UNIVERSIDADE DA CORUÑA

New Algorithms and Methodologies for Building Information Retrieval Collections

Doctoral Thesis

David Otero Freijeiro

2024

David Otero: *New Algorithms and Methodologies for Building Information Retrieval Collections*, Doctoral Thesis, Universidade da Coruña, 2024.

New Algorithms and Methodologies for Building Information Retrieval Collections

David Otero Freijeiro

Doctoral Thesis / 2024

Advisors:

Álvaro Barreiro García

Javier Parapar López

PhD in Computer Science

Álvaro Barreiro García, Professor at the Department of Computer Science and Information Technologies of Universidade da Coruña,

and

Javier Parapar López, Associate Professor at the Department of Computer Science and Information Technologies of Universidade da Coruña,

HEREBY CERTIFY

that the present Doctoral Thesis, ***New Algorithms and Methodologies for Building Information Retrieval Collections***, submitted to the Universidade da Coruña by David Otero Freijeiro, has been carried out under our supervision and fulfils all the requirements for the award of the degree of *PhD in Computer Science with International Mention*.

Álvaro Barreiro García
Advisor

Javier Parapar López
Advisor

To Nuria,
you are my island in the stream.

Sólo el que se entrega magnánimamente a la tarea de la verdad encontrará algo. Confiad en vosotros mismos, la potencia del espíritu es inconmensurable.

— Antonio Escohotado

ACKNOWLEDGMENTS

A PhD thesis is the result of an intellectual maturing process. This process, delightfully satisfying in the end, would not have been possible without the extraordinary guidance of my advisors. I sincerely thank Álvaro Barreiro for his valuable supervision and guidance during all these years. I also wish to thank all the times Javier Parapar was by my side, being the most dedicated advisor one can ask for. I have learnt countless things from you. From the bottom of my heart, thank you.

I also want to thank all the support of the IRLab members, whom I now proudly call my friends. Alfonso, Eliseo, Manu, Jorge, Gilberto, Brais, Juan, and many others, I will never forget all the time we have shared during all these years. I owe a special mention to Anxo. Honestly, I do not know how to express with words what you mean to me. I hope to have you by my side many more years.

I wish to express my gratitude to the members of the Information Management Systems Research Group at the University of Padova. Thank you, Guglielmo, Stefano, Laura, Ornella, Fabio, Dennis, Alessandro and the rest of the group. I keep fond memories of my time there. I want to especially thank Nicola for giving me the opportunity of working with him.

If anyone deserves a mention, it is my family. I do not have words to express the love that I profess you. Without the support and incondictional love of my parents, my grandparents, and my sisters, I would not be writing these lines. And Nuria, cause even when there were no stars in sight, you were always my only guiding light.

Last, I would like to acknowledge the efforts of the PhD defence committee and the external reviewers. I also want to recognise the input of those anonymous reviewers who evaluated all my papers. I also acknowledge the financial support of CITIC, *Xunta de Galicia* and the Government of Spain for my PhD grant, the travel grants and the funding through research projects.

ABSTRACT

Information retrieval systems play a crucial role in addressing users' information needs by aiding their exploration of vast collections of information. This thesis is framed in a critical information retrieval research aspect: evaluation. In particular, we propose new approaches for creating annotated test collections. Such collections are essential for evaluating retrieval systems' effectiveness in controlled experiments. Reflecting real-world conditions accurately in these test collections is pivotal for progress in the field.

We aim to introduce innovative techniques for efficiently assembling reliable test collections, facilitating broader research and development in information retrieval. The thesis first proposes a new method for building new pooled test collections without requiring costly evaluation campaigns. This approach simplifies and economizes the process of building new benchmarks. Then, we introduce a novel adjudication method for determining which pooled documents warrant human judgment, aiming to reduce the need for extensive expert assessments. This method is both cost-effective and efficient. Additionally, the thesis presents a fresh perspective on evaluating adjudication methods, emphasizing statistical significance, an aspect often overlooked in previous document adjudication research. As a demonstration of the methods explored in this thesis, we applied them to develop a new test collection whose construction process we describe here as an example of the use of reduced-budget methods.

In summary, this thesis integrates established information retrieval knowledge with new methodologies to create annotated collections that are both cost-effective and reliable. This fusion is crucial for advancing the development of more effective retrieval systems.

RESUMEN

Los sistemas de recuperación de información desempeñan un papel crucial a la hora de satisfacer las necesidades de información de los usuarios, ayudándoles a explorar vastas colecciones de información. Esta tesis se enmarca en un aspecto crítico de la investigación en recuperación de información: la evaluación. En concreto, proponemos nuevos enfoques para crear colecciones de prueba. Éstas son esenciales para evaluar la eficacia de los sistemas de recuperación en experimentos controlados. Reflejar con precisión las condiciones del mundo real en estas colecciones es fundamental para avanzar en este campo.

