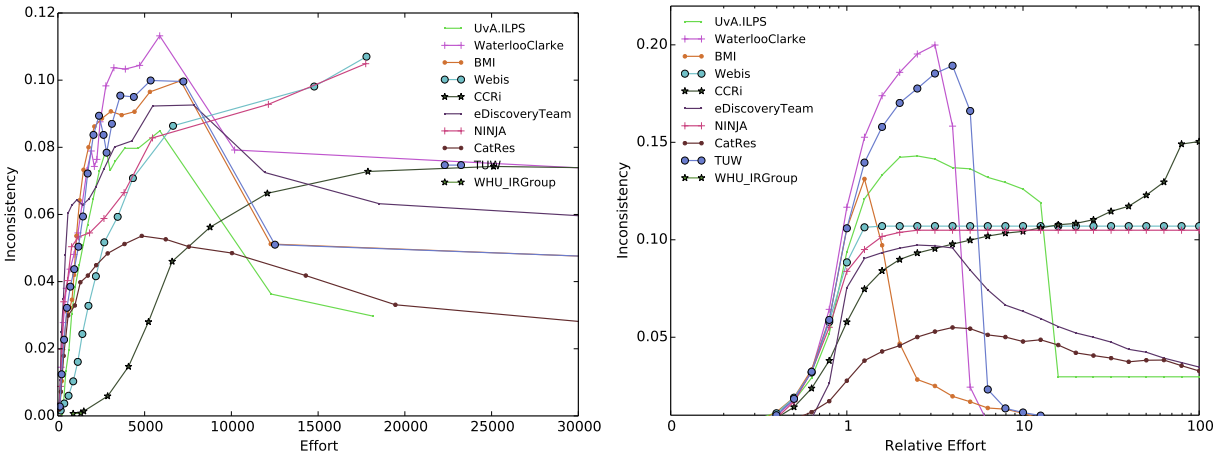
This is not to say that the systems were perfectly consistent on athome2. Indeed, the results from a perfectly consistent system would indicate that RMSE is equal to $1 - avg(\text{recall})$. Using this fact and some simple math, we can calculate the difference from perfection and reality, and referring to this difference as *consistency*. Consistency is related to standard deviation in that as one increases, the other increases, but they are numerically distinct. For example, in Table 5.1, the standard deviation for System B is 25.0 while the consistency is $|25 - 35.4| = 10.4$. Though we note that for perfectly consistent systems (e.g., System A in Table 5.1), standard deviation and consistency would both be 0.

We show system consistency for the At-Home collections in Figures 5.12, 5.13, and 5.14. These plots show that, by-and-large, athome2 does result in more consistent (i.e., closer to perfect consistency) system performance but is, by no means, perfect. Counter to our underlying hypothesis throughout this chapter, measuring systems based upon relative effort does not appear to guarantee a more consistent evaluation than measuring based upon absolute effort. However, relative effort does appear to at least help distinguish similar systems for athome2 and athome3.

When comparing the RMSE and gain curves for all the At-Home collections, it appears that an average recall exceeding 90% is necessary to achieve high consistency. While not an earth-shattering result, it does appear to indicate that high-recall retrieval systems can have more variable behaviour than may be desirable.

Based upon these curves, we would suggest that statistical tests comparing system performance should *n*ot be conducted at a relative effort of less than 1, as this generally

corresponds to the highest inconsistency. Indeed, a series of paired t-tests reveals that the BMI is not significantly different from the other best-performing runs (TUW, WaterlooClarke, eDiscoveryTeam, UvA.ILPS) for any relative effort less than 2. With respect to absolute effort, a recommendation is less clear, since the peaks appear to be corpus dependent rather than determined by a particular relative effort value. However, it does appear that the peaks tend to occur around the average number of relevant documents in the corpus (i.e., 1R).



(a) Effort

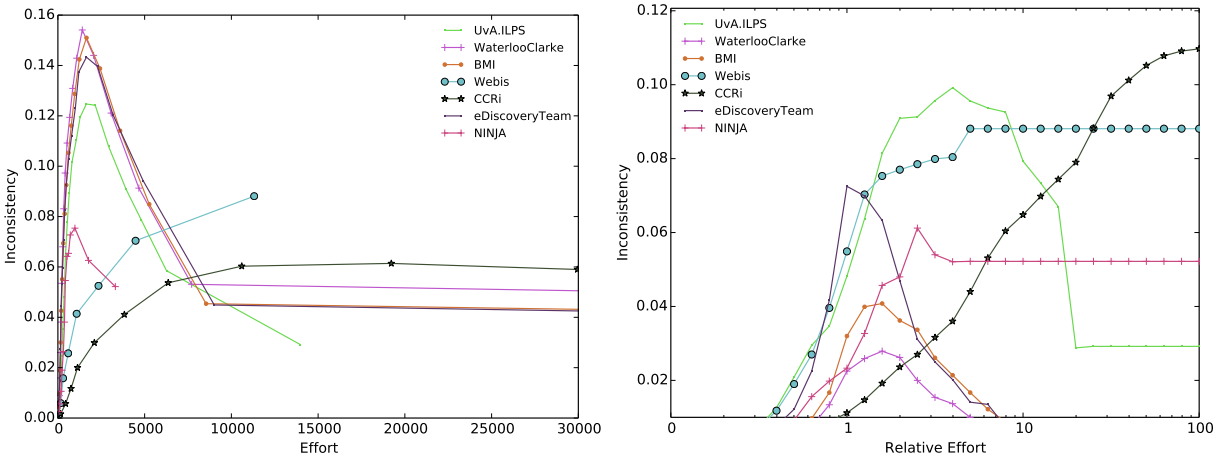
(b) Relative Effort

Figure 5.12: Consistency plots for the `athome1` collection. These plots illustrate the disparity between the root-mean-square error and the average depth/recall.

If we look at consistency of systems on the `Sandbox` collections (Figure 5.15)[5], we see that the systems are surprisingly more consistent than on the `At-Home` collections, especially in the case of `Kaine`. This behaviour may be a result of the slightly different tasks performed on these collections, namely, identifying records with a particular ICD-9 code (`MIMIC`) and information governance[6] (`Kaine`). Accordingly, there is not necessarily the issue of topical relevance at play *per se*, which may contribute to inconsistency in other collections. In addition, the relatively small size of `MIMIC` and the 4 topics of `Kaine` may have had some effect on these results. It is noteworthy that `Kaine` is as consistent as it is, given that it has two very high prevalence topics and two not so high prevalence topics.

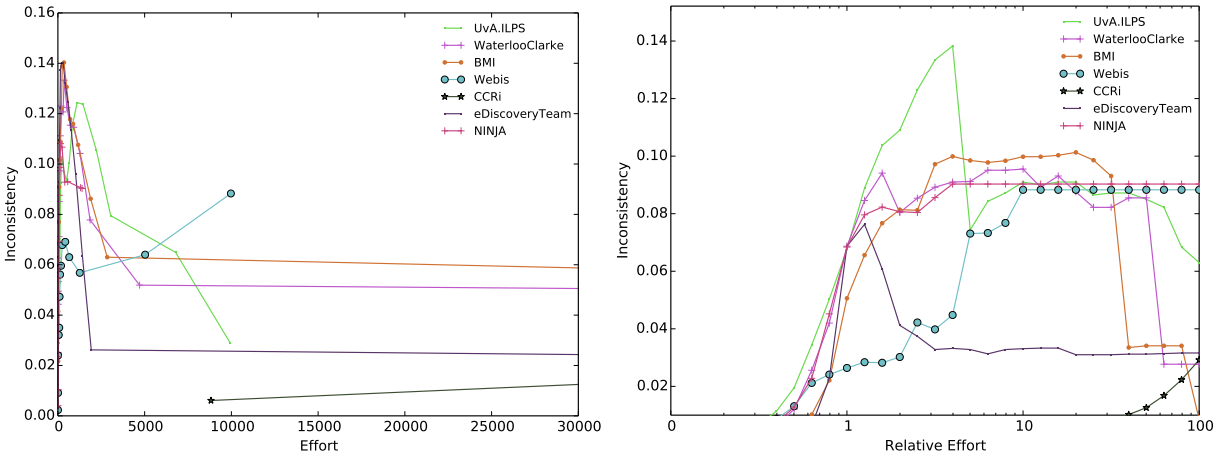
[5]For brevity, we show only the curve for relative effort.

[6]Determining archival record type.



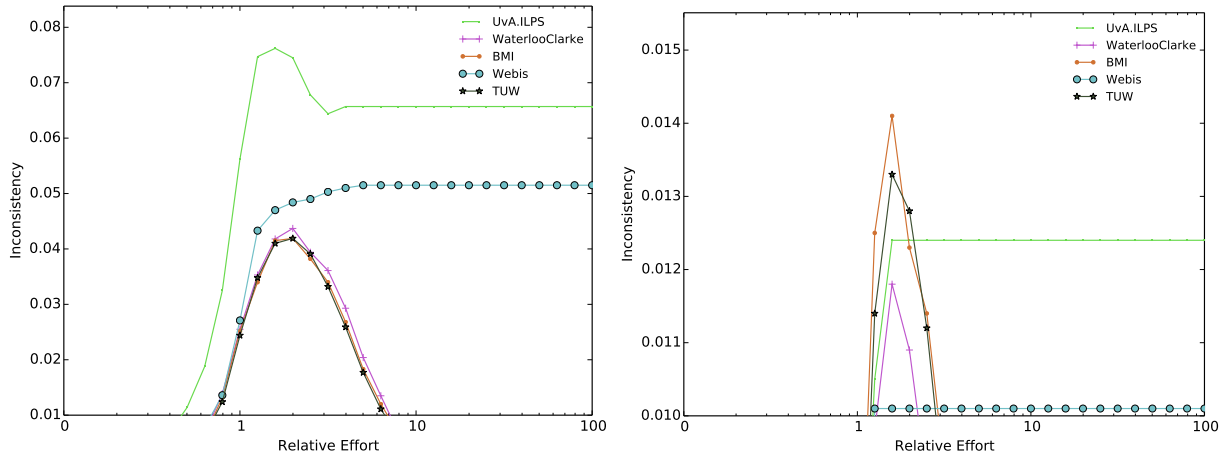
(a) Effort (b) Relative Effort

Figure 5.13: Consistency plots for the `athome2` collection. These plots illustrate the disparity between the root-mean-square error and the average recall.



(a) Effort (b) Relative Effort

Figure 5.14: Consistency plots for the `athome3` collection. These plots illustrate the disparity between the root-mean-square error and the average recall.



(a) MIMIC (b) Kaine

Figure 5.15: Plots showing the RMSE difference from the hypothetical "perfect" consistency for the Total Recall 2015 `Sandbox` collections.

5.4 Discussion

This chapter has presented a series of experiments exploring the evaluation space of high-recall retrieval tasks. The initial goal of this exploration was to determine whether any measure could consistently distinguish systems across topics and collections. It is unclear that this goal was successfully met, but we have seen a variety of behaviours that may help direct future investigations into the evaluation of high-recall retrieval systems.

One of the results that may have the most impact is that the RCV1 collection may not be a good collection for comparing high-recall retrieval systems. Indeed, our experiments showed that the state of the art (i.e., AutoTAR) was competitively challenged by random sampling on a wide variety of measures. This apparent competitiveness of random sampling persisted across a wide variety of metrics for RCV1. We believe that much of this behaviour can be attributed to the wide variation in topic prevalence in the corpus and the generally high prevalence overall (around 25,000 documents). Accordingly, we would caution against a default assumption that the RCV1 test collection is suited to distinguishing high-recall retrieval systems at arbitrary points of measurement. More directed measurement (e.g., at various stopping points) may be more valid, since these have a directed purpose.

To ensure that this result was not the by-product of a faulty experimental set-up, we replicated the experiment on the TREC-6 ad hoc test collection. As expected, AutoTAR

was substantially and significantly superior to random sampling across all evaluation measures. While this provides reassurance that our RCV1 results are accurate, it does not help us determine whether one evaluation measure is better than the others at distinguishing systems since they all worked.

The baseline experiments (RCV1 and TREC-6) did show that measuring recall in terms of relative effort may be beneficial in teasing out variations in system performance that would otherwise be masked by using absolute gain. Such a behaviour may be useful when test collections have a wide variation in topical prevalence (as was the case in RCV1). When tested on more realistic high-recall systems, similar behaviour was witnessed when comparing gain and relative gain curves. Though the "real-world" validation experiments showed that, contrary to our hypothesis, relative gain curves do not result in a more consistent measurement of average recall when compared to standard gain curves.

One of the more surprising results of the validation experiments was that even among the best-performing systems, there was a great deal of inconsistency in the recall they achieved for the same amount of (relative) effort. Such a result may indicate that significance testing based upon (relative) gain curves requires careful consideration, since systems appear to still be very inconsistent even after great effort has been expended. Though based upon the Total Recall `Sandbox` test collections, we judge that this may be dependent on the test collection and the particular task model (i.e., topical relevance versus information governance).

As was eluded to in Section 5.1, we could have computed relative effort with some non-zero fixed overhead (i.e., using the b parameter in $\frac{\text{effort}-b}{R}$). This could result in removing some system inconsistency, as we would then account for the variability that can occur in low-prevalence topics when systems attempt to overcome the cold start. However, it is unlikely to make systems substantially more consistent unless this fixed overhead were very large.

Furthermore, it may be the case that the arithmetic mean is not the ideal mean for averaging across topics with large differences in prevalence. Additional investigation is warranted into the applicability and utility of the geometric and harmonic means for these measures. However, it may also be the case that something like a percentile measurement (similar to those described in Section 4.5.2 of Chapter 4) may be more applicable in the cases of measuring the lower bound on performance.

Ultimately, we have seen that when users employ systems to achieve high-recall, there will be high variability in the recall achieved for commiserate amounts of effort, irrespective of how this effort is measured (i.e., relative or absolute). It is noteworthy that in most cases, "real-world" systems are able to achieve high-recall (˜90%) while requiring relatively

little review of the document collection on average. Also worthy of note is that our "real-world" results confirm Sormunen's observation that the last few points of recall require inordinate amounts of effort [141].

Chapter 6

Effects of High-Recall Retrieval Protocols on Assessing Behaviour

High-recall retrieval is often formulated, at least in part, as a (semi-)supervised text-classification task [42, 40, 97, 163] where an assessor judgments documents as relevant or not relevant for the purposes of training or refining a classifier. While the cited studies make use of supervised learning, it is not unreasonable for a system to not require complete supervision. For example, the TREC Spam track, as discussed in Chapter 2, offered a delayed feedback task where systems could not expect to receive a relevance assessment for all requests. Regardless of the specific task model, the training documents may be selected through any number of stragies, which can include random sampling (a passive approach) or active learning approaches, such as uncertainty sampling [95, 94] or relevance sampling (e.g., CAL [40]/AutoTAR [41] approaches). Most high-recall retrieval studies are simulations based upon existant test collections, and, thus, previously rendered judgments. Such experiments are simply following the Cranfield paradigm [153] which was employed in Chapters 3, 4, and 5.

What these studies often neglect to account for is that the presentation order and the prevalence of documents being reviewed can have an effect on the judging behaviour of the assessor. Such issues have been investigated in various forms for over 30 years [63, 82, 115, 137, 139] but not as they directly relate to high-recall retrieval and the different possible approaches.

This chapter describes a study into how training set selection can impact judging behaviour in high-recall settings. We accomplished this by conducting a paid user study, involving 36 university personnel, where they judged pre-generated training sets based

upon random, uncertainty, and relevance sampling strategies. We begin by describing this user study and its associated experimental methodology (Section 6.1). The user study made use of 9 topics from the TREC-6 ad hoc test collections and the associated assessments generated by the University of Waterloo, both of which were described in Chapter 3.

The primary result found is that the uncertainty and random sampling approaches yield a significantly higher likelihood of an assessor's marking a document as relevant than does relevance sampling (Section 6.2). On the other hand, time to render an assessment does not appear to be significantly affected by the strategy employed (Section 6.3). Such results are in line with those reported by Smucker and Jethani, who investigated the effect of prevalence on judging behaviour [137, 138]. We conclude with a discussion of the implications of this work, some limitations, and potential follow-up studies to investigate the effects in more realistic contexts—that is, when participant assessments affect the documents that appear later.

6.1 Experimental Methodology

This section provides the methodological details of the user study conducted in the Fall of 2015 at the University of Waterloo. We begin by describing the basic set-up of the study and go into more details in subsequent sections. In our design, assessors would judge the relevance of batches of 100 documents with respect to one of nine topics. For a topic, each batch contained the same 12 known documents, and 88 documents which were selected using one of random sampling, uncertainty sampling, or relevance sampling. Due to the nature of the data available (Section 6.1.1), only 9 topics were used, which resulted in $3 * 9 = 27$ batches being created in total. Each batch was then assessed by three assessors.

Due to a small error in our experimental apparatus that went undetected until the user study had been completed, for some topics the batches contained 10 or 11 of the same known documents rather than all 12. This difference does not degrade the validity of our experiment, merely its statistical power and the conclusions that can be drawn from the results.

The following is a brief summary of terms that we will use throughout the remainder of this chapter to discuss various aspects of our user study:

- **Context**: The manner in which documents, other than known documents, were selected.

- **Known documents**: 12 common documents presented for assessment across all contexts.
- **Batch**: The 100 documents presented to participants for review, which consisted of known documents and context documents.
- **NIST assessments**: Relevance assessments rendered by NIST for the TREC-6 ad hoc task [150].
- **Waterloo assessments**: Relevance assessments rendered by the University of Waterloo during participation in the TREC 6 ad hoc task [38].
- **Rel**: The relevance class of a document, rendered either by NIST, Waterloo, or some combination thereof.
- **Relevance (sampling)**: Context in which documents are selected iteratively using relevance sampling as implemented in the CAL protocol [40].
- **Uncertainty (sampling)**: Context in which documents are selected iteratively using uncertainty sampling as implemented in the SAL protocol [40].
- **Random (sampling)**: Context in which documents are selected using random sampling as implemented in the SPL protocol [40].

6.1.1 Documents and Labels

The document collection used was the TREC-6 ad hoc test collection, which has been described elsewhere in this thesis (Chapter 3) and in other relevance assessment studies [152, 121, 53]. As discussed in Chapter 3, the University of Waterloo generated a set of ternary relevance assessments ("relevant", "not relevant", and "iffy"), using interactive search and judging [38]. NIST, on the other hand, used only "relevant" and "not relevant" for their assessing. Each set of assessments has an additional implicit category, "unjudged," which corresponds to those documents that were not assessed by that group (NIST or Waterloo). This results in 3 NIST relevance classes and 4 Waterloo relevance classes, such that every document in the collection has one of each group's labels for every topic. Thus, a document belongs to one of 12 categories resulting from the Cartesian product of these two sets of labels.

Our selection of topics was such that only those which had at least 1 document belonging to each of these 12 combination classes were used. Out of the 50 official topics, 9 topics had

at least 1 document in all 12 classes. For each of these 9 topics, we selected 1 document at random from these 12 relevance classes to function as our 12 known documents. Using each of the sampling strategies, discussed below, a list of 90 documents was created for each context. The 12 known documents were then inserted at fixed positions, chosen at random, into these lists of 90 documents. Any documents appearing beyond the 100th position were discarded to maintain consistent effort across experimental conditions.

In the random sampling context, the list of 90 documents was a simple random sample of the TREC-6 document collection, excluding those documents that had been selected as known documents. This is identical to Cormack and Grossman's simple passive learning protocol [40]. For the active learning strategies, uncertainty and relevance sampling, we employed the simple active learning and continuous active learning protocols of Cormack and Grossman [40] such that the top 10 (rather than the top 1,000) documents were selected for training at each iteration. Recall that for uncertainty sampling, the top N documents are those the classifier is most uncertain about (i.e., closest to the relevance threshold); and for relevance sampling, the top N documents are those the classifier thinks are most likely to be relevant. Figure 2.1 in Chapter 2 and the associated discussion provide a more detailed description of these protocols.

The `sofia-ml`[1] package was used as the classifier of choice and was configured to minimize logistic loss (i.e., the classifier performed logistic regression). Furthermore, `sofia-ml` was provided with the trainining options given to BMI (specified in Section 4). For feature engineering, tf-idf features were generated for purely alphabetic terms that had been case normalized and Porter stemmed. The tf-idf scores were generated for each document using the formula $(1 + log(\text{term frequency in document})) * log(\frac{\text{Corpus size}}{\text{\# of documents term appears in}})$. These are the same features and scoring function as used in Chapter 3. Following the AutoTAR approach [41], each active learning strategy had an initial training set consisting of a positively labelled pseudo-document, the topic statement, and 100 negatively labelled documents, selected at random without regard to their true relevance. Due to the generally low prevalence of the corpus (as reported in Chapter 3), this should not have introduced many false negatives.

With respect to uncertainty sampling, the top 10 documents were those that had a score closest to 0 (and, therefore a likelihood of relevance close to 0.5); for relevance sampling, the top 10 were those with the greatest likelihood of relevance (i.e., highest score). Waterloo assessments were used for training such that: "relevant" and "iffy" were labelled as positive; "non-relevant" and "unjudged" were labelled negative. The choice to use the Waterloo assessments and to train "iffy" documents as positive was influenced by the results and

[1] Available from `https://code.google.com/archive/p/sofia-ml/`.

Topic	Corpus Prevalence	Context Count		
		Relevance	Uncertainty	Random
301	0.24	40	6	5
304	0.13	26	4	4
306	0.11	41	6	6
307	0.14	65	5	5
319	0.17	63	6	4
324	0.09	74	5	5
332	0.08	41	5	5
337	0.11	83	6	6
343	0.12	33	5	6

Table 6.1: Corpus prevalence and number of positive documents for each context, where Waterloo "relevant" and "iffy" assessments are considered positive. The counts for each batch include known documents.

observations of Chapter 3. Our reasoning was that using the Waterloo assessments would provide an independent set of training assessments that would not produce a classifier biased towards the official NIST assessments. In addition, we have seen that training "iffy" documents as a positive assessments can produce a better classifier.

Each list of 90 context documents was generated through 9 iterations of the particular active learning protocol or from 1 random sample in the case of passive learning. Table 6.1 depicts the prevalence of positive (Waterloo "relevant" or "iffy") documents in the corpus, as well as the number in each batch, including known documents. The number of positive documents produced by uncertainty sampling is somewhat lower than one might expect,[2] but this is likely the result of the fact that uncertainty sampling is tailored more toward Support Vector Machine (SVM) optimization than towards logistic regression. This discrepancy was not realized until after the conclusion of the user study and will be discussed further in Section 6.4.

6.1.2 Assessment Protocol

Following approval from the University of Waterloo Office of Research Ethics, 36 participants were recruited at large from the University of Waterloo through internal mailing

[2]This expectation may arise from Cormack and Grossman's results [40] showing that SAL can be competitive with CAL.

lists. These participants included undergraduate students, graduate students, and faculty, with no other demographic information being requested or retained.

Each participant was initially assigned to a context and topic at random, and was told that they would be remunerated $20 for assessing all 100 documents. They were advised that at any point they could terminate their participation and would receive $0.50 for each document judged rather than the $20. They were also notified that if they took less than 2 hours to judge all 100 documents and were sufficiently accurate in their assessments, they would receive an additional $10 and have the opportunity to assess up to two additional batches (subject to the same bonus and continuation criteria). Internally, the accuracy criteria was satisfied by achieving 25% recall and 25% precision with respect to the NIST assessments. In order to prevent particupants from trying to "game the system," they were not told the exact criteria. The criteria itself was used to limit the repeated inclusion of "bad" assessors (e.g., those who simply labelled everything as relevant or not relevant) in the study. By trying to limit the influence of "bad" assessors, we hoped to increase the statistical power of our study without requiring an overly large pool of participants.

For participants who met the criteria and wished to continue, a new context and topic were selected at random with the restriction that participants would never assess the same context or topic twice. While no participant chose to do so, had a participant wished to leave the study during a subsequent iteration, they would have received all remuneration earned up until that batch (i.e., including the additional $10) and $0.50 per document judged in that batch. Participants whose assessments did not meet the criteria were paid $20 and not invited to continue. With respect to the time component, we observed that the minimum time to finish a batch was 2.58 minutes, the maximum was 109.45 minutes, and the average was 25.8 minutes. Based on such timing information, it would appear that generally participants took their time to perform the task. While some potentially took too much time and others too little, these differences were likely to balance out overall during the analysis of the judging behaviour.

Of the 36 participants: 19 assessed 3 batches; 7 completed 2 batches; and 10 completed 1 batch. No participant failed to continue due to the time restriction and 1 participant declined to continue even though they were able to do so. In total, 3 batches representing each context and each topic were assessed. Note that the random assignment of topic and context without repetition was intended to mitigate any potential learning effects without requiring a full Graeco-Latin square. Furthermore, a full Graeco-Latin square would have required many more participants to complete all of the necessary batches, which would have greatly increased the cost of the study, both monetarily and in time spent running the study.

6.1.3 User Interface

Figure 6.1 depicts the full-screen HTML interface that was presented to participants for the assessment process. Participants, upon registering interest in the user study and after having the task explained, were directed to a Web page that provided them with a "Consent to Participate" form (Figure 6.2). This form merely collected their acknowledgment that they agreed to participate in the study. If they changed their mind, the form directed them to close the browser. Upon acceptance of the form, participants were directed to a version of the judging interface that was configured for their particular assessing task.

The assessment interface (Figure 6.1) was broken into three panes: a top left, bottom left, and right pane. The top left pane contained the full TREC topic description (i.e., the title, description, and narrative) and was always visible to the participant. Participants were allowed to familiarize themselves with the topic before starting the official assessment. Their choice to start assessing was signified by clicking a "Start" button in the right pane. Figure 6.3 provides a screenshot of the interface prior to the "Start" button's being clicked.

Upon starting the assessment process, the lower left pane contained the current document to be judged with the topic's title terms highlighted wherever they appeared in the document. This highlighting was performed because we believed that any reasonable system should have this basic minimum feature to aid in assessment. The intent was to minimize the need for participants to conduct "Ctrl-F" searches of the document, since terms were already highlighted. No additional highlighting or snippet-generation functionality was provided due to the potential for such functionality to confound the results of the study.

The right pane was used for informational and navigational purposes. Participants were told how many documents they had assessed, the time elapsed on the current document, the average time taken, and a target time per document. This timing information was provided as an explicit motivator to keep participants mindful of the time taken, but we do not believe that providing this information unduly influenced participant behaviour. The target time, in particular, corresponded to a total elapsed time of 2 hours, though it was believed that no reasonable participant should take that long[3] (though as noted

[3]It is worth noting, that there is an existing body of research [144, 54, 55] that has found thattime pressure and time constraints can impact user behaviour on search tasks. However, it was our belief that providing a time limit but providing no gauge for time elapsed would have potentially hindered our results. Indeed, in the eDiscovery domain there is a (potentially high) cost to taking too long to assess documents. Accordingly, by having such a timer, we were trying to keep the participants motivated and moving along on the task. Furthermore, the maximum envisioned time of 2 hours far exceeds the time limits in the cited studies.

above, one participant got close). Finally, the right pane contained single-click action buttons to render an assessment and proceed to the next document. The interface was intentionally designed to prohibit backtracking to change previous assessments since that would complicate the experimental set-up and could have resulted in different behaviours that were not under investigation.

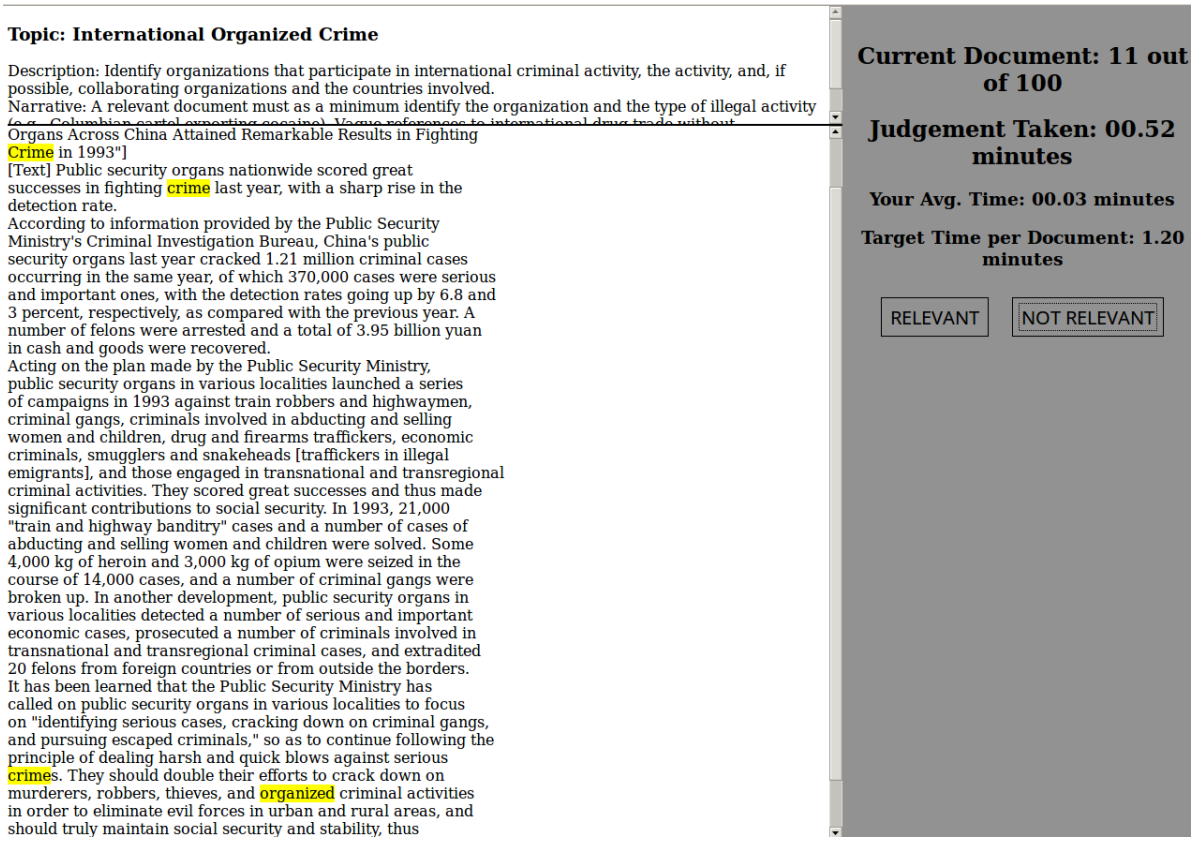


Figure 6.1: Screenshot of the document assessment interface during the assessment process.

6.1.4 Evaluation

Our interest was in the probability that an assessor would render a positive assessment for a document, $\Pr[\text{User}^+]$. The primary predictor variable in our study was the context in which the document was assessed. We measured the conditional probability, $\Pr[\text{User}^+|\text{Context}]$, so that we might test the hypothesis that $\Pr[\text{User}^+|\text{Relevance}] < \Pr[\text{User}^+|\text{Uncertainty} \vee$

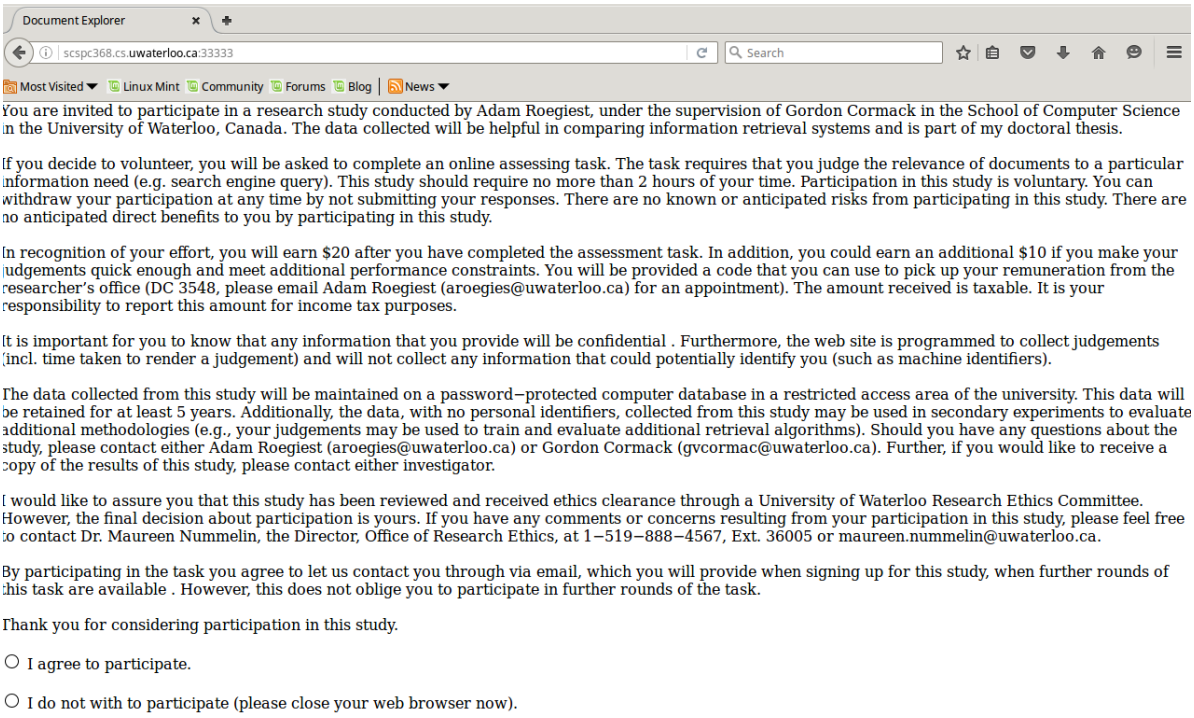


Figure 6.2: Screenshot of the document assessment interface with the "Consent to Participate" form.

Random]. If such a hypothesis was supported, we then determined whether, individually, $\Pr[\text{User}^+|\text{Relevance}] < \Pr[\text{User}^+|\text{Uncertainty}]$ and $\Pr[\text{User}^+|\text{Relevance}] < \Pr[\text{User}^+|\text{Random}]$.

A secondary predictor variable was the relevance class, Rel, of the document, which was determined by the NIST or Waterloo assessments or some combination thereof. To preserve statistical power, we restricted this class to be W-RI, and its complement, W-NU, where W-RI denotes that Waterloo judged the document as "relevant" or "iffy"; thus, W-NU denotes those documents that were unjudged or judged as "not relevant" by Waterloo.

Prior to conducting any statistical tests, we assumed that the hypothesis that $\Pr[\text{User}^+|\text{W-RI}] > \Pr[\text{User}^+|\text{W-NU}]$ was extremely unlikely to be rejected. Accordingly, we concerned our investigation with the following two hypotheses:

(1) $\Pr[\text{User}^+|\text{Relevance} \wedge \text{W-RI}] < \Pr[\text{User}^+|(\text{Uncertainty} \vee \text{Random}) \wedge \text{W-RI}]$

(2) $\Pr[\text{User}^+|\text{Relevance} \wedge \text{W-NU}] < \Pr[\text{User}^+|(\text{Uncertainty} \vee \text{Random}) \wedge \text{W-NU}]$

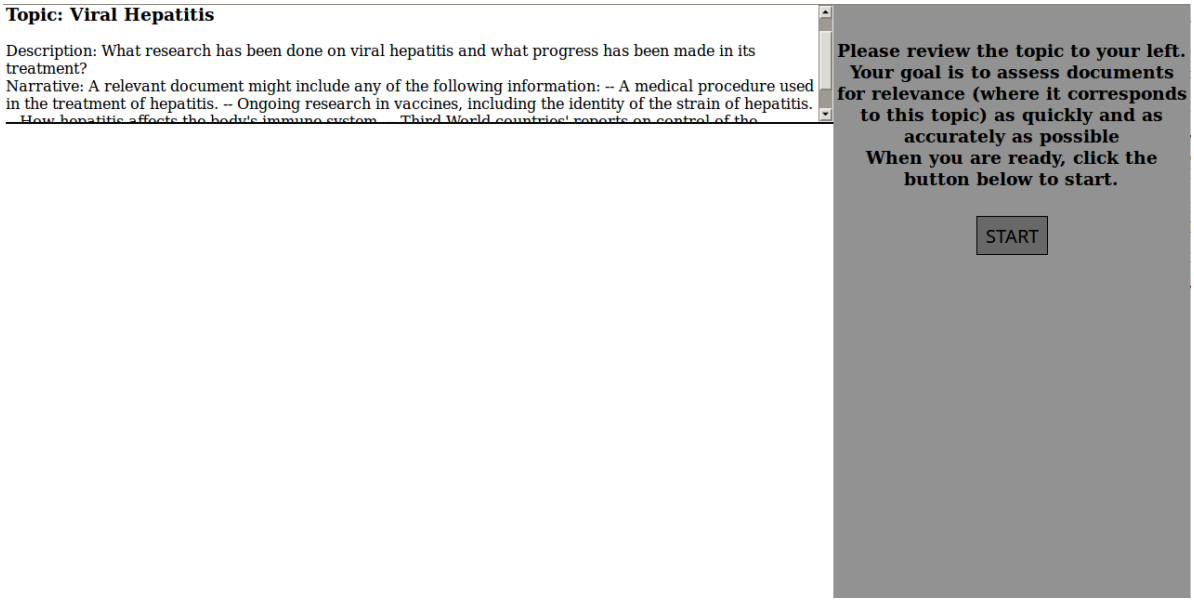


Figure 6.3: Screenshot of the document assesssment interface before the assessment process.