Nuestro objetivo es introducir técnicas innovadoras para construir colecciones anotadas que sean fiables, y facilitar así la investigación y el desarrollo en el campo de la recuperación de información. En primer lugar, la tesis propone un nuevo método para crear nuevas colecciones de prueba sin necesidad de costosas campañas de evaluación, simplificando y economizando el proceso. A continuación, presentamos un nuevo método de adjudicación para determinar qué documentos merecen un juicio humano, con el objetivo de reducir el numero de juicios expertos necesarios. Este método es rentable y eficiente. Además, la tesis presenta una nueva perspectiva de la evaluación de los métodos de adjudicación, haciendo hincapié en la significancia estadística, un aspecto que a menudo se pasa por alto en anteriores investigaciones sobre adjudicación de documentos. Finalmente, aplicamos los métodos explorados en esta tesis para construir una nueva colección de prueba, cuyo proceso de construcción describimos, para demostrar la utilidad de nuestras propuestas.

En resumen, esta tesis integra conocimiento establecido en el campo con nuevas metodologías para así crear nuevas colecciones de prueba fiables y con bajo coste. Esta combinación es crucial para avanzar en el desarrollo de sistemas de recuperación de información más efectivos.

RESUMO

Os sistemas de recuperación de información desempeñan un papel crucial á hora de satisfacer as necesidades de información dos usuarios, axudándolles a explorar vastas coleccións de información. Esta tese enmárcase nun aspecto crítico da investigación en recuperación de información: a avaliación. En concreto, propoñemos novos enfoques para crear coleccións de proba. Estas son esenciais para avaliar a eficacia dos sistemas de recuperación en experimentos controlados. Reflectir con precisión as condicións do mundo real nestas coleccións é fundamental para avanzar neste campo.

O noso obxectivo é introducir técnicas innovadoras para construír coleccións anotadas que sexan fiables, e facilitar así a investigación e o desenvolvemento no campo da recuperación de información. En primeiro lugar, a tese propón un novo método para crear novas coleccións de proba sen necesidade de custosas campañas de avaliación, simplificando e economizando o proceso. A continuación, presentamos un novo método de adxudicación para determinar que documentos merecen un xuízo humano, co obxectivo de reducir o numero de xuízos expertos necesarios. Este método é rentable e eficiente. Ademais, a tese presenta unha nova perspectiva da avaliación dos métodos de adxudicación, facendo fincapé na significancia estatística, un aspecto que a miúdo se pasa por alto en anteriores investigacións sobre adxudicación de documentos. Finalmente, aplicamos os métodos explorados nesta tese para consruir unha nova colección de proba, cuxo proceso de construción describimos, para demostrar a utilidade das nosas propostas.

En resumo, esta tese integra coñecemento establecido no campo con novas metodoloxías para así crear novas coleccións de proba fiables e con baixo custo. Esta combinación é crucial para avanzar no desenvolvemento de sistemas de recuperación de información máis efectivos

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACRONYMS

AA	Active Agreements
AD	Active Disagreements
AP	Average Precision
CAS	Computational Archival Science
CDS	Clinical Decision Support
DL21	Deep Learning 21
DMM	Divergence Minimization Model
FWER	Family-Wise Error Rate
HSD	Honestly Significant Difference
IDF	Inverse Document Frequency
IR	Information Retrieval
ISJ	Interactive Searching and Judging
KLD	Kullback-Leibler Divergence
LOGO	Leave-One-Group-Out
MA	Mixed Agreements
MAP	Mean Average Precision
MD	Mixed Disagreements
MEDMM	Maximum-Entropy Divergence Minimization Model
MLE	Maximum Likelihood Estimate
MM	MaxMean
MMNS	MaxMean Non-Stationary
MTF	MoveToFront
NDCG	Normalized Discounted Cumulative Gain
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NTCIR	NII Testbeds and Community for Information Access Research
PRF	Pseudo-Relevance Feedback

RF	Relevance Feedback
RM	Relevance Model
RM1	Relevance Model 1
RM3	Relevance Model 3
TREC	Text REtrieval Conference
TS	Thompson Sampling
TSNS	Thompson Sampling Non-Stationary

Part I

PROLOGUE

1

INTRODUCTION

Search engines are essential tools for exploring vast collections of information. Locating relevant information in these collections is nearly impossible without them. Users rely on search engines to sift through collections of documents, including emails, online commerce products, streaming videos, web pages, and news. In a more broad sense, we can say that search engines support human cognition with documents. Underlying every search engine, a ranking system is responsible for identifying the most pertinent documents based on the user's information needs. Information Retrieval (IR) is a field of computer science dedicated to developing systems that assist users in finding information tailored to their needs. The importance of these ranking systems has grown tremendously, especially as they become integral parts in areas like e-commerce, recommender systems, and social media. The ranking quality profoundly influences the user experience, making the system's effectiveness vital to user satisfaction. Consequently, the performance of these systems is paramount for both users and providers, leading to extensive efforts in their evaluation and improvement.