To approximate the conditional probabilities above, we computed the fraction of positive assessments for documents satisfying the corresponding predictor variables. To determine the significance of the hypothesized difference, we applied a paired binomial test, where possible, to equivalent batches based upon topic, relevance class, and assessor experience. Due to the experimental design and use of a fixed number of known documents, there were an equal number of batches for each context and relevance class present in the above hypotheses. For the purposes of conducting paired significance tests, it was easy enough to pair batches based upon the specific combination of topic and relevance class. If we had not invited particiants to asses more than one batch, it would have been simple to pair based upon assessor since we could simply have chosen a random pairing that also met the above criteria.

Since we invited some participants to assess more than one batch, we attempted to mitigate any potential learning effects by aligning assessors as best as possible based upon their prior experience in the user study, such that: the first batch reviewed by one participant was matched to the first batch reviewed by another; the second batch reviewed by one assessor was matched to the second batch reviewed by another participant; and so on. Due to the disparity in number of batches assessed (discussed in Section 6.1.2), some batches were matched with discordant assessor experience. We do not believe that this had a material impact on the results.

6.2 Judging Behaviour

Predictor	Pr[User$^+$|Predictor]	p-value
Context: Relevance	0.42 (0.36,0.48)	-
Context: Uncertainty	0.54 (0.48,0.59)	0.0002
Context: Random	0.54 (0.48,0.60)	0.0002
Rel: W-RI	0.63 (0.58,0.68)	< 0.0001
Rel: W-NU	0.39 (0.34,0.43)	

Table 6.2: Probability of a study participant making a positive assessment, with 95% confidence intervals, for the primary predictors. For context, p-values were computed using a two-tailed paired binomial test; for relevance, p-values were computed using a z-test for difference in proportions.

Predictor	Pr[User$^+$|Predictor]	p-value
Relevance and W-RI	0.52 (0.43,0.61)	-
Uncertainty and W-RI	0.67 (0.58,0.75)	0.0037
Random and W-RI	0.70 (0.62,0.78)	0.0005
Relevance and W-NU	0.33 (0.26,0.41)	-
Uncertainty and W-NU	0.42 (0.34,0.50)	0.0288
Random and W-NU	0.40 (0.32,0.48)	0.1214

Table 6.3: Probability of a study participant making a positive assessment, with 95% confidence intervals, for combined predictors. p-values were computed relative to CAL, using a two-tailed paired binomial test.

From Table 6.2, we can see that uncertainty sampling and random sampling individually yield a higher probability of a positive assessment than relevance sampling, where the differences are substantial and statistically significant. Furthermore, the W-RI relevance class yields a higher probability of positive assessment than does W-NU, which is also substantial and statistically significant. These results indicate that context and relevance class separately affect the probability of a positive assessment, which is in accord with our primary and secondary hypotheses.

When examining the combined effect of context and relevance class, as reported in Table 6.3, we see that when the underlying documents were from the W-RI relevance class and uncertainty or random sampling was used, a substantially and significantly higher

probability of positive assessment was rendered than when relevance sampling was used. On the other hand, while the differences of uncertainty and random sampling are materially different from those of relevance sampling with the W-NU relevance class, only the difference between relevance and uncertainty sampling is significant. However, applying Bonferroni correction for multiple hypothesis testing renders that result not significant.

We find that the difference: $\Pr[\text{User}^+|\text{Relevance} \wedge \text{W-NU}] - \Pr[\text{User}^+|(\text{Uncertainty} \vee \text{Random}) \wedge W - NU)]$ is significant ($p < 0.0288$), by the following argument: for the null hypothesis to be true, it would be necessary for
$\Pr[\text{User}^+|\text{Relevance} \wedge \text{W-NU}] \geq \Pr[\text{User}^+|\text{Uncertainty} \wedge \text{W-NU})]$ *and*
$\Pr[\text{User}^+|\text{Relevance} \wedge \text{W-NU}] \geq \Pr[\text{User}^+|\text{Random} \wedge \text{W-NU})]$. It cannot be the case that the probability of this combination's occurring exceeds the probability of either event's occurring separately, which means that $p < \min(0.0288, 0.1214) = 0.0288$.

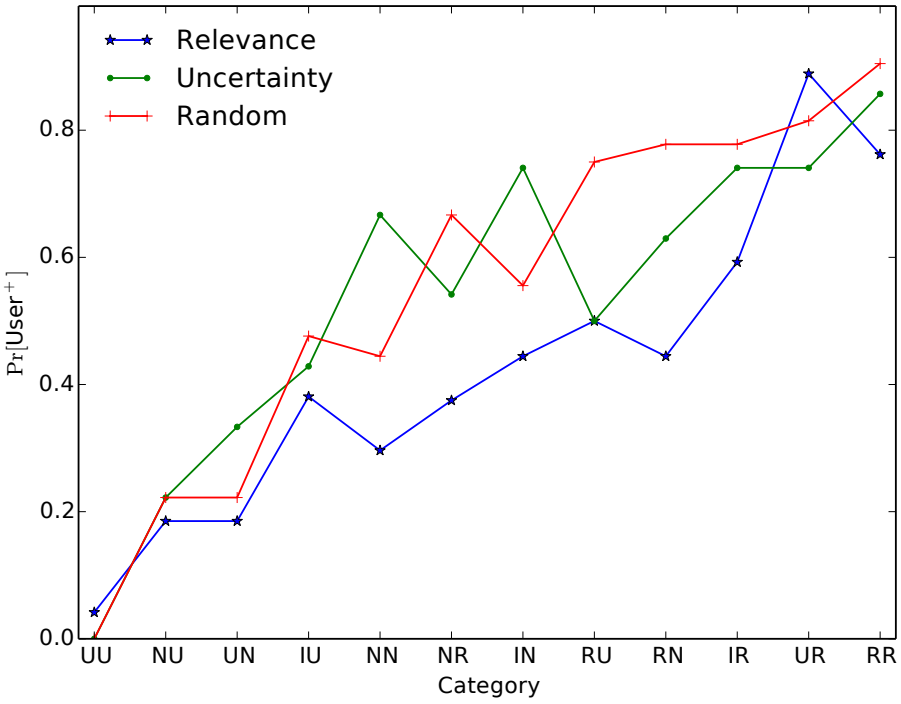


Figure 6.4: Probability of positive assessment given a context and elementary relevance class. Relevance classes are denoted by xy where $x \in R, I, N, U$ denotes Waterloo relevant, iffy, non-relevant and unjudged, and $y \in R, N, U$ denotes NIST relevant, non-relevant, and unjudged.

For all combinations of Waterloo and NIST assessments, Figure 6.4 plots the probability

of a positive user assessment. The resulting curves for random sampling and uncertainty sampling are generally higher than the curve for relevance sampling. In cases where one of the Waterloo or NIST assessments is "relevant" and the other is "relevant" or "unjudged" (i.e., they do not contradict each other), there does not appear to be a substantial difference between relevance sampling and the two other sampling strategies. A similar trend appears when Waterloo assessed a document as "iffy" and NIST rendered an assessment for the document. Interestingly, the Waterloo "iffy" but NIST "unjudged" combination has a similar probability of positive assessments to other cases where one of the two sets thought the document relevant. Such an occurrence could be indicative of a missed relevant document but might also just be a quirk of the document selection process employed by Waterloo. Whether these observations reflect chance or the effect of context is a good candidate for further investigation. In general, these curves support the above statistical results and provide additional evidence that our hypotheses hold.

6.3 Assessment Time

Predictor	Time Taken Per Doc	p-value
Rel: W-RI	26.56 (23.86,29.27)	0.1775
Rel: W-NU	23.92 (21.19,26.64)	
Context: Relevance	26.44 (22.69, 30.19)	-
Context: Uncertainty	23.53 (20.58, 26.48)	0.1984
Context: Random	25.46 (22.21, 28.72)	0.6781

Table 6.4: Average time, in seconds, taken to assess documents under each condition, with 95% confidence intervals, for all documents and for known documents only. For context, p-values are with respect to a paired two-tailed t-test against the relevance sampling predictor. For relevance, p-values are from Welch's t-test.

For the purposes of measuring candidacy for the continuation criteria, we collected the time a participant took to judge a document. This time was calculated as the difference in time from the document's being loaded into the page and the user's making an assessment. Table 6.4 shows that neither context nor relevance class appears to have significant or substantial impact on the time to render an assessment. Figure 6.5 plots the time taken against all combinations of Waterloo and NIST relevance classes. Unlike in Figure 6.4, there does not appear to be any substantial differences among document contexts. Interestingly, this plot does seem to suggest that documents which Waterloo *and* NIST deemed "not

Avg. Time to Judge	Correct	False Positive	False Negative
Normalized Time	-0.06 (-0.08, -0.04)	0.26 (0.16, 0.35)	0.40 (0.29, 0.51)
Raw Times in Seconds	14.30 (13.54, 15.06)	19.27 (17.70, 20.85)	25.20 (22.82, 27.57)

Table 6.5: Replication of Smucker and Jethani's analysis [139] of errors and time to make judgments in the context of all documents and all participants. Normalized time results from normalizing all raw times to have a mean of 0 and a standard deviation of 1.

relevant" take longer to judge. The cause of such an occurrence is not readily apparent, but it would be an interesting avenue of future research to determine whether this behaviour is a manifestation of Smucker and Jethani's observation [139] that assessors take longer to make incorrect assessment.

To shine some light on this issue, we reproduce in Table 6.5 the analysis of Smucker and Jethani [139], using all of the documents assessed and all participants from our user study. Smucker and Jethani reported statistics for the raw time spent reviewing as well as statistics after the times had been normalized to have a mean of 0 and a standard deviation of 1. Unlike Smucker and Jethani, we report 95% confidence intervals rather than p-values, primarily from author preference. Our results generally agree with those observed by Smucker and Jethani, where making incorrect assessments does generally take longer to make. In a reversal of Smucker and Jethani's results, our results indicate that it takes longer to make a false negative mistake compared to making a false positive. The immediate cause is not apparent, but it could stem from differences in experimental set-up.

6.4 Discussion

We have seen that assessors are less likely to judge a document as relevant (i.e., a positive assessment) when they are presented in the context of relevance sampling than when documents are presented based upon uncertainty or random sampling. This behaviour holds regardless of whether the document was assessed as relevant or "iffy" by the University of Waterloo's archival assessments for TREC-6. Further investigation is required to determine whether this result applies solely to marginally relevant documents or to all relevant documents. Neither context nor relevance class appears to have a significant effect on the time required to judge a document for topical relevance, though our results appear to suggest that incorrect judgments take longer to make, which is in accordance with the results of Smucker and Jethani [139].

Our results have ramifications beyond the selection of training documents for the pur-

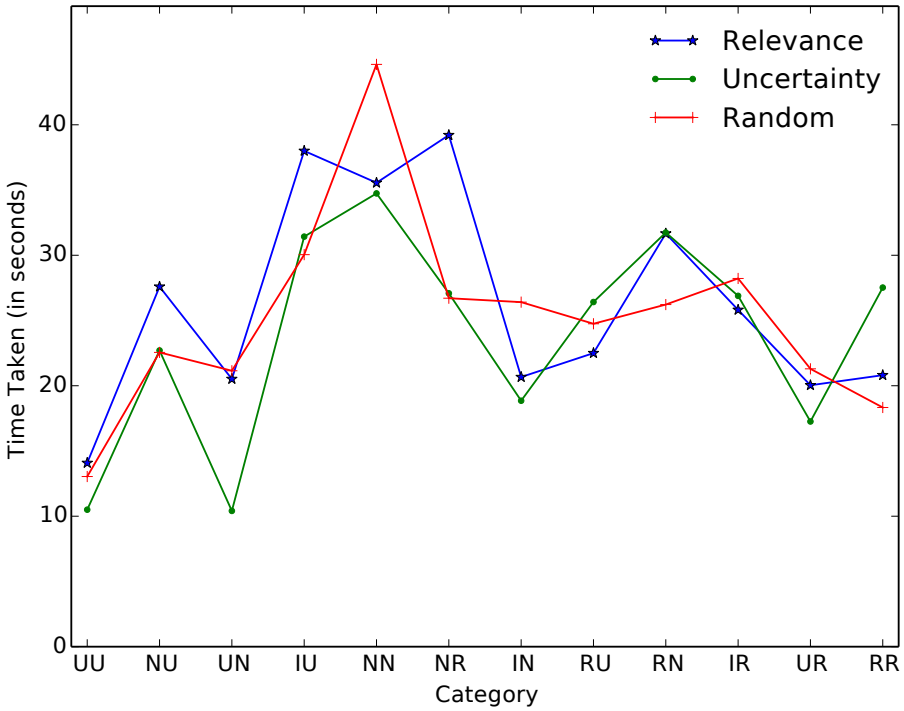


Figure 6.5: Average time for assessment given a context and elementary relevance class. Relevance classes are denoted by xy where $x \in R, I, N, U$ denotes Waterloo relevant, iffy, non-relevant and unjudged, and $x \in R, N, U$ denotes NIST relevant, non-relevant, and unjudged.

poses of inducing a classifier. In technology-assisted review (i.e., electronic discovery), every document with a positive classification is assessed and labelled by a human. These results would indicate that a superior (higher-precision) classifier may yield fewer documents labelled relevant than an inferior (lower-precision) classifier, in spite of the fact that the former may identify more relevant documents overall than the latter.

Furthermore, under the Cranfield paradigm [153] of information retrieval evaluation, a set of relevance assessments based upon a depth pool may produce substantially different results from a set of assessments sampled from the whole collection. Accordingly, we would expect that comparing the evaluation results of one evaluation using depth pooling with one using stratified sampling to yield noticeable differences even when the systems and topics are held constant.

The results presented in this chapter and in Chapter 3, in conjunction with previous

studies [63, 82, 115, 137, 139], call into question the practice of deeming one individual's assessments to be authoritative for the purposes of evaluation. All together, these results indicate that any authoritative assessor will be influenced by the context, the (underlying "true") relevance class, and by the apparent prevalence of the documents they are to judge. Changing any one of these factors can produce substantial changes to not only the assessments themselves but also the resulting evaluations that make use of the assessments.

Due to these potential issues, the community would benefit greatly from having more accurate models of assessing behaviour and, in particular, of how different judging contexts affect the rates of relevant and non-relevant assessment. These models could be used in ways similar to that of assessor error [162, 149, 157] to improve estimates of effectiveness. While we cannot infer such models from the user study presented in this chapter, it has provided evidence that such models are needed and could be beneficial in more accurately assessing system performance.

The user study presented herein is not without its own limitations. First and foremost, the task conducted by the participants was not representative of the real-world scenario where participants' assessments would drive the active learning strategies and influence the next document(s) shown to them. The choice to employ predetermined (i.e., static or fixed) batches was driven by the desire to maintain consistency across assessors; that is, we can more carefully control the contextual documents than would be possible in a more realistic setting. In particular, this prohibited participants from "going down rabbit holes," if they mislabelled documents, which could have resulted in vastly different experiences for participants even when the underlying sampling strategy was the same. Additional research is necessary to determine whether or not these results would be replicated in a "live fire" scenario and, if so, to what degree they would manifest themselves.

Such an experiment could also facilitate adjusting the protocol with a more appropriate learning algorithm for uncertainty sampling (i.e., SVMs). While maintaining the logistic regression classifier would maintain inter-experimental consistency, changing to SVM classification would result in a more ecologically valid application of uncertainty sampling. That is, SVMs and uncertainty sampling both attempt to improve the decision boundary between classes. In the case of SVMs, it is done by attempting to maximize the distance between the separating hyperplane and the points on either side of it. In the case of uncertainty sampling, this is done by clarifying which side of the decision boundary "borderline" documents actually fall on. Conversely, in the case of logistic regression, the goal is to maximize the log-odds of positive classification, which does appear to align more closely with relevance sampling. It is worth noting that removing the `--roc` option, which was inherited from the Total Recall track's BMI and optimizes ranking performance, may also provide a more amenable setting for using uncertainty sampling with logistic regression.

That being said, we do not have reason to believe that the results presented for uncertainty sampling are invalid or without merit, simply that they do not portray a best-case scenario. It may also be sufficient to conduct a small replication study consisting only of SVM-based uncertainty sampling and to compare those results with those observed here to determine whether the observed behaviours align.

Chapter 7

Future Work

This chapter is dedicated to various extensions and continuations of work presented elsewhere in thesis and to ideas that stem from actually conducting those experiments. Accordingly, there may be some thematic dissonance with the rest of the thesis due to the fact that many of these are related more to implementation concerns rather than to evaluation.

7.1 Bad Feedback Is Better Than No Feedback

In Chapter 3, we investigated how surrogate assessments compare to authoritative assessments when it comes to system training and evaluation. One of the major limitations of that work was its sole focus on a single batch step of passive supervised learning that is related to Cormack and Grossman's simple passive learning protocol [40]. Cormack and Grossman have shown simple passive learning to be generally inferior to active learning protocols, such as simple active learning or continuous active learning. A logical follow-up to our study would be to investigate the impact of "bad" feedback on active learning protocols for high-recall retrieval.

In such a study, the goal would not necessarily be to compare active and passive learning strategies, but to examine the effect that surrogate assessments have on such protocols when compared to the authority. Fundamentally this forms the next logical step to Chapter 3 and should be relatively straightforward to implement.[1] Additionally, it would desirable to extend such experiments to the use of real-world high-recall systems and some inroads have

[1] Indeed, some preliminary work has been done but would likely have to be redone to ensure experimental consistency with any new work.

been made on that front. Such collaboration would likely make use of similar architecture to that of the Total Recall track, further highlighting its extensibility and reusability.

In terms of hypotheses for such follow-up experiments, we might reasonably hypothesize that when comparing within protocols (i.e., CAL to CAL), that the results of the previous study will hold and the authority will result in a superior classifier. If we do between-protocol comparison (i.e., CAL to SPL), then it would be reasonable to hypothesize that a surrogate trained active learning protocol may exceed the performance of an authoritatively trained passive system. Furthermore, continuous active learning, given its generally strong performance, will likely outperform simple active learning in a similar experimental set-up. Should such hypotheses be confirmed by the experiments, additional evidence would exist that the "garbage in, garbage out" phenomenon does not occur when using surrogate assessments in these ways[2]. This confirmation would also provide additional evidence that active learning strategies are preferable to passive ones.

7.2 Relative Review Cost/Quality versus Quantity

Chapter 3 operated under the assumption that surrogate assessments are often portrayed as being cheaper and less effective than authoritative assessments. Accordingly, we argued that under such a model taking the union of two sets of surrogate assessments would still likely be cheaper than having an authoritative assessor render such judgments. Following this line of reasoning, we may be reasonably be able to achieve better performance with less cost by using more surrogate assessments than by using authoritative assessments. For example, we might instead wish to have a single surrogate judge twice as many documents as the authoritative assessor for the same cost and achieve better performance due to the greater number of training documents. We use the term relative review cost to reflect the number of document assessments that can be rendered by a surrogate assessor for the same cost as that of an authoritative one.

This study is interested in investigating relative review cost of surrogate assessments as it relates to high-recall retrieval protocols (e.g., CAL, SAL, SPL). For example, a relative review cost of 2 would imply that an authoritative assessment is worth 2 surrogate assessments. Following this line, an IR expert might have a relative review cost of 8 when compared to a senior lawyer, while a PhD student might have a relative review cost of 5 to that same IR expert. Accordingly, the PhD student might then have a relative review cost of 40 when compared to the senior lawyer.

[2]We readily admit that a machine learner trained with poorly labelled data will yield garbage. However, we do not believe that surrogate assessments necessarily fall under such a category.

It is worth noting that empirical validation is required to determine at what relative review costs surrogate assessments become an attractive alternative to authoritative assessments (e.g., is it as low as 2 or as high as 20?). Similarly, validation is required to determine whether combinations of surrogates ever become lucrative when compared to single surrogate and authoritative assessors. In addition, this work is concerned with the training costs of high-recall retrieval protocols, as many applications (e.g., eDiscovery, systematic review) may require a final review by some form of authoritative assessor. Thus, there is little point in trying to replace the final review assessor with a surrogate.

With sufficient foresight, this study and the study discussed above (Section 7.1) can be done at the same time, as only the evaluation criteria are different; the classification tasks and training data are identical between the two experiments. Furthermore, the results of this study and the previous study may provide more insight into high-recall retrieval evaluation (discussed in Chapters 4 and 5).

7.3 Preference Ratios and High-Recall Evaluation

One of the interesting results from Chapter 5 is that while the BMI was routinely superior across the 103 topics of the RCV1 corpus (according to the corresponding gain curves) when compared to random sampling, most evaluation measures did not report a substantive difference between the two systems. Clarke and Smucker [34] proposed the *preference ratio* for two systems, A and B, when evaluated by a single user, i, over a set of topics, T, as follows:

$$r = \frac{1}{|\mathrm{T}|} \sum_{t \in \mathrm{T}} \mathrm{Pref}(s_{Ati}, s_{Bti})$$

where s_{Xti} denotes the score for system X on topic t according to user i, and Pref is defined as:

$$Pref(s_{Ati}, s_{Bti}) = \begin{cases} 1 & \text{if } s_{Ati} > s_{Bti} \\ 0.5 & \text{if } s_{Ati} = s_{Bti} \\ 0 & \text{if } s_{Ati} < s_{Bti} \end{cases}$$

They show that the average of the preference ratio across a set of users can be used as a measure for determining which system users would prefer on average. By constructing a bootstrap distribution over the set of topics, Clarke and Smucker showed that the preference ratio can then be used to determine statistical significance between pairs of systems.

In their work, Clarke and Smucker simulated users to determine how much time is spent reading relevant material (i.e., in their jargon, time that is "well spent") as a means

of evaluating systems. No such conception is present in high-recall retrieval evaluation, though it might be possible to adapt such a measure. However, the underlying assumption of an evaluation measure is that it reflects some (potentially unknown) user model. Accordingly, we could treat the measures in Chapters 4 and 5 as the "users" in a preference ratio calculation. This interpretation preference ratio is simply a barometer for how well systems compare across a set of measures, condensed into a single value. Since computing this measure is relatively straightforward and could be done alongside the more "traditional" bootstrap evaluation presented in Chapter 5, we believe that this avenue of future research is worth investigating. We note, however, that such a preference ratio merely tells us which system is better and sidesteps issues with absolute performance summarization. Additional work is still needed to a measure that can do perform both evaluation goals.

7.4 A General Purpose Platform for IR Experimentation

The SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR) [13] sought to explore how reproducible and generalizable the results of IR evaluation actually are. As part of this, they ran yhe Open-Source Information Retrieval Reproducibility Challenge [98],[3] which challenged IR researchers to submit baseline ad hoc search implementations from their in-house search software with the goal of providing reference implementations to help reviewers and other researchers determine what acceptable baseline results should be. This challenge required researchers to create shell scripts that would run their systems against various past TREC test collections on a Amazon EC2 instance which they could configure as desired. One of the reasons was to allow other researchers to download these scripts and replicate the results on their own machines. Unsurprisingly, not even a year later, some these scripts no longer work out of the box and require some manual tweaking[4]. Rather than having researchers submit shell scripts, it may have been wiser to have them submit virtual machines (or Docker containers) that interfaced with an ad hoc search API, potentially providing the corpus and topics. Such an approach could perhaps increase the longevity of these archival baseline systems.

Such a proposal is intentionally worded to sound reminiscent of the Total Recall architecture (presented in Chapter 4). Furthermore, the underlying evaluation architecture of the Real-Time Summarization track[5] [123] shares a common underlying infrastructure with

[3]Results can be found at: `https://github.com/lintool/IR-Reproducibility`

[4]This is derived from the author's anecdotal experience.

[5]Track details can be found at: `http://trecrts.github.io`

the Total Recall track. This is not surprising, since the author of both architectures is the author of this thesis. During the development of both platforms, it became increasingly apparent that many IR tasks can be operationalized as a simple API that systems could communicate with over the Web (or locally). However, having to rewrite the same, or very similar, code multiple times is tedious, time consuming, and puts the onus of development on a select group of individuals.

The fact that these two fairly different tasks (Total Recall and Real-Time Summarization) can share a large base of underlying code leads us to believe that the creation process could be automated by a small creation tool. This creation tool would take in a specification detailing various required aspects (e.g., whether to provide relevance feedback, or to provide endpoints to send documents to assessors) and automatically create the necessary Web server which provides the specified API. While there may still need to be custom code written to support unforeseen or one-off evaluation tasks, such a tool would have the potential to greatly speed up deployment of such servers. Additionally, that custom code would be written by the IR researcher and not necessarily the author of this thesis. The exact details of how to actually create this tool have not made it past planning stages, but progress will be made in this direction.

7.5 Extrapolating Relevance for Unjudged Documents

In Chapter 3, we used only assessments that corresponded to documents judged by all assessors, which in some cases dramatically limited the number of assessments available. At the same time in Chapter 4, great effort was taken to create as complete an assessment effort as possible for the test collections used. In these types of cases, the ability to infer a relevance judgment for previously unjudged documents would be desirable and beneficial. For example, in the case of the TREC-4 surrogate assessments [152], the two surrogate assessors judged only a subset of the total pool. Being able to extend their judgments to the entire pool may yield more accurate experimental results. A similar case can be made for the TRE- 6 assessments generated by NIST and the University of Waterloo: extending both sets of judgments to comprise the entire set of reviewed documents may provide a more complete evaluation framework when using the TREC-6 data set.

There are several potential options when attempting extrapolation of one set of judgments to another:

1. No extrapolation. Consider only the intersection of the evaluated documents.

- This approach has been used in the experiments discussed in this thesis and forms the baseline approach.

2. Benefit of the doubt. For any unjudged document that is judged by one of the assessors, use the assessment rendered by that assessor.

   - This approach is highly optimistic and may not reflect actual agreement.

3. For any unjudged document assume that it is not relevant for that assessor.

   - This is highly pessimistic and may not reflect the assessor's previous behaviour.

4. Ignore any document that is assessed relevant by one assessor but unjudged by the other. Assume that unjudged and not relevant would have resulted in not relevant.

   - The idea here is that we want to minimize false positive for the unjudged documents, since there are many fewer relevant than non-relevant documents.

5. Interpolate. Use machine learning and train on agreements/disagreements for multiply judged documents and then infer judgment based upon classification results.

   - This can be done per class or just general agreement/disagreement.

It is not clear which of these approaches (if any) is best. However, the ability to extend relevance assessments based upon multiply assessed subsets would provide potentially more "complete" relevance assessments, which would further the ability of researchers to evaluate their algorithms with alternate assessors. Alternate assessors are useful, since, as discussed in Chapter 3, much of the effect of the "authoritative" TREC assessor is derived from the fact that they are used for both training and evaluation. Complete alternate sets of surrogate assessments would facilitate attempts to ensure algorithms are not just biased towards the de facto authoritative assessor.

There is an outstanding issue: How can we determine which of these approaches is better? The solution is not clear. We could attempt to maintain the true positive and false positive rates for the alternate assessor, or we could look at system-ranking agreements or significant differences between systems when comparing partial and complete alternate assessments. A simple but likely effective approach would be to follow standard machine learning protocol and perform cross-validation on the documents for which we know the actual alternate assessment. This, at least, would tell us how effective the above algorithms are for predicting the correct label but, further investigation will invariably be required.

Chapter 8

Conclusions

High-recall retrieval is a vast domain, ranging from implementation to evaluation to effects on assessing behaviour and beyond. The domain is important not only in academia (e.g., test collection creation in information retrieval) but has become extremely important in the real world (e.g., electronic discovery in civil litigation), where inferior systems and evaluation can result in wasted time, money, and effort. This thesis has been chiefly focused on increasing awareness of evaluation issues and creating a framework for replicable high-recall retrieval experiments. At its core, this thesis attempts to lay a base upon which best practices for both academic and industrial research can be built.

We began this thesis with an investigation into the presumed "infallibility" of gold standard or authoritative assessors. More precisely, we examined the "garbage in, garbage out" belief, common in eDiscovery settings [68, 5], that derides the use of cheaper surrogate assessors, in spite of evidence [112, 163, 152, 162] that challenges this belief. We have shown across several test collections that authoritative assessors possess no special attribute that makes them more suited or capable of the task, other than having been selected to be the authority. Indeed, when surrogates have more diverse or liberal interpretations of relevance, we have seen that they tend to perform competitively with other surrogates and in some cases rival the authoritative assessor. If the belief of "garbage in, garbage out" were true, such interpretations, which often result in more false positives ("garbage"), should perform worse, but this was not the case in our experiments. Our findings indicate that when evaluating the results of a classifier trained by one assessor using an independent set of assessments, a large part of what is measured is a difference of opinion rather than a magnification of error. Accordingly, this leads us to conclude that it is necessary to ask, when determining the effectiveness of a system, "according to whom?" Performing well

according to an independent third party is a resounding triumph but unremarkable when self-assessed.

Following this, a framework for the evaluation of high-recall retrieval systems was presented. The framework was designed to take into account several issues that had arisen in several previous TREC tracks, including interaction with the gold standard assessor biasing results; trouble in automating and distributing systems that must provide a particular interface; and the inability to distribute more realistic test collections, since they can contain personally identifiable or other sensitive information. To mitigate these issues, we designed a Web service that provided a REST(ful) API that encapsulated document and topic distribution while providing document-at-a-time relevance assessment. Furthermore, the framework was designed so that retrieval systems could be encapsulated in a virtual machine and run against this Web service in an air-walled environment with only summary results being produced. By creating such a framework, we have attempted to mitigate those aforementioned issues as well as to facilitate the replication of results through the standardization of an evaluation framework.

The proposed evaluation framework was validated through its use in the TREC 2015 Total Recall track, where participants designed high-recall retrieval systems to interact with this Web service. To further aid system designers, a baseline system was provided in the form of a virtual machine that they were encouraged to use as an exemplar system and, if they desired, as a template for their own system. As part of the Total Recall track, 3 test collections were labelled as completely as possible for 30 topics. These 3 collections provided the basis for over-the-Internet participation and were provided entirely through the Web service. Two private collections, also completely and independently labelled, were used as a validation of the framework but also for the results from the over-the-Internet experiments. Three industrial teams used manually directed systems and 7 academic teams produced purely automatic systems in the track's first iteration. By and large, we found that the manual and automatic runs were competitive with each other and that no system was consistently better across all topics, collections, or evaluation measures. To further investigate this result, we conducted a pilot facet-based evaluation but failed to find any substantial differences from the main results. We posit that this may be due to some combination of less than ideal facets, binary relevance grades, and the particular facet-based evaluation measures used.

Based upon the Total Recall evaluation results, we investigated the suitability of several possible evaluation metrics in terms of distinguishing between systems and the consistency of system performance on those measures. We began by comparing two strategies that have been shown to be substantially different [40]: relevance and random sampling. When comparing these strategies on the Reuters Collection Volume 1 collection, we surprisingly

did not find any significant differences between the two strategies. This result appears to be a result of high per-topic variance in the number of relevant documents in conjunction with the generally high occurrence of relevant material in the collection. To validate our experimental methodology, we re-ran the experiment on the TREC-6 ad hoc test collection and found significant and substantial differences across all measures used. While such a result did not necessarily determine that one measure was materially better than other, these validation results indicate that the results on RCV1 may be due to properties of the test collection and not the experiment itself. As part of this experiment, we proposed a new measure, the relative gain curve, which measures recall as a function of the effort expended relative to the number of relevant documents: $\frac{effort}{R}$. When compared to the gain curve, which measures recall as a function of raw effort, we found that the relative gain curve reports behaviour that is hidden by the the gain curve.

Based upon the baseline results, we then compared the gain curve and the relative gain curve on the 2015 Total Recall track submissions. Across these collections, the relative gain curve displayed behaviour that was different from the gain curve and reminiscent of the results observed in the precision-recall curves used in the official evaluation of the track. To our surprise, we found that the measurements of recall were no more consistent and (sometimes) less consistent when using relative effort rather than absolute effort. The cause of this is not apparent and requires further investigation, perhaps by accounting for a fixed level of overhead in the computation of relative effort (e.g., to mitigate variance caused by the cold-start problem). On the other hand, the `Sandbox` collections appear to result in more consistent estimates than any of the `At-Home` collections, which may indicate that part of the observed variability stems from some interaction between topic prevalence, the collection itself, and the task being simulated. However, these are preliminary hypotheses and more investigation is necessary before firm conclusions can be made and the implications on evaluation understood.

This thesis, as well as much of the literature on information retrieval, has followed the Cranfield paradigm, which relies on the underlying assumption that the relevance assessments used will be rendered identically regardless of how they are presented to users. For example, this assumption means that regardless of which Total Recall track system we choose as the basis for selecting documents to assess, the assessor will make the same judgments as those used in the track. This assumption is made due to the fact that collecting new judgments is inefficient and would compromise the replicability and reusability of test collections. When these judgments are used solely to evaluate a system, rather than train it, such an assumption is reasonable; otherwise test collections would be one-off creations. That being said, there have been many investigations [82, 63, 115, 86, 137, 138, 139, 119, 15, 132, 165, 10, 11] into the effect that

presentation order has in an ad hoc search setting, but little investigation has been done in high-recall settings. This thesis presents the design, execution, and the results of a user study exploring the effect that three high-recall training strategies can have on assessing behaviour.

Primarily, we found that relevance sampling is significantly and substantially more likely to induce a lower probability of marking a document relevant when compared to random and uncertainty sampling. Interestingly, there appears to be no difference in the time to make judgments among these three strategies, though much as in earlier work [139], we found that assessors take longer to make incorrect assessments. We also find that the underlying relevance (i.e., whether the document was trained as relevant) appears to significantly affect the probability of a relevant assessment. Furthermore, the interaction between sampling strategy and underlying relevance is significant when the underlying relevance is positive, though this may be affected by the presence of marginally relevant documents. Our results indicate that it would be questionable to equate experimental validation when different relevance assessing strategies are used (e.g., stratified sampling versus depth pooling).

References

[1] Workshop on supporting search and sensemaking for electronically stored information in discovery proceedings (desi workshop). http://www.umiacs.umd.edu/~oard/desi-ws/, 2007.

[2] Second international workshop on supporting search and sensemaking for electronically stored information in discovery proceedings (desi ii workshop). http://web.archive.org/web/20080414173119/http://www.cs.ucl.ac.uk/staff/S.Attfield/desi/, 2008.

[3] Global e-discovery/e-disclosure workshop (desi iii workshop). https://web.archive.org/web/20130902124718/http://www.law.pitt.edu/DESI3_Workshop, 2009.

[4] Workshop on setting standards for searching electronically stored information in discovery proceedings (desi iv workshop). http://www.umiacs.umd.edu/~oard/desi-ws/, 2011.

[5] *Da Silva Moore v. Publicis Groupe*, 2012. 287 F.R.D. 182, S.D.N.Y.

[6] Workshop on standards for using predictive coding, machine learning, and other advanced search and review methods in e-discovery (desi v workshop). http://www.umiacs.umd.edu/~oard/desi5/, 2013.

[7] Workshop on using machine learning and other advanced techniques to address legal problems in e-discovery and information governance (desi vi workshop). http://www.umiacs.umd.edu/~oard/desi6/, 2015.