Traditionally, research and development in this field has relied on human-labelled data in the form of expert annotations: experts assess the relevance of documents to specific information needs, creating annotated datasets that serve to train and evaluate systems. Constructing new specific collections for the always-increasing number of different tasks related to IR is very expensive due to this needed human effort. This cost hinders the creation of new benchmarks and, thus, the broad research and development of new ideas in the field.

This thesis introduces novel methods for efficiently and reliably constructing annotated collections for information retrieval.

1.1 MOTIVATION

Information retrieval has a long and rich tradition of experimentation. This tradition dates back to the 1950s and 1960s, marked by Cyril Cleverdon's pioneering work. Before his experiments, debates on different approaches to IR were largely anecdotal and philosophical. Cleverdon was the first to perform an empirical, formally scientific experiment to compare different indexing schemes. Cleverdon's experiments, controversial at the time, were the first to use *test collections*: a common set of documents, a common set of information needs, and a common set of assessments. This approach allowed for more controlled experiments and better comparisons, reducing the variability that was pervasive in previous similar attempts. This methodology, now known as the Cranfield paradigm (named after the UK village where Cleverdon worked), set a precedent for IR evaluation. The use of test collections to evaluate the effectiveness of retrieval systems has been the *de facto* standard since then.

As IR evolves, document collections expand and new tasks emerge, systems face fresh challenges. Test collections must adapt to mirror the environments of operational systems. Creating larger and representative test collections is crucial for accurately evaluating systems in real-world conditions, but it is also a very costly process. Advancing retrieval system development becomes increasingly difficult without adequate tools and methodologies to develop new experimental test collections.

1.2 AIM AND SCOPE

This thesis focuses on developing innovative methods for constructing new collections for IR. We aim to create high-quality and cost-effective benchmarks, addressing a crucial need in the field. Building specific collections for various IR tasks is resource-intensive, primarily due to the extensive human effort required for relevance assessments. This requirement makes the process expensive. We aim to introduce techniques that reduce these costs, thereby supporting IR's broader development and research.

Our approach combines established IR knowledge with novel methods for selecting documents from a corpus for expert assessment. Additionally, we explore recent methodologies for evaluating the methods for building collections and propose a new approach. Beyond proving our methods' effectiveness through empirical, laboratory-based experiments, we emphasize their real-world applicability. Toward the thesis's conclusion, we demonstrate how our contributions can be applied to construct a practical, real-world collection.

1.3 OVERVIEW

This manuscript is divided into five parts with seven chapters. The current chapter presents the introduction to this work. Chapter 2 introduces basic information retrieval concepts and related work. Chapters 3 to 6 contain the novel contributions of this thesis. We aimed at making these chapters as self-contained as possible, so that they are easy to understand only with the information presented in Chapter 2, to easy their readability. Finally, Chapter 7 wraps up by compelling the main findings of this thesis and proposing some ideas for future work. In more detail, the contents of each part and each chapter are:

PART I This first part includes Chapter 1, which presents this thesis, and Chapter 2, which introduces pertinent basic concepts and discusses relevant background work.

PART II In this part, we explore some novel methods for building new retrieval test collections in situations with scarce resources. On the one hand, in Chapter 3, we present a novel methodology for creating synthetic runs from which to pool documents to obtain new relevance judgements when no real participants are available. On the other hand, in Chapter 4, we explore the use of real relevance feedback to prioritize the pooled documents, with the aim of reducing the number of assessments needed to build a set of reusable judgements.

PART III We propose, in Chapter 5, a new way of evaluating adjudication methods for building new IR test collections. In particular, we argue that existing methods miss a part of the whole picture by looking only at how these methods are able to preserve the ranking of pooled submissions. To fill this void, we propose to focus on the preservation of the statistical significances between the evaluated systems.

PART IV In this part, which includes Chapter 6, we explore a novel application of the contributions presented in previous parts to build a new collection. By building upon the contributions presented in Parts II and III, we create a new collection that includes relevant social media content about the patrimonialization processes suffered by cultural heritage entities.

PART V In this last part, which includes Chapter 7, we summarize our main findings and contributions, and suggest some ideas for future work.

2

SETTING THE STAGE

In this chapter, we introduce basic and fundamental concepts of IR and its evaluation. We also focus on explaining how collections for IR evaluation are built. Since we already provide more advanced background content in each contribution chapter, this is intentionally left as a brief introduction. An advanced reader, with experience in the foundations of IR, may skip this chapter.