[8] *Global Aerospace Inc., et al., v. Landow Aviation L.P., et al.*, Apr. 23, 2012. No. CL 61040 (Va. Cir. Ct.).

[9] Eugene Agichtein, David Carmel, Yuval Pinter, and Donna Harman. Overview of the TREC 2015 LiveQA Track. In *Proceedings of the Twenty-Fourth Text REtrieval Conference (TREC 2015)*, 2015.

[10] Aiman L. Al-Harbi and Mark D. Smucker. User expressions of relevance judgment certainty. In *Proceedings of the Seventh Annual Workshop on Human-Computer Interaction and Information Retrieval*, HCIR '13, 2013.

[11] Aiman L. Al-Harbi and Mark D. Smucker. A qualitative exploration of secondary assessor relevance judging behavior. In *Proceedings of the 5th Information Interaction in Context Symposium*, IIiX '14, pages 195–204, 2014.

[12] James Allan. HARD Track Overview in TREC 2005. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, 2005.

[13] Jaime Arguello, Fernando Diaz, Jimmy Lin, and Andrew Trotman. Sigir 2015 workshop on reproducibility, inexplicability, and generalizability of results (rigor). In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 1147–1148, 2015.

[14] Mossaab Bagdouri, William Webber, David D. Lewis, and Douglas W. Oard. Towards minimizing the annotation cost of certified text classification. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, CIKM '13, pages 989–998, 2013.

[15] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. Relevance assessment: Are judges exchangeable and does it matter? In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 667–674, 2008.

[16] Thomas Barnett and Svetlana Godjevac. Faster, better, cheaper legal document review, pipe dream or reality? In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Law (DESI IV Workshop)*, 2011.

[17] Thomas Barnett, Svetlana Godjevac, Jean-Michel Renders, Caroline Privault, John Schneider, and Robert Wickstrom. Machine learning classification for document review. In *Proceedings of the Twelveth International Conference on Artificial Intelligence and Law (DESI I Workshop)*, 2009.

[18] Jason R. Baron, David D. Lewis, and Douglas W. Oard. TREC 2006 Legal Track overview. In *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)*, 2006.

[19] Robert S. Bauer, Dan Brassil, Christopher Hogan, Gina Taranto, and John Seely Brown. Impedance matching of humans $\leftrightarrow$ machines in high-Q information retrieval systems. In *2009 IEEE International Conference on Systems, Man and Cybernetics*, pages 97–101, 2009.

[20] Robert S. Bauer, Teresa Jade, and Mitchell P. Marcus. STIR: Simultaneous achievement of high precision and high recall through socio-technical information retrieval. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Law (DESI I Workshop)*, 2007.

[21] David C. Blair. Some thoughts on the reported results of TREC. *Information Processing and Management*, 38(3), 2002.

[22] David C. Blair and M. E. Maron. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3):289–299, 1985.

[23] Dan Brassil, Christopher Hogan, and Simon Attfield. The centrality of user modeling to high recall with high precision search. In *2009 IEEE International Conference on Systems, Man and Cybernetics*, pages 91–96, 2009.

[24] Carla E. Brodley and Mark A. Friedl. Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11:131–167, 1999.

[25] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. Bias and the limits of pooling for large collections. *Information Retrieval*, 10(6):491–508, 2007.

[26] Chris Buckley and Stephen Robertson. Relevance feedback track overview: TREC 2008. In *Proceedings of the Seventeenth Text REtrieval Conference (TREC 2008)*, 2008.

[27] Stefan Büttcher, Charles L. A. Clarke, and Gordon V. Cormack. *Information Retrieval - Implementing and Evaluating Search Engines*. MIT Press, 2010.

[28] David Carmel, Elad Yom-Tov, and Ian Soboroff. SIGIR workshop report: Predicting query difficulty-methods and applications. *ACM SIGIR Forum*, 39:25–28, 2005.

[29] Ben Carterette, James Allan, and Ramesh Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 268–275, 2006.

[30] Ben Carterette and Mark D. Smucker. Hypothesis testing with incomplete relevance judgments. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 643–652, 2007.

[31] Ben Carterette and Ian Soboroff. The effect of assessor error on ir system evaluation. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 539–546, 2010.

[32] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011.

[33] Jianlin Cheng, Amanda Jones, Caroline Privault, and Jean-Michel Renders. Soft Labeling for Multi-Pass Document Review. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law (DESI V Workshop)*, 2013.

[34] Charles L. A. Clarke and Mark D. Smucker. Time well spent. In *Proceedings of the 5th Information Interaction in Context Symposium*, IIiX '14, pages 205–214, 2014.

[35] Cyril W. Cleverdon. The effect of variations in relevance assessments in comparative experimental tests of index languages. *Cranfield University Technical Report*, Oct. 1970.

[36] Gordon V. Cormack. TREC 2006 Spam Track Overview. In *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)*, 2006.

[37] Gordon V. Cormack. TREC 2007 Spam Track Overview. In *Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007)*, 2007.

[38] Gordon V. Cormack, Charles L. A. Clarke, Christoper R. Palmer, and Samuel S. L. To. Passage-Based Refinement (MultiText Experiments for TREC-6). In *Proceedings of the Sixth Text REtrieval Conference (TREC-6)* , 1997.

[39] Gordon V. Cormack and Maura R. Grossman. The Grossman-Cormack Glossary of Technology-Assisted Review. *Federal Courts Law Review*, 7(1):1–34, 2013.

[40] Gordon V. Cormack and Maura R. Grossman. Evaluation of Machine-learning Protocols for Technology-assisted Review in Electronic Discovery. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 153–162, 2014.

[41] Gordon V. Cormack and Maura R. Grossman. Autonomy and reliability of continuous active learning for technology-assisted review. *CoRR*, abs/1504.06868, 2015.

[42] 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*, SIGIR '15, pages 763–766, 2015.

[43] Gordon V. Cormack and Maura R. Grossman. Waterloo (Cormack) Participation in the TREC 2015 Total Recall Track. In *Proceedings of the Twenty-Fourth Text REtrieval Conference (TREC 2015)*, 2015.

[44] Gordon V. Cormack and Maura R. Grossman. Engineering quality and reliability in technology-assisted review. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 75–84, 2016.

[45] Gordon V. Cormack and Maura R. Grossman. Scalability of continuous active learning for reliable high-recall text classification. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 1039–1048, 2016.

[46] Gordon V. Cormack, Maura R. Grossman, Bruce Hedin, and Douglas W. Oard. Overview of the TREC 2010 Legal Track. In *Proceedings of the Nineteenth Text REtrieval Conference (TREC 2010)*, 2010.

[47] Gordon V. Cormack and Aleksander Kolcz. Spam Filter Evaluation with Imprecise Ground Truth. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 604–611, 2009.

[48] Gordon V. Cormack and Thomas R. Lynam. Spam Corpus Creation for TREC. In *Proceedings of the Second Conference on Email and Anti-Spam (CEAS 2005)*, 2005.

[49] Gordon V. Cormack and Thomas R. Lynam. TREC 2005 Spam Track Overview. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, 2005.

[50] Gordon V. Cormack and Thomas R. Lynam. Statistical precision of information retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 533–540, 2006.

[51] Gordon V. Cormack and Thomas R. Lynam. Power and bias of subset pooling strategies. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 837–838, 2007.

[52] Gordon V. Cormack and Mona Mojdeh. Machine Learning for Information Retrieval: TREC 2009 web, relevance feedback and legal tracks. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, 2009.

[53] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. Efficient construction of large test collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 282–289, 1998.

[54] Anita Crescenzi, Diane Kelly, and Leif Azzopardi. Time pressure and system delays in information search. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 767–770, 2015.

[55] Anita Crescenzi, Diane Kelly, and Leif Azzopardi. Impacts of time constraints and system delays on user experience. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pages 141–150, 2016.

[56] Vivek Dhand. Efficient semantic indexing via neural networks with dynamic supervised feedback. In *Proceedings of the Twenty-Fourth Text REtrieval Conference (TREC 2015)*, 2015.

[57] William C. Dimm. Information retrieval performance measurement using extrapolated precision. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Law (DESI VI Workshop)*, 2015.

[58] Kroll Discovery. eDiscovery around the globe. `http://www.ediscovery.com/ediscovery-around-the-globe/`.

[59] J. Stephen Downie. The music information retrieval evaluation exchange (2005 – 2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

[60] J. Stephen Downie, Andreas F. Ehmann, Mert Bay, and M. Cameron Jones. The music information retrieval evaluation exchange: Some observations and insights. *Advances in Music Information Retrieval*, pages 93–115, 2010.

[61] Harris Drucker, Behzad Shahrary, and David C. Gibbon. Support vector machines: Relevance feedback and information retrieval. *Information Processing and Management*, 38(3):305–323, 2002.

[62] Michael Eisenberg. *Mangitude Estimation and the Measurement of Relevance*. PhD thesis, Syracuse University, 1986.

[63] Michael Eisenberg and Carol Barry. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *Journal of the American Society for Information Science*, 39:293–300, 1988.

[64] John R. Frank, Max Kleiman-Weiner, Daniel A. Roberts, Ellen Voorhees, and Ian Soboroff. Evaluating stream filtering for entity profile updates in trec 2012, 2013, and 2014. In *Proceedings of the Twenty-Third Text REtrieval Conference (TREC 2014)*, 2014.

[65] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014.

[66] Norton Rose Fulbright. eDiscovery around the globe (White Paper). `http://www.nortonrosefulbright.com/files/20150529-ediscovery-around-the-globe-whitepaper-129403.pdf`, May 2015.

[67] Norton Rose Fulbright. eDiscovery around the globe: 2015 in review (White Paper). `http://www.nortonrosefulbright.com/files/20160223-ediscovery-around-the-globe-2015-in-review-137422.pdf`, February 2016.

[68] Dean Gonsowski. A look into the e-discovery crystal ball. *Inside Counsel*, 2011.

[69] Maura R. Grossman and Gordon V. Cormack. Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review. *Richmond Journal of Law and Technology*, 17:1–48, 2010.

[70] Maura R. Grossman and Gordon V. Cormack. Comments on "The Implications of Rule 26 (g) on the Use of Technology-Assisted Review". *Federal Courts Law Review*, 8:285–313, 2014.

[71] Maura R. Grossman, Gordon V. Cormack, Bruce Hedin, and Douglas W. Oard. Overview of the TREC 2011 Legal Track. In *Proceedings of the Twentieth Text REtrieval Conference (TREC 2011)*, 2010.

[72] Matthias Hagen, Steve G´oring, Magdalena Keil, Olaoluwa Anifowose, Amir Othman, and Benno Stein. Webis at TREC 2015: Tasks and Total Recall Tracks. In *Proceedings of the Twenty-Fourth Text REtrieval Conference (TREC 2015)*, 2015.

[73] Donna Harman. Overview of the Fourth Text REtrieval Conference (TREC-4). In *Proceedings of the Fourth Text REtrieval Conference (TREC-4)* , 1995.

[74] Donna Harman and Chris Buckley. Overview of the reliable information access workshop. *Information Retrieval*, 12(6):615–641, 2009.

[75] Bruce Hedin and Douglas W. Oard. Replication and automation of expert judgments: Information engineering in legal e-discovery. In *Proceedings of the 2009 IEEE International Conference on Systems, Man and Cybernetics*, SMC'09, pages 102–107, 2009.

[76] Bruce Hedin, Stephen Tomlinson, Jason R. Baron, and Douglas W. Oard. Overview of the TREC 2009 Legal Track. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, 2009.

[77] William Hersh. TREC 2002 interactive track report. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, 2002.

[78] Julian PT Higgins, Sally Green, et al. *Cochrane Handbook for Systematic Reviews of Interventions*, volume 4. John Wiley & Sons, 2011.

[79] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, pages 289–296, 1999.

[80] Christopher Hogan, Robert S. Bauer, and Dan Brassil. Automation of legal sensemaking in e-discovery. *Artificial Intelligence and Law*, 18(4):431–457, 2010.

[81] Christopher Hogan, Dan Brassil, Shana M. Rugani, Jennifer Reinhart, Misti Gerber, and Teresa Jade. H5 at TREC 2008 legal interactive: User modeling, assessment & measurement. In *Proceedings of the Seventeenth Text REtrieval Conference (TREC 2008)*, 2008.

[82] Mu-hsuan Huang and Hui-yu Wang. The influence of document presentation order and number of documents judged on users' judgments of relevance. *Journal of the American Society for Information Science*, 55(11):970–979, 2004.

[83] Gaya K. Jayasinghe, William Webber, Mark Sanderson, and J. Shane Culpepper. Extending test collection pools without manual runs. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 915–918, 2014.

[84] Gaya K. Jayasinghe, William Webber, Mark Sanderson, and J. Shane Culpepper. Improving test collection pools with machine learning. In *Proceedings of the 19th Australasian Document Computing Symposium*, ADCS '14, 2014.

[85] Chandra Prakash Jethani. *Effect of Prevalence on Relevance Assessing Behaviour*. PhD thesis, University of Waterloo, 2011.

[86] Chandra Prakash Jethani and Mark D. Smucker. Modeling the time to judge document relevance. In *Procedings of the SIGIR 2010 Workshop on the Simulation of Interaction*, pages 11–12, 2010.

[87] T. Joachims. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*, chapter 11. MIT Press, 1999.

[88] T. Joachims. Transductive inference for text classification using support vector machines. In *International Conference on Machine Learning (ICML)*, volume 99, pages 200–209, 1999.

[89] Karen Spärck Jones and C.J. van Rijsbergen. Report on the need for and provision of an 'ideal' information retrieval test collection. Technical report, University of Cambridge, 1975.

[90] Paul Kantor, K Myung-Ho, Ulukbek Ibraev, and Koray Atasoy. Estimating the number of relevant documents in enormous collections. In *Proceedings of the Annual Meeting of the American Society for Information Science*, volume 36, 1999.

[91] Kenneth A. Kinney, Scott B. Huffman, and Juting Zhai. How evaluator domain expertise affects search result relevance judgments. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 591–598, 2008.

[92] Aleksander Kolcz and Gordon V. Cormack. Genre-based Decomposition of Email Class Noise. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 427–436, 2009.

[93] Robert Lemos. Researchers reverse Netflix anonymization. `http://www.securityfocus.com/news/11497`.

[94] David D. Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. *SIGIR Forum*, 29(2):13–19, 1995.

[95] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 3–12, 1994.

[96] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, pages 361–397, 2004.

[97] Cheng Li, Yue Wang, Paul Resnick, and Qiaozhu Mei. ReQ-ReC: High Recall Retrieval with Query Pooling and Interactive Classification. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 163–172, 2014.

[98] Jimmy Lin, Matt Crane, Andrew Trotman, Jamie Callan, Ishan Chattopadhyaya, John Foley, Grant Ingersoll, Craig Macdonald, and Sebastiano Vigna. Toward reproducible baselines: The open-source IR reproducibility challenge. In *Proceedings of the 28th European Conference on Information Retrieval Research*, ECIR '16, pages 408–420, 2016.

[99] Ralph Losey, Jim Sullivan, and Tony Reichenberger. e-Discovery Team at TREC 2015 Total Recall Track. In *Proceedings of the Twenty-Fourth Text REtrieval Conference (TREC 2015)*, 2015.

[100] Peter Lubell-Doughtie and Kenneth Hamilton. Heliod at TREC Legal 2011: Learning to Rank from Relevance Feedback for e-Discovery. In *Proceedings of the Twentieth Text REtrieval Conference (TREC 2011)*, 2011.

[101] Peter Lubell-Doughtie and Katja Hofmann. Learning to rank from relevance feedback for e-discovery. In *Proceedings of the 34th European Conference on Advances in Information Retrieval Research*, ECIR '12, pages 535–539, 2012.

[102] Jiyun Luo, Christopher Wing, Hui Yang, and Marti Hearst. The water filling model and the cube test: Multi-dimensional evaluation for professional search. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 709–714, 2013.

[103] Mihai Lupu. TUW at the First Total Recall Track. In *Proceedings of the Twenty-Fourth Text REtrieval Conference (TREC 2015)*, 2015.

[104] Thomas R. Lynam and Gordon V. Cormack. On-line spam filter fusion. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 123–130, 2006.

[105] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. Judging relevance using magnitude estimation. In Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr, editors, *Proceedings of the 37th European Conference on Information Retrieval Research*, ECIR '15, pages 215–220, 2015.

[106] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[107] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[108] Douglas W. Oard, Jason R. Baron, Bruce Hedin, David D. Lewis, and Stephen Tomlinson. Evaluation of information retrieval for e-discovery. *Artificial Intelligence and Law*, 18(4):347–386, 2010.

[109] Douglas W. Oard, Bruce Hedin, Stephen Tomlinson, and Jason R. Baron. Overview of the TREC 2008 Legal Track. In *Proceedings of the Seventeenth Text REtrieval Conference (TREC 2008)*, 2008.

[110] Douglas W. Oard and William Webber. Information retrieval for e-discovery. *Foundations and Trends® in Information Retrieval*, 7(2–3):99–237, 2013.

[111] Supreme Court of the United States of America. *Federal Rules of Civil Procedure*.

[112] Jeremy Pickens. In TAR, wrong decisions can lead to the right documents (a response to Ralph Losey). `http://www.catalystsecure.com/blog/2014/02/in-`

tar-wrong-decisions-can-lead-to-the-right-documents-a-response-to-ralph-losey/, February 2014.

[113] Jeremy Pickens. An exploratory analysis of control sets for measuring e-discovery progress. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Law (DESI VI Workshop)*, 2015.

[114] Jeremy Pickens, Tom Gricks, Bayu Hardi, and Mark Noel. A Constrained Approach to Manual Total Recall. In *Proceedings of the Twenty-Fourth Text REtrieval Conference (TREC 2015)*, 2015.

[115] Lorraine M. Purgailis Parker and Robert E. Johnson. Does Order of Presentation Affect Users' Judgment of Documents? *Journal of the American Society for Information Science*, 41(7):493–494, 1990.

[116] Dan Regard and Tom Matzen. A Re-Examination of Blair and Maron. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law (DESI V Workshop)*, 2013.

[117] Stephen Robertson and Ian Soboroff. The TREC 2002 Filtering Track Report. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, 2002.

[118] Joseph J. Rocchio. Relevance feedback in information retrieval. In Gerard Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, chapter Chapter 14. Prentice-Hall, Inc., 1971.