2.1 INFORMATION RETRIEVAL

Information retrieval is a computer science area aimed at developing systems to satisfy the information needs of its users. There have been many attempts to formally define what IR is. We believe that the one given by Gerard Salton in his textbook is still accurate nowadays: "Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information." (Gerard Salton 1968). Web search engines are probably one of the most known examples of a retrieval system. Without a search engine, it would be impossible to find something relevant within the vast collection of information that is the web. The inventions of the internet and, more importantly, of the World Wide Web, have fostered an exponential growth in the development of new retrieval systems.

IR deals with the development of many components that form a retrieval system: crawlers, indexes, user interfaces, query processors, among others. But, underlying every retrieval system, there is the core retrieval model, responsible for generating the list of ranked documents that is presented to the user. The effectiveness of this model directly affects the user experience: a model not able to deliver relevant information to the user can make the entire system useless. In this thesis, we focus on building new collections for

training and evaluating these retrieval models. Thus, we now explain the basis of how these systems work to produce a ranking of documents.

2.1.1 *Ad hoc Retrieval*

Ad hoc retrieval represents the most obvious search task: a user has an information need, sits down in front of a system, and interacts with it to fulfil this need by searching for relevant information in a given collection. Usually, the information need of the user is introduced in the form of a short textual description consisting of a few keywords called *query*. The job of the system is to match this query with every document using a retrieval model, and thus produce a ranked list of documents.

There are a lot of ways in which a model can match a query to a document to estimate its relevance. The research of new retrieval models is one of the most prominent lines in IR research since its beginnings, and is still nowadays. It is not our intention to make an extensive review of these models here, but some of the most prominent models have been the boolean model, the vector space model (G. Salton et al. 1975), probabilistic models such as those based on language modelling (Ponte and Croft 1998), BM25 (Robertson and Zaragoza 2009), and, more recently, neural models based on the transformer architecture (Lin et al. 2021). Although all these models work quite differently under the hood, at the end, the search abstraction is the same for all: producing a list of relevant document given an information need in the form of a query. Thus, the evaluation framework used to assess the effectiveness of these models is common to all.

2.2 INFORMATION RETRIEVAL EVALUATION

As we already said in the previous chapter, IR has a rich tradition of experimentation that began with the pioneering work of Cyril Cleverdon (Cleverdon 1962). His work was the first to use test collections for comparing different techniques—he was comparing different indexing schemes—under the same experimental conditions, reducing much of the variability of the experiment. This constitutes the roots of what nowadays we know as the Cranfield paradigm (Voorhees 2002, 2019).

2.2.1 *Cranfield and TREC*

This paradigm is applied as follows to evaluate retrieval systems. A test collection comprises a set of documents (i.e. retrieval units), a set of information

needs (also called *topics*), and a set of relevance assessments (also known as relevance judgements or *qrels*) that stablish which documents are relevant with respect to every information need. A practitioner or researcher wanting to evaluate the performance of a given retrieval system, runs this system against every topic in the collection, producing a ranked list of the documents of the collection for each topic (this is a *run* or a *submission*). The ranking of documents for each topic is a list by decreasing order of estimated relevance, representing the idea of the system of which is relevant to the specific topic. Then, it is possible to compute a given effectiveness metric of choice for each topic using the relevance assessments of the collection. The overall quality of the system is measured as the average metric score for all topics. Different systems produce runs for the exact same set of topics on the exact same corpus of documents, and thus systems can be compared using the average scores. Systems producing better scores are considered more effective systems than the others. This paradigm is an abstraction of a typical user search task, making a few, but very important assumptions: i) relevance can be approximated by topical similarity, meaning that all relevant documents are equally important; ii) relevance is independent of the user, meaning that a single set of assessments is valid for any user; and, iii) relevance assessments are complete, that is, all the relevant documents for a topic are known. Obviously, these assumptions are not always true, but abstracting the process of a search task allows us to greatly reduce the variability of an experiment.

In the early days of IR, the collections that were used by the community were small enough that allowed for *completeness*: every document got its judgement with respect to every topic (Harman 2011). However, these collections were small in comparison to what commercial systems faced in the real world. The problem is that relevance judgements are made by human experts. While this allows a great deal of assessments to be made, it is certainly impossible to manually review every document in a modern test collection. During the 1970s, Karen Spärck Jones and others argued for the need of an 'IDEAL' collection (Spärck Jones and Rijsbergen 1975). There was an interest in building a general-purpose large test collection that served to evaluate a wide range of different systems. The 'IDEAL' collection was never built, but they proposed the main ideas of the current methodology used to build new retrieval test collections: *pooling*. Pooling is nowadays the standard methodology for building new test collections, and works as follows. Given the rankings output by a range of different systems for a set of topics, that is, given a series of runs, the idea is to judge only the union of the top ranked documents of each one, for each topic. This union is called the *pool*. Since we are just judging the top-k documents of each ranking, this method is also known as top$k$ pooling.