[119] Adam Roegiest and Gordon V. Cormack. Impact of review-set selection on human assessment for text classification. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 861–864, 2016.

[120] Adam Roegiest, Gordon V. Cormack, Charles L. A. Clake, and Maura R. Grossman. TREC 2015 Total Recall Track Overview. In *Proceedings of the Twenty-Fourth Text REtrieval Conference (TREC 2015)*, 2015.

[121] Adam Roegiest, Gordon V. Cormack, Charles L. A. Clarke, and Maura R. Grossman. Impact of surrogate assessments on high-recall retrieval. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 555–564, 2015.

[122] Adam Roegiest, Gordon V. Cormack, Charles L. A. Clarke, and Maura R. Grossman. TREC 2015 total recall track overview. In *Proceedings of the Twenty-Fourth Text REtrieval Conference (TREC 2015)*, 2015.

[123] Adam Roegiest, Luchen Tan, Jimmy Lin, and Charles L. A. Clarke. A platform for streaming push notifications to mobile assessors. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 1077–1080, 2016.

[124] Herbert L. Roitblat, Anne Kershaw, and Patrick Oot. Document categorization in legal electronic discovery: Computer classification vs. manual review. *Journal of the American Society for Information Science and Technology*, 61(1):70–80, 2010.

[125] José A. Sáez, Mikel Galar, Julián Luengo, and Francisco Herrera. Analyzing the presence of noise in multi-class problems: Alleviating its influence with the one-vs-one decomposition. *Knowledge and Information Systems*, 38(1):179–206, 2014.

[126] Mark Sanderson and Justin Zobel. Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 162–169, 2005.

[127] Tefko Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in Information Science. *Advances in Librarianship*, 6:79–138, 1976.

[128] Tefko 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 Techngology*, 58(13):1915–1933, 2007.

[129] Tefko Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in Information Science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Techngology*, 58(13):2126–2144, 2007.

[130] Tefko Saracevic. Why is relevance still the basic notion in information science? (despite great advances in information technology). In *Proceedings of the 14th International Symposium on Information Science*, ISI '15, 2015.

[131] Karl Schieneman and Thomas C. Gricks III. Implications of rule 26 (g) on the use of technology-assisted review. *Federal Courts Law Review*, 7:239–274, 2013.

[132] Falk Scholer, Diane Kelly, Wan-Ching Wu, Hanseul S. Lee, and William Webber. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 623–632, 2013.

[133] Falk Scholer, Eddy Maddalena, Stefano Mizzaro, and Andrew Turpin. Magnitudes of relevance: Relevance judgements, magnitude estimation, and crowdsourcing. In *Proceedings of Sixth International Workshop on Evaluating Information Access (EVIA2014), a Satellite Workshop of the 11th NTCIR Conference*, NTCIR '14, pages 9–16, 2014.

[134] Johannes C. Scholtes, Tim van Cann, and Mary Mack. The Impact of Incorrect Training Sets and Rolling Collection on Technology Assisted Review. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law (DESI V Workshop)*, 2013.

[135] D. Sculley and Gordon V. Cormack. Filtering Email Spam in the Presence of Noisy User Feedback. *Proceedings of the Fifth Conference on Email and Anti-Spam (CEAS 2008)*, 2008.

[136] Fabrizio Sebastiani. Machine Learning in Automated Text Categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47, 2002.

[137] Mark D. Smucker and Chandra Prakash Jethani. Human performance and retrieval precision revisited. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 595–602, 2010.

[138] Mark D. Smucker and Chandra Prakash Jethani. Measuring assessor accuracy: A comparison of nist assessors and user study participants. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1231–1232, 2011.

[139] Mark D. Smucker and Chandra Prakash Jethani. Time to judge relevance as an indicator of assessor error. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 1153–1154, 2012.

[140] Ian Soboroff and Stephen Robertson. Building a filtering test collection for trec 2002. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 243–250, 2003.

[141] Eero Sormunen. Extensions to the STAIRS Study–Empirical Evidence for the Hypothesised Ineffectiveness of Boolean Queries in Large Full-Text Databases. *Information Retrieval*, 4(3-4):257–273, 2001.

[142] Stanley Smith Stevens. *Psychophysics*. Transaction Publishers, 1986.

[143] Genichi Taguchi. *Introduction to Quality Engineering: Designing Quality Into Products and Processes*. Quality Resources, 1986.

[144] Anastasios Tombros, Ian Ruthven, and Joemon M. Jose. How users assess web pages for information seeking. *Journal of the American society for Information Science and Technology*, 56(4):327–344, 2005.

[145] Stephen Tomlinson, Douglas W. Oard, Jason R. Baron, and Paul Thompson. Overview of the TREC 2007 Legal Track. In *Proceedings of the Sixteenth Text REtrieval Conference (TREC 2006)*, 2007.

[146] Andrew Trotman and Dylan Jenkinson. IR evaluation using multiple assessors per topic. *Proceedings of the 12th Australasian Document Computing Sumposium*, 2007.

[147] Andrew Turpin, Falk Scholer, Stefano Mizzaro, and Eddy Maddalena. The benefits of magnitude estimation relevance assessments for information retrieval evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 565–574, 2015.

[148] David van Dijk, Zhaochun Ren, Evangelos Kanoulas, and Maarten de Rijke. The University of Amsterdam (ILPS.UvA) at TREC 2015 Total Recall Track. In *Proceedings of the Twenty-Fourth Text REtrieval Conference (TREC 2015)*, 2015.

[149] Jyothi K. Vinjumur, Douglas W. Oard, and Jiaul H. Paik. Assessing the reliability and reusability of an e-discovery privilege test collection. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 1047–1050, 2014.

[150] Ellen Voorhees and Donna Harman. Overview of the Sixth Text REtrieval Conference (TREC-4). In *Proceedings of the Sixth Text REtrieval Conference (TREC-6)* , 1997.

[151] Ellen Voorhees and Donna Harman. Overview of the Eighth Text REtrieval Conference (TREC-8). In *Proceedings of the Eigthth Text REtrieval Conference (TREC-8)* , 1999.

[152] Ellen M Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36(5):697–716, 2000.

[153] Ellen M. Voorhees. The Philosophy of Information Retrieval Evaluation. In *Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370, 2002.

[154] Ellen M. Voorhees and Donna K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[155] Jianqiang Wang and Dagobert Soergel. A user study of relevance judgments for e-discovery. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–10, 2010.

[156] Jun Wang and Jianhan Zhu. Portfolio theory of information retrieval. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 115–122, 2009.

[157] William Webber. Re-examining the effectiveness of manual review. In *Proceedings of the SIGIR Information Retrieval for E-Discovery Workshop*, page 2, 2011.

[158] William Webber. Approximate recall confidence intervals. *ACM Transactions on Information Systems*, 31(1):1–33, 2013.

[159] William Webber. Random vs active selection of training examples in e-discovery. `http://blog.codalism.com/index.php/random-vs-active-selection-of-training-examples-in-e-discovery/`, 2014.

[160] William Webber, Mossaab Bagdouri, David D. Lewis, and Douglas W. Oard. Sequential testing in classifier evaluation yields biased estimates of effectiveness. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 933–936, 2013.

[161] William Webber, Praveen Chandar, and Ben Carterette. Alternative assessor disagreement and retrieval depth. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 125–134, 2012.

[162] William Webber, Douglas W. Oard, Falk Scholer, and Bruce Hedin. Assessor error in stratified evaluation. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 539–548, 2010.

[163] William Webber and Jeremy Pickens. Assessor disagreement and text classifier accuracy. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 929–932, 2013.

[164] William Webber and Jeremy Pickens. Assessor disagreement and text classifier accuracy. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 929–932, 2013.

[165] William Webber, Bryan Toth, and Marjorie Desamito. Effect of written instructions on assessor agreement. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 1053–1054, 2012.

[166] Chuan Wu, Wei Lu, and Ruixue Wang. WHU at TREC Total Recall Track 2015. In *Proceedings of the Twenty-Fourth Text REtrieval Conference (TREC 2015)*, 2015.

[167] Cheng Xiang Zhai, William W. Cohen, and John Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 10–17, 2003.

[168] Haotian Zhang, Wu Lin, Yipeng Wang, Charles L. A. Clarke, and Mark D. Smucker. WaterlooClarke: TREC 2015 Total Recall Track. In *Proceedings of the Twenty-Fourth Text REtrieval Conference (TREC 2015)*, 2015.

[169] Xingquan Zhu and Xindong Wu. Class noise vs. attribute noise: A quantitative study of their impacts. *Artificial Intelligence Review*, 22(3):177–210, 2004.

[170] Justin Zobel, Alistair Moffat, and Laurence A. F. Park. Against recall: Is it persistence, cardinality, density, coverage, or totality? *SIGIR Forum*, 43(1):3–8, 2009.

APPENDICES

Appendix A

Total Recall Per-Topic Gain Curves

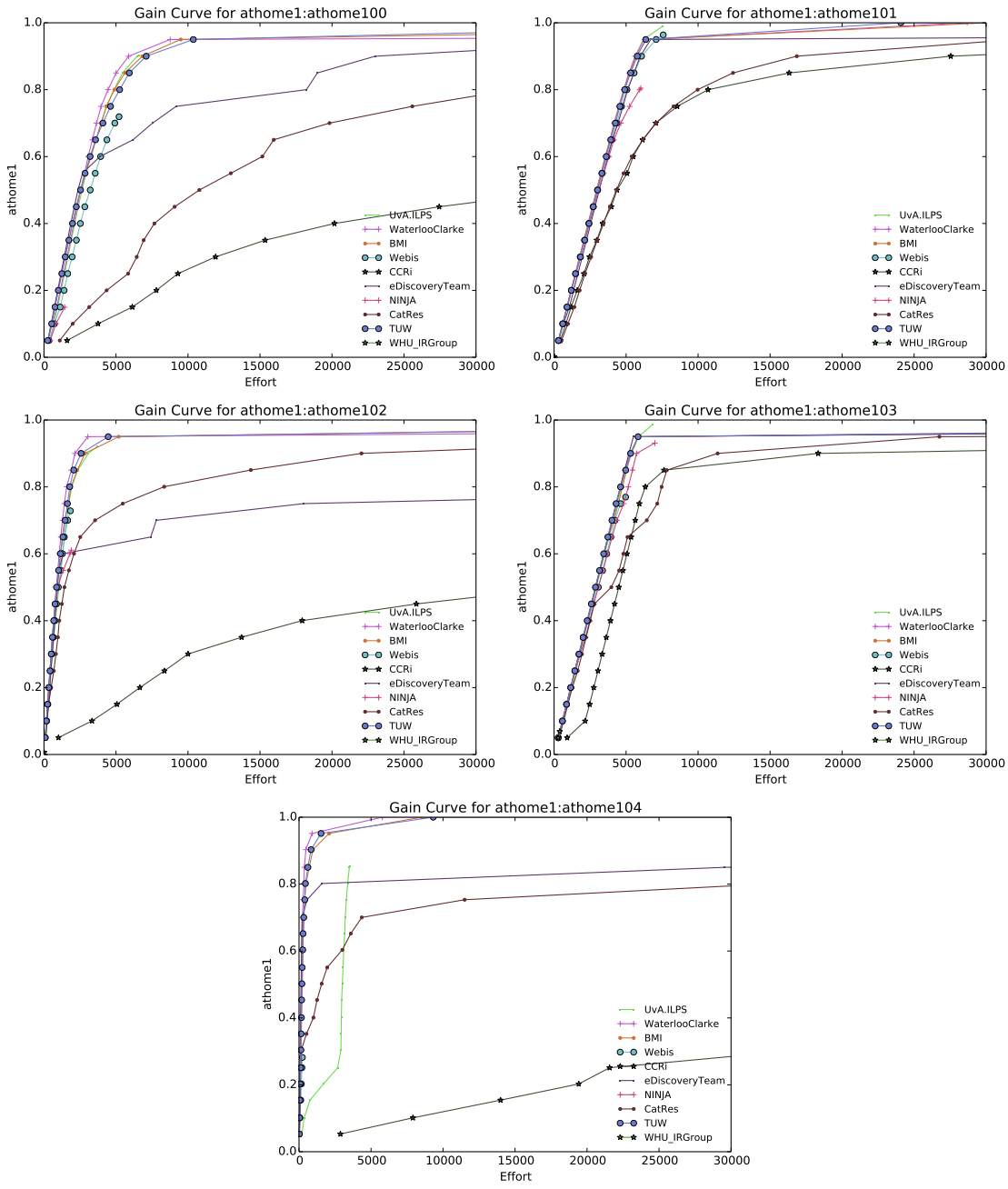


Figure A.1: Per-topic gain curves for Total Recall 2015 submissions.

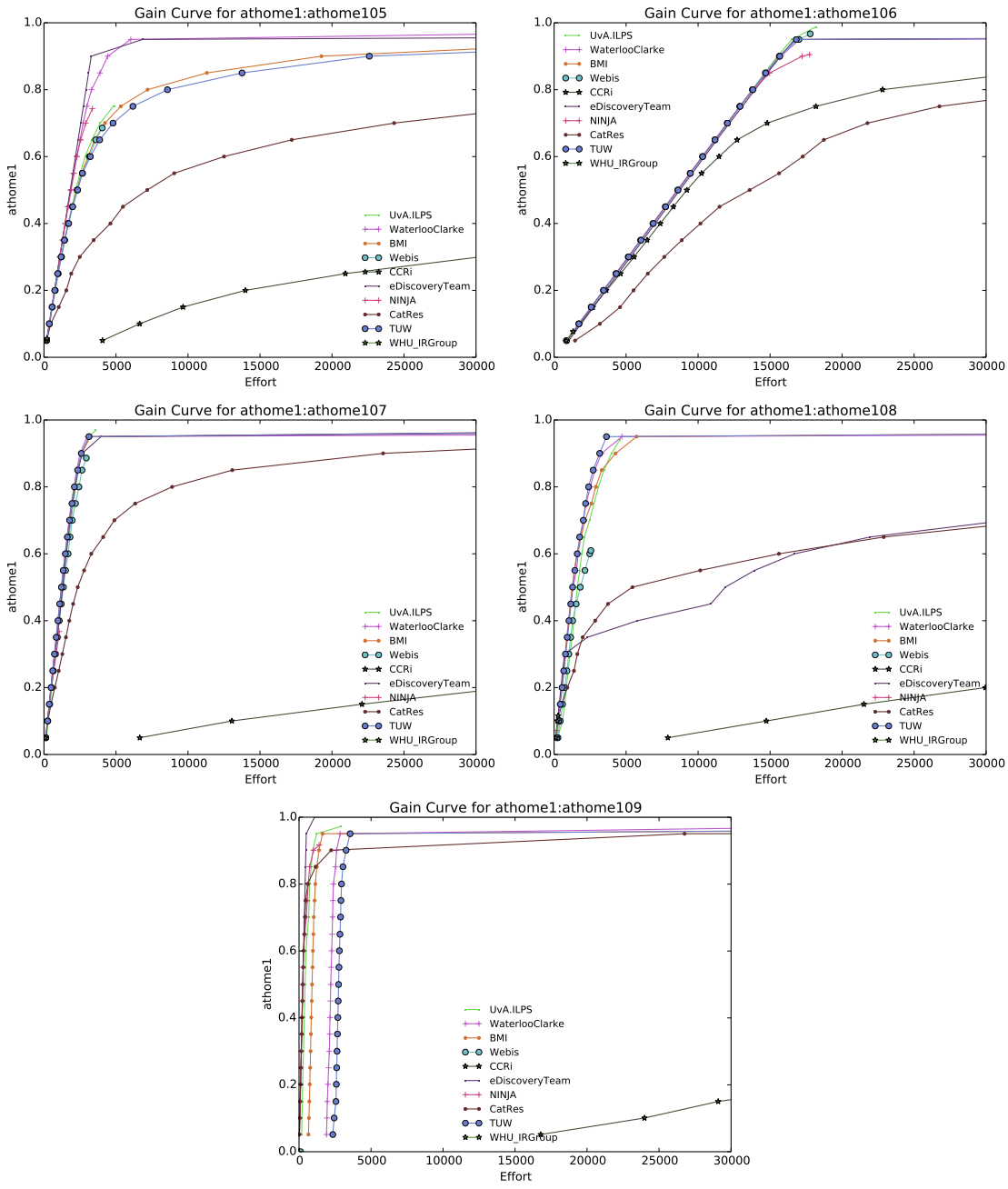


Figure A.2: Per-topic gain curves for Total Recall 2015 submissions.