Since each ranking will try to push relevant documents to the top, we can assume that most relevant documents will enter the pool, and that those that did not are not relevant. Although it is obvious that we are going to miss some relevant documents, extensive evaluation has shown that the evaluation of systems is equivalent to having copmlete judgements (Voorhees 2002; Zobel 1998),

Due to the need of crafting larger collections, the United States National Institute of Standards and Technology (NIST) founded the Text REtrieval Conference (TREC) with the aim of constructing a realistic genera-purpose collection for IR research. TREC has been running since then, and many new and different collections have been built through pooling. However, collections are again small compared with the real-world conditions that operational systems face. Pooling it is dependent on the size of the corpus of documents and, as this corpus grows, it is increasingly difficult to find most relevant documents. Many and very diverse attempts have tried new approaches to tackle these problems. Since the number of judgements is usually the limiting factor for building collections, a thread of work has proposed several approaches that look at how to efficiently allocate the budget of assessors. The main idea here is that, given a limited number of documents that assessors can judge per topic, how we can decide which documents merit those judgements. Methods that *actively* decide which methods merit human judgement are called adjudication methods.

2.2.2 *Adjudication Methods*

One of the first attempts in this direction is *MoveToFront* (MTF) (Cormack, Palmer et al. 1998). This method dynamically chooses the next document to assess based on the previous set of judged documents. The idea is to favour runs that have recently retrieved relevant documents, avoiding sampling documents from poor runs. In a posterior publication, Losada, Parapar et al. (2016) showed that MTF is just an instance of what is called multi-armed bandit-based methods. In the Reinforcement Learning field, multi-armed bandit methods have been studied for decades (Sutton and Barto 2018). In the most basic formalization, the multi-armed bandit problem consists of an algorithm that interacts with an environment during T rounds. At each round, the algorithm chooses one among a set of K different actions that it can perform, i.e. arms, and receives a reward for that arm. With this reward, the algorithm refines the policy that chooses the arm at each round, with the aim of maximizing the cumulated reward after T rounds. Therefore, the algorithm has to balance *exploration*, i.e. choosing different arms to

gather new information, with *exploitation*, choosing always the arms it knows that perform better. Losada, Parapar et al. (2016) applied this framework to the task of document adjudication. Instead of arms, we have rankings (for each topic). At each round, we can choose a ranking to pick its top-ranked document, and assess the relevance of this document. Using the relevance of the documents that each ranking provides, we can refine a policy to choose more documents from rankings that provided more relevant documents in the past, avoiding poor rankings. There different policies that we can devise to choose among rankings. In the original paper, they proposed four: MaxMean (MM), MaxMean Non Stationary (MMNS), Thompson Sampling (TS), and Thompson Sampling Non Stationary (TSNS). We refer the reader to the original paper (Losada, Parapar et al. 2016) for a mathematical formulation of how these methods works. As we said, all these methods prioritize the pooled documents with the aim of focusing the assessor's judgements on documents that have higher chances of being relevant. Indeed, experiments haven shown that most of relevant documents can be obtained with many fewer judgements than traditional top-k pooling.

2.2.3 *Evaluation of Adjudication Methods*

When evaluating retrieval systems, we are not usually interested in knowing the *absolute*[1] value that a metric may give to a system, but in a good estimate of the system performance, so we can *reliably* distinguish the best systems. This is the reason why we can assume that leaving some possible relevant documents out of the pool, as long as we can reliably distinguish between systems, this is not a problem. Because of this, the performance of adjudication methods is traditionally measured on its ability of *reliably* rank systems. If the adjudication method allows us to construct new judgements with fewer costs, and we still reliably distinguish between systems with those judgements, we can trust on the adjudication method and use it to build new judgements in the future. Commonly, the *reliability* of adjudication methods is evaluated with simulation. We can simulate, on an existing collection, that we gather a new set of judgements by running the given method on the existing collection. Using these new judgements, we compute evaluation scores for each system, and rank them according to this score. Now we can compare this ranking with the official evaluation of TREC. If both rankings are similar, we assume that the adjudication method is reliable. To measure this similarity, we usually use a ranking correlation, in particular Kendall's τ (Kendall 1938, 1948). There

[1] By absolute we mean the value we would get if we had complete relevance judgements.

are other strategies to measure the reliability of these methods but all of them share that they look at ranking position of systems, that is, if one particular system is ranked the same in both rankings.

Additionally, the ability to distinguish between systems should also hold for systems that did not participate in the pool or even did not exist when the collection was built. This is the *reusability*: the ability to correctly evaluate *unseen* systems. Reusability is commonly evaluated by performing *leave-out-uniques* experiments. Each run that contributes to the pool is evaluated both with the official assessments from the collection and the set of relevant documents produced by removing the relevant documents uniquely retrieved by that run. The idea is to simulate that the run did not contribute to the pool. Thus, if both evaluations are similar, one can conclude that the adjudication method is reusable.

2.2.4 *Significance Testing*

When evaluating and comparing the performance of multiple IR systems, we want to promote systems that are truly better than the others rather than systems that performed better by chance. There is noise inherent to the evaluation paradigm we have explained before: topics are just a sample of the whole universe of possible topics; the assessors that judge documents' relevance are humans, and thus their behaviour will be very variable; also, the set of documents that form the collection are not all the possible documents that a system can face. Because of all these issues, there is a chance that the results we see when comparing just average scores are not "really true", but due to the noise in the experimental setting. To tackle this problem, a usual evaluation pipeline includes performing a statistical test. Obviously, this is no free of problems, the output of a significance test is also subject to random chance. However, a significance test saying that a system is significantly better than other gives us more confidence in our results.

A significance test considers both a null hypothesis ($H0$) and an alternative hypothesis ($H1$). In IR, the most common approach is to run paired two-sided tests. In this case, the null hypothesis corresponds with the situation where the scores of both systems are from the same population, while the alternative hypothesis considers that the came from different ones. The significance test computes the probability of observing a difference of means that is, at least, as large as the difference computed from the observed means. If that probability, known as the *p-value*, is small, we have more chances that the difference we are seeing is "real", and not by chance: a lower p-value means more probability of rejecting the null hypothesis. There are several tests we can employ here,

differentiated by how they compute the p-value. The question of which test better for IR evaluation is still open in the field (Ferro and Sanderson 2019, 2022; Parapar, Losada and Barreiro 2021; Parapar, Losada, Presedo-Quindimil et al. 2020; Urbano, Lima et al. 2019; Urbano, Marrero et al. 2013).

Part II

BUILDING NEW COLLECTIONS

3

THE WISDOM OF THE RANKERS

As we have seen in the previous chapter, evaluating IR systems through test collections allows researchers and practitioners to compare different systems in a controlled and reproducible experimental setting. This methodology, known as the Cranfield paradigm, has fostered a tremendous number of advancements in recent decades (Voorhees 2019; Voorhees and Harman 2005), and is still common practice for competition-like initiatives such as TREC (Craswell, Mitra, Yilmaz, Campos and Lin 2021; Craswell, Mitra, Yilmaz, Campos, Voorhees et al. 2021; Voorhees and Harman 2005) and NTCIR (Kato et al. 2019).

Building a new test collection is expensive, mainly from collecting relevance assessments for topic-document pairs. To alleviate this cost, it is common practice to select the documents that merit human judgements by pooling over a set of rankings provided by the competition participants. This methodology, with its caveats (Buckley, Dimmick et al. 2007), has been put under consideration many times, and many times has been shown its benefits (Altun and Kutlu 2020; Cormack and Lynam 2007a; Cormack, Palmer et al. 1998; Lipani, Losada et al. 2019; Lipani, Zuccon et al. 2016; Losada, Parapar et al. 2016; Moffat, Webber et al. 2007; Otero, Parapar and Barreiro 2023; Otero, Parapar and Ferro 2023; Voorhees 2002; Voorhees, Soboroff et al. 2022; Zobel 1998). The main problem we want to highlight here is the dependency of this model to the existence of participant-generated rankings from where to pool the documents of each topic. This is a handicap in two ways. On the one hand, not everyone has access to enough resources to organize and run a competition-like initiative such as TREC. On the other hand, there are situations in which it would be useful to have human-annotated data before running a new task, for example, training data for a new niche retrieval task, where abundant and enough data is usually not available.

In this chapter, our main aim is to develop a new method to gather new pooled judgements when no real participant systems are available. We propose a method for creating simulated participant systems (Otero, Parapar and Barreiro 2021). Using automatically generated query variations and well-known retrieval models, we create new runs to pool documents from them. With this contribution, we give research and practitioners a useful tool to build new benchmarks when organizing an expensive evaluation campaign is not an option. The results we are presenting here have already been published (Otero, Parapar and Barreiro 2021; Otero, Valcarce et al. 2019). In this chapter, we perform some additional experimentation.

Section 3.1 introduces necessary content not introduced before, Section 3.2 explains our new proposed method in detail. In Section 3.3, we perform an extensive analysis of our method and provide a discussion on its advantages. Finally, in Section 3.4, we describe related work, then provide some concluding remarks in Section 3.5.