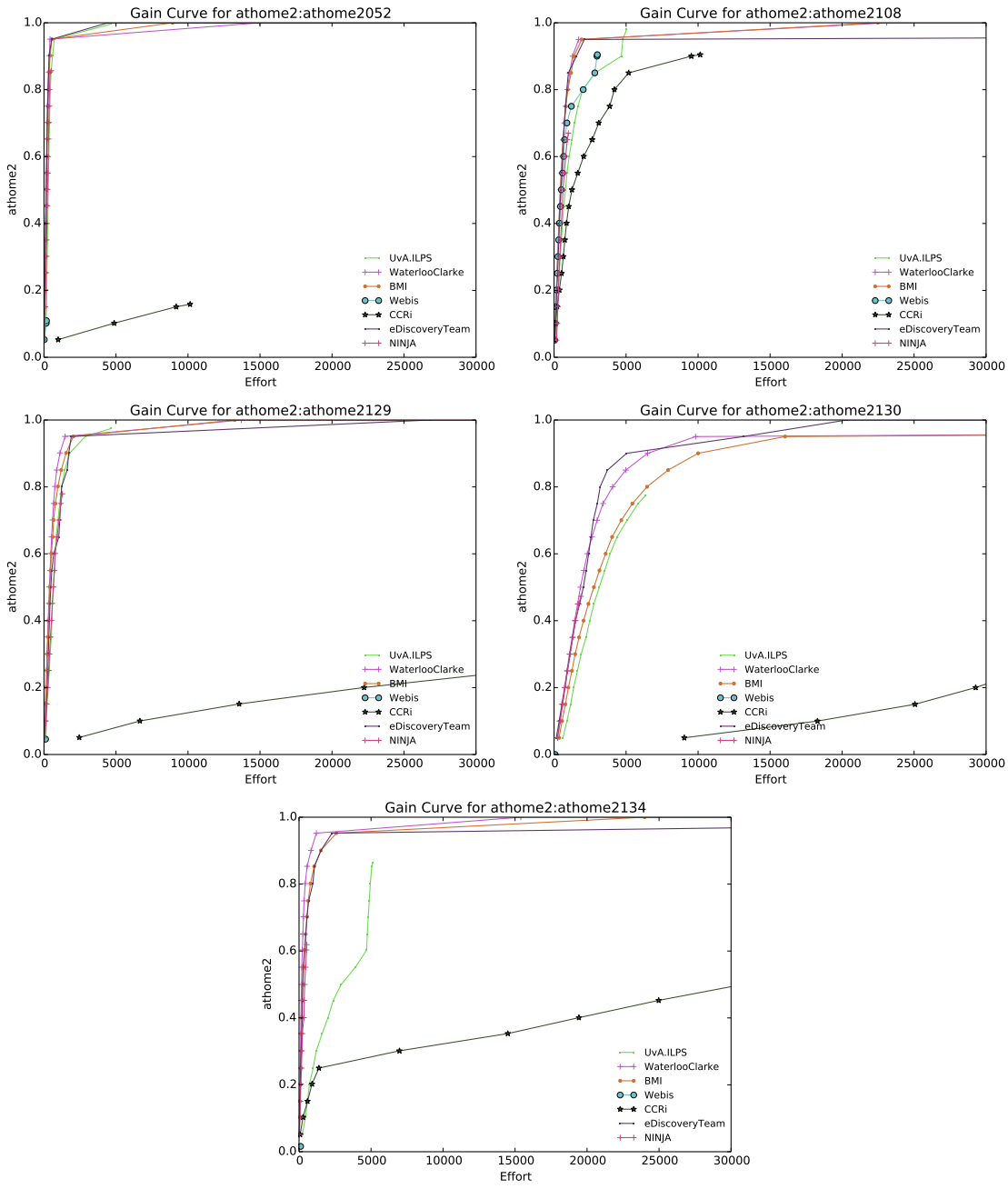


Figure A.3: Per-topic gain curves for Total Recall 2015 submissions.

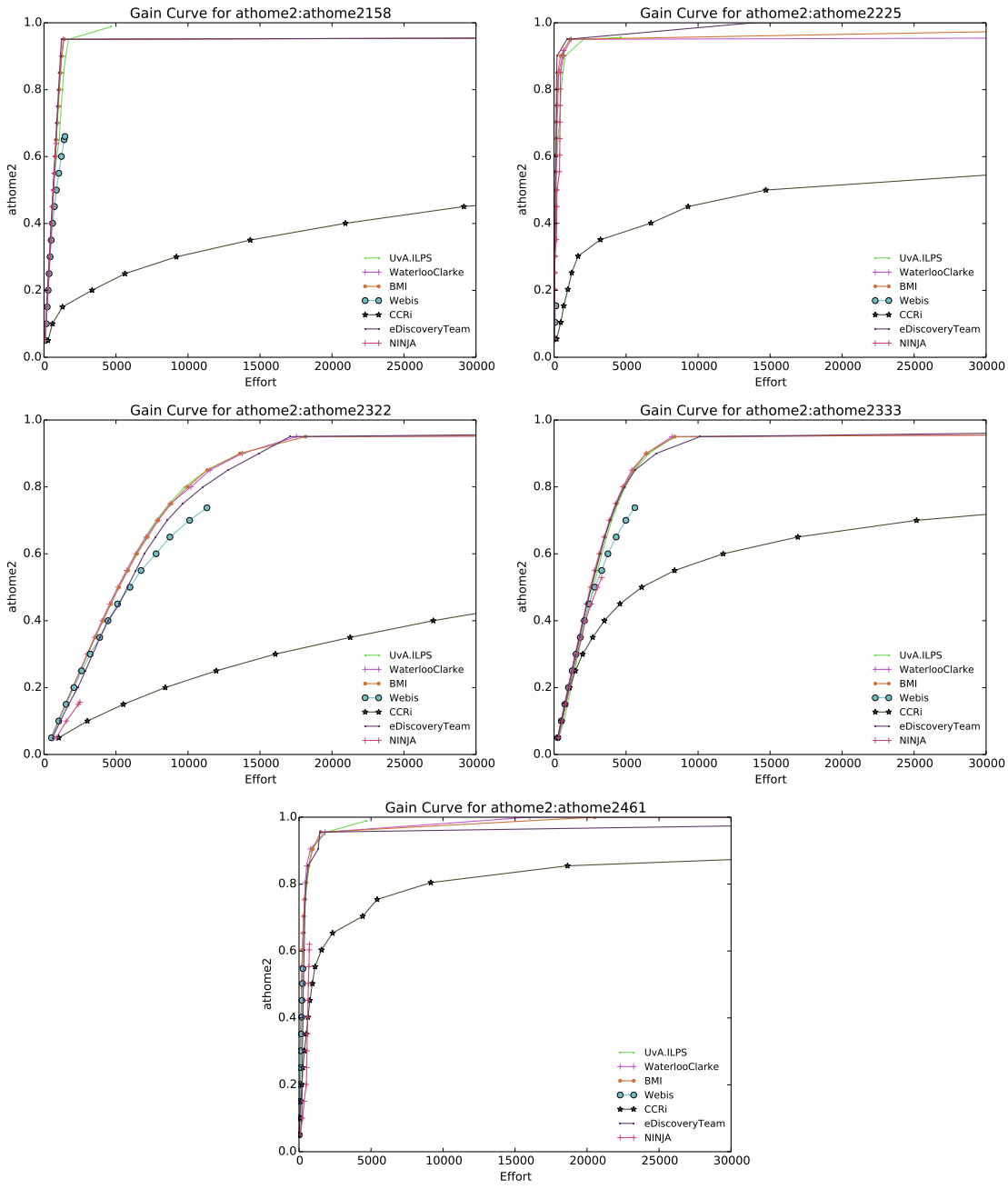


Figure A.4: Per-topic gain curves for Total Recall 2015 submissions.

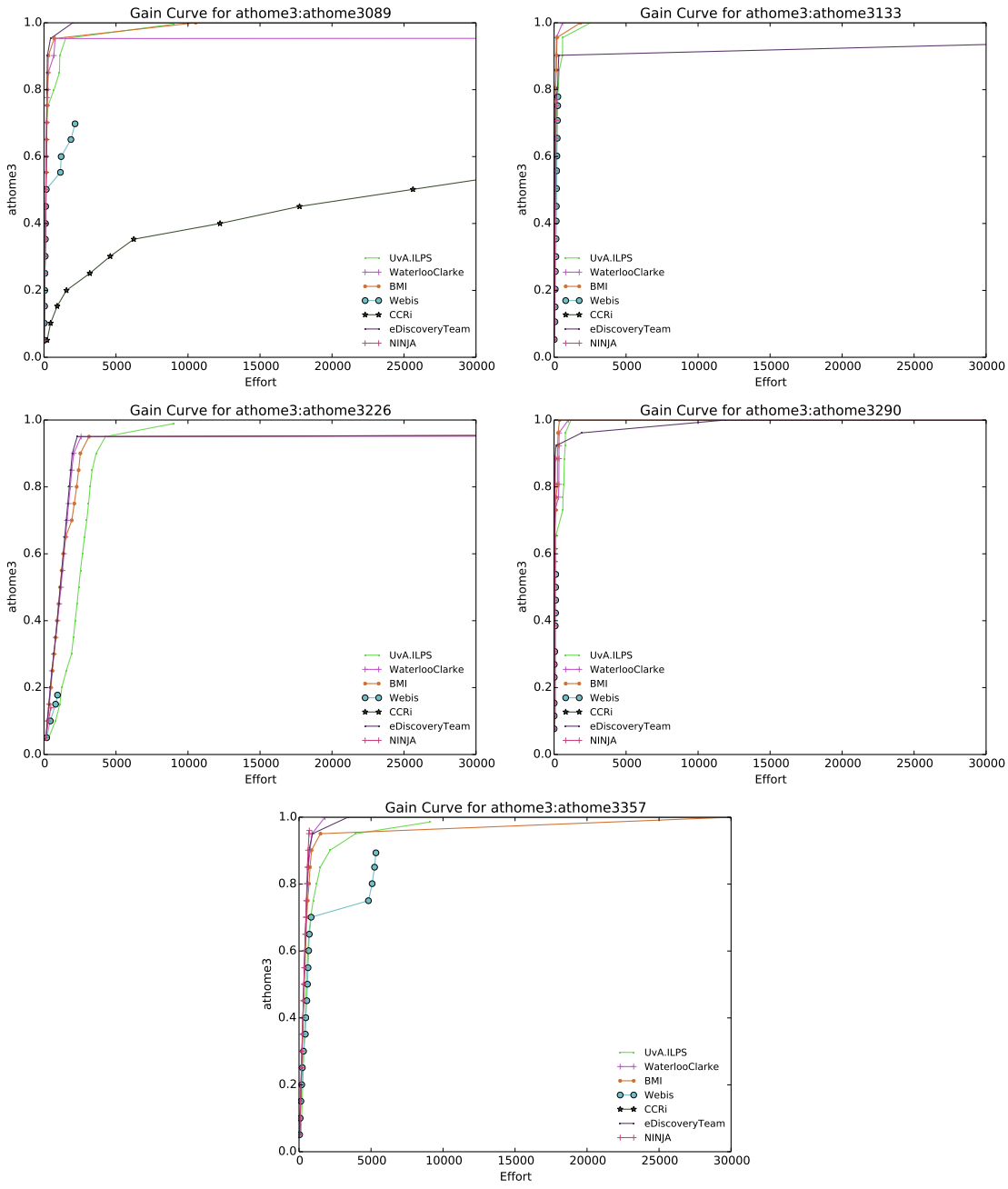


Figure A.5: Per-topic gain curves for Total Recall 2015 submissions.

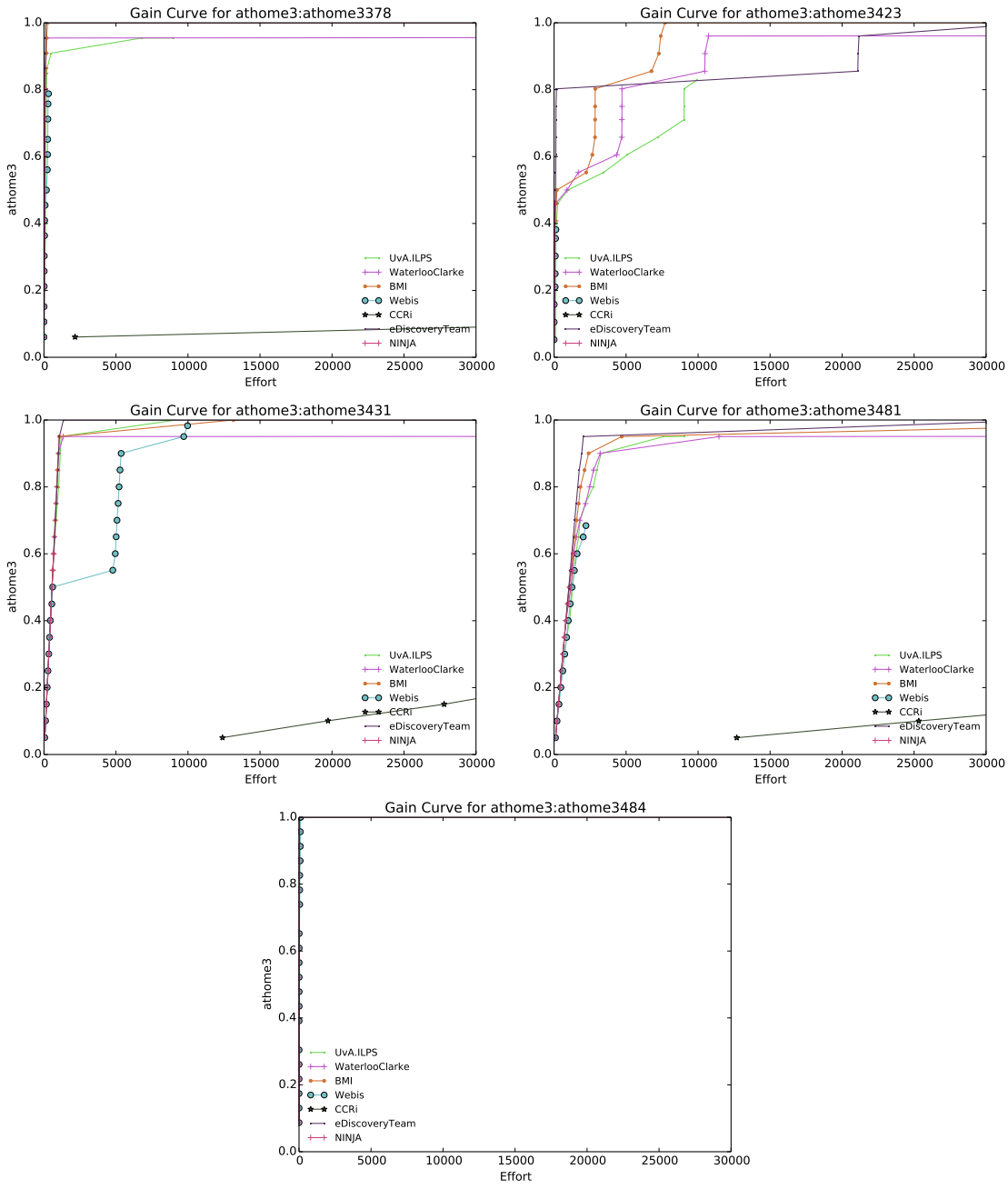


Figure A.6: Per-topic gain curves for Total Recall 2015 submissions.

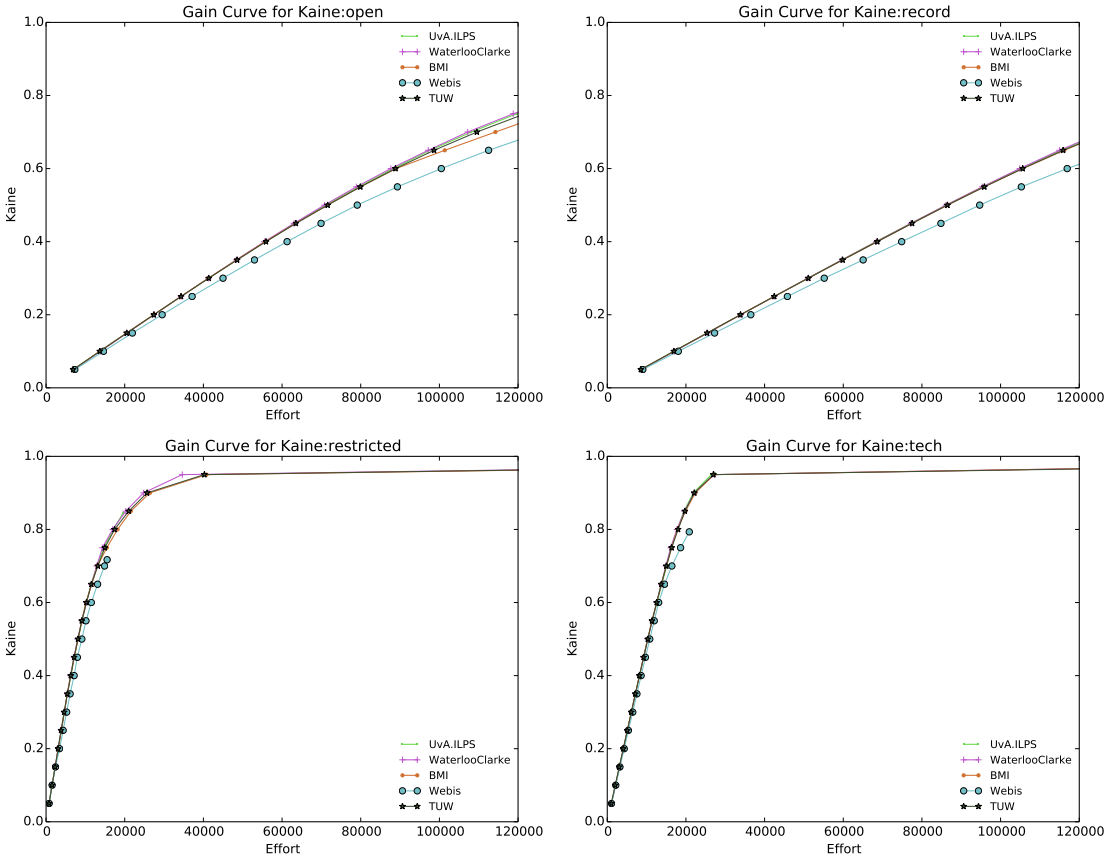


Figure A.7: Per-topic gain curves for Total Recall 2015 submissions.