3.1 INTRODUCTION

Early works on IR research and evaluation were about ad hoc document retrieval (Cleverdon 1962; Kent et al. 1955). Although still nowadays a lot of research revolves around this core task, we have a lot of other tasks not related to document ranking. These tasks usually appear from new commercial needs (e.g., conversational assistants and recommender systems), and also from new research efforts from the community (Balog and Neumayer 2013; Basu et al. 2018; Ghanem et al. 2019; Losada, Crestani et al. 2017, 2019, 2020; Otero, Martin-Rodilla et al. 2021; Pérez et al. 2022). All these tasks cover and work on a wide range of different problems, with each of them having its own particularities. Still, to some degree, they all rely on the standard Cranfield paradigm and test collections for developing and evaluating new advancements. Thus, they also suffer from the same problems we outlined in the previous section. In this case, we refer to the need for participant-generated runs from which to pool the documents that merit human assessments. This problem becomes particularly severe for tasks and cases with very niche problems that do not have the support and resources of mainstream research.

We propose a new methodology to make it easier for researchers and practitioners to build retrieval test collections in scenarios with scarce resources and when organizing an expensive evaluation campaign is not an option. In the following section, we explain the details of our proposal.

3.2 CONSTRUCTING BENCHMARKS WITHOUT PARTICIPANT SYSTEMS

The main idea of our proposal is to create pools from which to select the documents in way that we do not need real participants like in a TREC-like competition. This is not because we do not acknowledge the usefulness of these campaigns, but because we want to develop a method to gather pooled judgements when organizing these campaigns is not possible.

Our approach relies mainly on two ideas. The first one is to generate, automatically, query variations for the same topic, with the aim of representing the fact that the same topic might be expressed in different ways by different users. The second idea is to employ off-the-shelf ranking models that are known to be robust and well-performing in different scenarios. By combining these two elements, we can generate a set of runs that will serve as the input to the pooling. Overall, we want our method to be simple and easily deployable, since our main aim is to ease the building of new test collections.

3.2.1 *Query Variants Generation*

We have devised three different simple ways of, given a specific topic, generating different query variations for the same topic. The first way that we propose is simply to use the *title* of the topic. This title expresses the main information need of the topic in just a few keywords, thus we think is a good idea as a starting point.

The second approach that we use is similar to the previous one. In addition to the *title*, in TREC, topics also have a *description* section that explains the information need in a more verbose way. Our idea is to use this *description* as the topic's query.

The third and final approach is as follows. Parting from the text of the topic's *title*, we create query variations by appending to this text the top-ranked term of the *description+narrative* sections, ranked by IDF. In other words: we take every term from the union of the terms in the *description* and *narrative* sections. Then, we compute the IDF of each term, and rank them by decreasing IDF. To generate query variations, we append each one of these terms to the topic's *title*. The idea here is to diversify the documents obtained with each variation by taking very "rare" terms, i.e. those that have the highest IDF.

Finally, to correctly ground and compare the results of these three proposals, we have also used a set of manually-generated variations. The idea here is to compare our three approaches with human-generated variants, which are supposed to perform better than automatically-generated ones. These

variations, that are available for the topics 301-450 and 600-700[1], were created by a team participating in the 2017 TREC CORE Track (Benham et al. 2017).

3.2.2 *Off-The-Shelf Ranking Methods*

The second tool that we use to generate synthetic runs are off-the-shelf ranking models. It is not our aim to use here every state-of-the-art model, but a set ranking models that represent different approaches to retrieval, with the aim of diversifying the pools. Our decision here was a very basic one: employ every retrieval model available through Apache Lucene at the time we run the experiments. Apache Lucene[2] is an open-source, Java library, that provides indexing and search features, and that is the most-used library for research purposes in IR. At the time we run these experiments, we used 72 different retrieval models. Among them, there is BM25 (Robertson and Zaragoza 2009), query likelihood model (Ponte and Croft 1998), with Dirichlet and Jelinek-Mercer smoothing, the Vector Space Model with TF-IDF weighting schema (G. Salton et al. 1975).

3.2.3 *Synthetic Pools*

From the combination of the four ways we have of generating query variations for each topic, and the use of the 72 ranking systems that we have available off-the-shelf, we have produced four different synthetic pools for each collection. We use the word synthetic to distinguish our pools from those created used real participant runs from a TREC competition. These four different pools are created as follows.

- *title pool*. The first pool that we create is what we name *title* pool. We run the *title* of each topic as a query against each of the 72 different models that we have, creating 72 different runs. These 72 runs form the *title* pool.