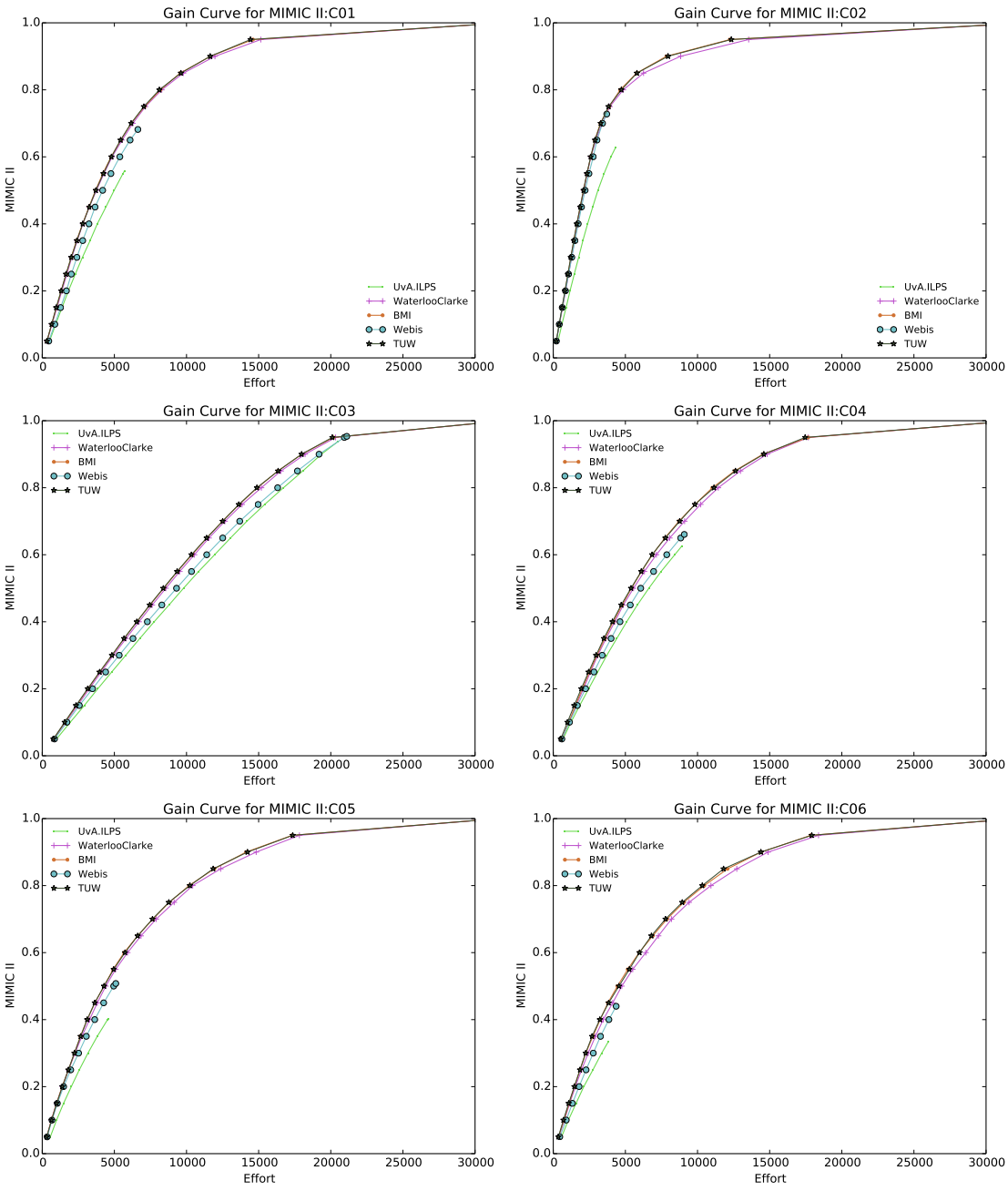


Figure A.8: Per-topic gain curves for Total Recall 2015 submissions.

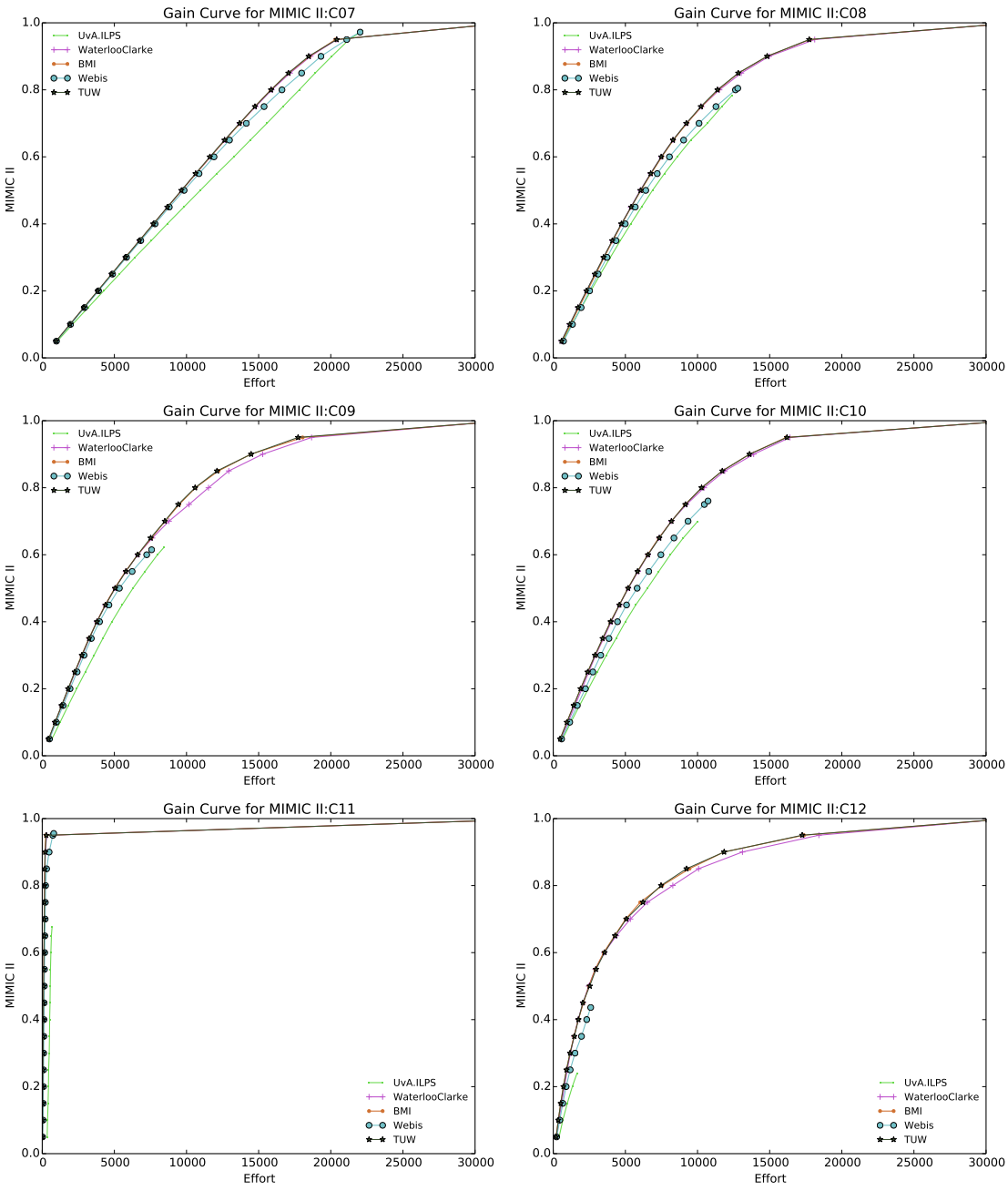


Figure A.9: Per-topic gain curves for Total Recall 2015 submissions.

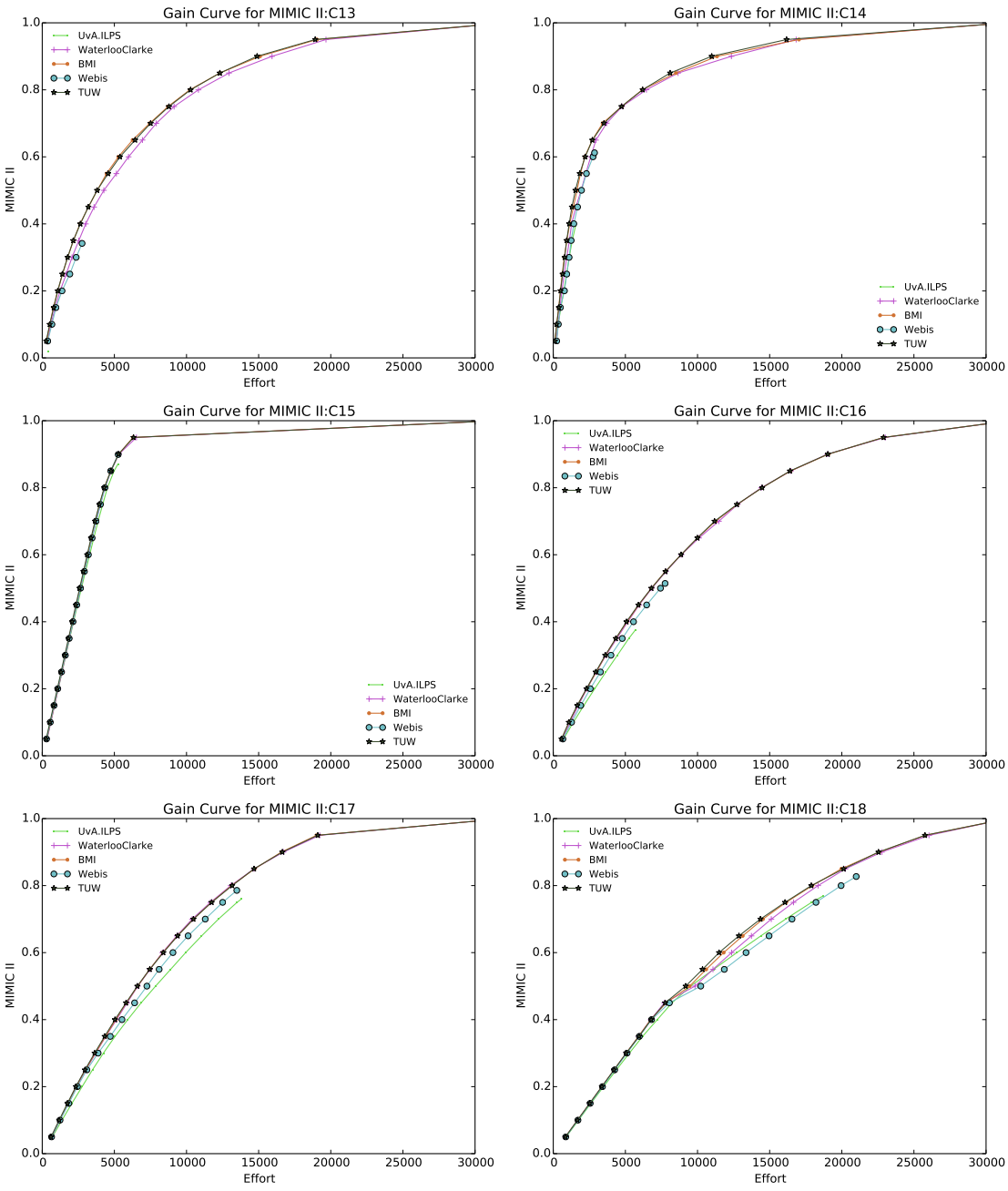


Figure A.10: Per-topic gain curves for Total Recall 2015 submissions.

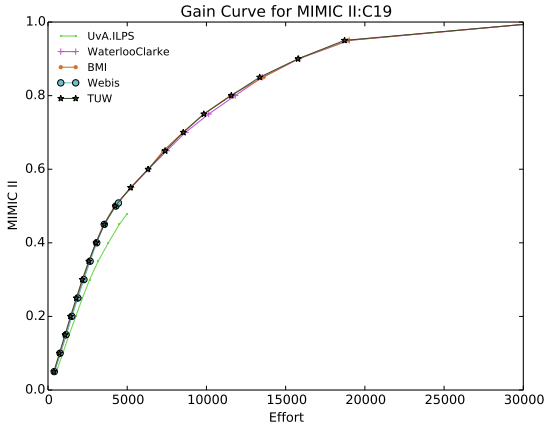


Figure A.11: Per-topic gain curves for Total Recall 2015 submissions.

Glossary

This section provides a brief glossary of terms used throughout this thesis. Entries marked with †are adapted from "The Grossman-Cormack Glossary of Technology-Assisted Review" [39], which provides a more comprehensive and general glossary of terms.

Active learning† is an iterative process where the training set for a machine learning algorithm is repeatedly expanded with documents selected by the algorithm or a human assessor.

Area under the ROC curve (AUC)† is, as the name states, the area under the ROC curve. It can be thought of as the probability that a relevant document would be ranked before a non-relevant document in the ranked list. AUC can be computed using the formula: $\frac{X}{R \times N}$, where X is the number of times a relevant document is ranked above a non-relevant document in the ranked list and N is the number of non-relevant documents. Note that the computation of AUC requires a complete ranking of the collection.

Assessor is simply a human who assigns a relevance judgment to a document with respect to an information need.

Average precision (AP) is, for a single topic, the result of averaging each precision@k value, calculated using the rank of each relevant document (in the ranked list) as the cutoff k. Average precision for a ranked list is calculated as:

$$\frac{1}{\mathrm{R}} \cdot \sum_{i=1}^{k} rel(i) \cdot \text{Precision@}i$$

where $rel(i) = 1$ if the document at is relevant and 0 otherwise, and k is the length of the ranked list. AP is an approximation of the area under the uninterpolated precision-recall curve.

Cross validation is the process whereby a collection is divided into n partitions and each partition is used in turn as a test set for machine learning trained on the other $n-1$ partitions. The end goal is to provide a more accurate estimate of classifier performance when a test set does not exist or is too costly to create. A particular version of cross validation is used in Chapter 3.

Enron refers to the Enron email collection that was collected at the request of the Federal Energy Regulatory Committee and of which several versions have been released. More details can be found in Section 2.5.1.

F1 is the harmonic mean of Recall and Precision: $2 \times \frac{precision \times recall}{precision + recall}$.

F1@k is the F1 score when only the first k elements of the ranked list are used in the computation.

Facet is an arbitrary split of the document collection based upon some identifiable aspect. In traditional IR, subtopics would be considered facets under this definition. Our use of facet is more general and applies to any identifiable property (e.g., file type, sender or recipient of email, etc.).

False positive rate is the fraction of non-relevant documents that were incorrectly identified as relevant.

Features† are the aspects of a document that are used by a Machine Learning Algorithm to classify the document. For example, features can include single words or phrases as well as metadata such as sender or recipient of an email.

Feature engineering† is the process of selecting features for machine learning algorithms. This may itself involve some computation in an attempt to find the "best" features for a particular task.

Gain curve measures recall@k as a function of varying the cutoff, k, in a ranked list. Effectively, it reports the review effort required to achieve a particular level of recall.

Gold standard/ground truth is the final arbiter of truth about the relevance of documents with respect to an information. In electronic discovery, this is often the senior litigator in charge of a particular case.

Hypothetical F1 is the F1@k score such that k is chosen to maximize the F1 score. This is a post hoc evaluation measure and k is usually determined through a sweep of all possible cutoffs. Sometimes called max(imum) F1.

Information need is the information that a searcher desires to find (e.g., "Batman's main villain" or "the answer to life, the universe, and everything").

Interpolated precision at recall point X is the maximum precision attained for any recall point $\geq$ X.

Information governance is a set of processes and procedures that dictates which documents, both electronic and hard copy, should be maintained to meet he regulations that govern a particular organization (e.g., legal, environmental, regulatory).

Judgmental sample† is a sample of the document collection that was drawn using subjective factors, often to find the "most interesting" or "most useful" document or documents believed to have been missed. Note that a judgmental sample is not a statistical sample, and properties from this sample cannot be extrapolated to the document collection at large.

Logistic regression† is a supervised learning algorithm that uses the features of a document to estimate its probability of relevance.

Machine learning† is the use of a computer algorithm to organize or classify documents based upon analysis of their features.

Manual review† is the process of having humans manually review documents in a collection for relevance to a particular topic.

Mean average precision (MAP) is simply AP averaged over multiple topics.

Passive learning is the process of training a machine learning algorithm such that the algorithm has no input into the training documents selected solely, by the human trainer or other outside source.

Precision is the number of relevant documents retrieved divided by the number of documents retrieved.

Precision@k is the precision achieved when only the first k elements of a ranked list are used in the computation.

Precision-Recall curve is the plot of all possible precision@k and recall@k values attainable at possible values of k in a ranked list. An *interpolated* precision-recall curve is created by measuring interpolated precision at various recall values (e.g., 0.1, 0.2, ..., 1.0).

R is the number of relevant documents for a particular information need in a particular document collection.

R-Precision is simply the precision@R in a ranked list. At this point, precision, recall, and F1 are all equal. It is sometimes called the precision-recall break-even point.

Random sampling† is the process of selecting a subset of the document collection such that every document has an equal chance to be in the sample. When used for the purposes of training a machine learning algorithm, this is called simple passive learning (SPL).

Ranked list is the result of performing relevance ranking. It is a list of documents ranked by decreasing likelihood of relevance to a particular information need.

Recall is the fraction of relevant documents retrieved in a ranked list divided by R.

Recall@k is the recall achieved when only the first k elements of a ranked list are used in the computation.

Recall depth is the shortest prefix of a ranked list that needs to be examined to achieve a desired level of recall. Sometimes called recall effort.

Receiver operating characteristic (ROC) curve† is the curve resulting from plotting all true positive and false positive rates from all possible cutoffs in a ranked list.

Relevance/relevant† A document is considered relevant if it meets a particular information need. This is typically determined by a human.

Relevance ranking† is the process of assigning a score to a subset of the documents in a document collection to indicate the likelihood of relevance to a particular information need. The result is a list of documents ranked, in most cases, from most likely to least likely to be relevant. This list is called a ranked list.

Relevance sampling is an active learning technique where documents that the classifier believes are most likely to be relevant are added to the training set. This is sometimes called continuous active learning (CAL).

(Review) effort is the number of documents that must be reviewed for relevance assessment. Depending on the context this may include any training assessments (for machine learning) as well as the effort of reviewing a resulting ranked list for evaluation.

Reuters Collection Volume 1 (RCV1) is a document collection consisting of newswire articles collected and exhaustively labelled by Reuters with respect to a variety of relevance classes, including topic, country, and language. More precise details on the creation and use of this collection have been documented by Lewis et al. [96].

Surrogate is an assessor who is not the same as the assessor being used for evaluation (i.e., the gold standard or authority) in an experiment.

Supervised learning† is a machine learning process in which a machine learning algorithm is trained to determine the differences between relevant and non-relevant documents using labeled exemplar documents, called a training set.

Support vector machine (SVM)† is a supervised learning algorithm that maps documents in a hyperspace (i.e., features become points in a multi-dimensional space). The algorithm uses geometric methods to find a hyperplane that separates the relevant and non-relevant training examples. Documents outside of the training set are classified according to which side of the hyperplane they fall upon.

Topic is TREC jargon for an information need.

(Topic) authority is the term applied to the volunteer senior lawyer who acted as the gold standard for the TREC Legal track. In this thesis, we use the term more generally to refer to the assessments being used for evaluation even if they were not originally created as a gold standard.

True positive rate is the fraction of relevant documents that were correctly identified as relevant.

Uncertainty sampling† is an active learning technique where documents that the machine learning algorithm is most uncertain about are assessed and added to the training set. In terms of logistic regression, this would refer to documents that have a corresponding probability closest to 0.5. For SVM algorithms, these documents would be closest to the hyperplane. Sometimes called simple active learning (SAL).