- *title+description pool*. The second pool that we create is what we name *title+description* pool. We repeat the same procedure as before, but using the *description* of the topic, instead of the *title*, as the query. From this procedure, we create 72 new runs. This pool is formed by these new 72 runs, and the 72 runs from the previous step, thus having in total 144 different runs.

[1] http://culpepper.io/publications/robust-uqv.txt.gz
[2] https://lucene.apache.org

Table 3.1: Statistics of the synthetic pools.

	TREC6				TREC7			
	title	*title+desc*	*manual*	*IDF*	*title*	*title+desc*	*manual*	*IDF*
Topics	50	50	50	50	50	50	50	50
Runs	72	144	648	648	72	144	648	648
Avg. pool size	333	662	1618	1330	328	590	1333	1310
Max. pool size	604	1126	2987	2816	570	922	2467	2655
Min. pool size	7	309	529	386	63	263	543	466
Pool depth (*k*)	100	100	100	100	100	100	100	100
Avg. # of relevants	31	40	59	48	30	38	62	51
Max. # of relevants	100	127	196	180	92	128	220	206
Min. # of relevants	0	0	2	0	2	3	4	5
	TREC8				**ROBUST04**			
	title	*title+desc*	*manual*	*IDF*	*title*	*title+desc*	*manual*	*IDF*
Topics	50	50	50	50	49	49	49	49
Runs	72	144	648	648	72	144	648	648
Avg. pool size	312	551	1278	1227	333	562	1385	1192
Max. pool size	476	881	2151	2139	592	953	2452	2895
Min. pool size	85	282	524	501	144	241	367	374
Pool depth (*k*)	100	100	100	100	100	100	100	100
Avg. # of relevants	37	44	64	56	24	29	35	32
Max. # of relevants	143	162	223	220	69	81	112	102
Min. # of relevants	3	4	6	3	1	2	3	2

- *IDF pool*. The third pool that we create is what we called *IDF* pool. The procedure to create this pool is similar to the previous ones. We run each query variation against each retrieval model to create a new run. Following the approach to generate variations that we explained before, we have generated 8 different variations for each topic. Thus, in this case, we are creating $8 \times 72 = 576$ different runs. As before, we also enlarge this set runs by using also the *title* pool.

- *manual* pool. The fourth and last pool that we create is what we called the *manual* pool. As with the previous one, we create 576 different runs by using 8 different manual variations for each topic against the 72 retrieval models. In this pool and in the previous one, we have used the same number of variations per topic to make a fair comparison between both. This number is imposed by the fact that in the existing manual variations, 8 is the minimum number available for every topic.

We provide some statistics of this pools in Table 3.1. This table includes the number of topics, the number of runs that compose each pool, the depth of the pool we have used to compute the statistics and make the experiments, the average number of documents (also the maximum and the minimum) per topic. Also, we have included the number of average relevant documents, the maximum, and the minimum per topic, according to the relevance indicated in the official qrels of the respective collections.

3.3 EXPERIMENTS

In this section, we now give the details of how we have evaluated our proposal, the datasets we have used, and other details about our experimental setup.

3.3.1 *Evaluation Procedure*

We have evaluated our proposal from two perspectives. First, the ability to find more relevant documents with less assessor effort. Second, the ability to correctly rank real, existing participant runs from TREC tracks.

3.3.1.1 *Recall*

We are interested in knowing how many judgements we need to reach a particular ratio of relevant documents found, since the more productive use of assessor's time is when they review documents that are in fact relevant. Thus, we have computed the recall at different levels of number of judgements

Table 3.2: Statistics of the collections used for experimentation.

	TREC6	TREC7	TREC8	ROBUST04
Topics	50	50	50	49
Runs	46	84	71	110
Avg. pool size	1445	1611	1786	1126
Max. pool size	1902	2585	2646	2355
Min. pool size	914	1025	1114	511
Pool depth (k)	100	100	100	100
Avg. # of relevants	92	93	94	42
Max. # of relevants	471	354	346	161
Min. # of relevants	3	7	6	3

per topic. In particular, assuming that we have n judgements per topic, we compute the recall at each of these points. Then, we summarize the recall by computing the area under the curve (AUC) of this recall's curve.

3.3.1.2 *Reusability*

Additionally, we are also interested in knowing to which extent our synthetic pools are able to correctly rank real participant runs from TREC competitions. To evaluate this, we have computed the Kendall's τ correlation (Kendall 1938) between the official ranking from TREC, and the ranking computed with the qrels derived from our synthetic pools.

3.3.2 *Datasets*

We conducted experiments on six standard collections from TREC. Our selection of datasets is constrained to those that have the manual query variations. That is, we have used the datasets that include the topic for which manual query variations are available. These datasets are: TREC6, TREC7, TREC8 and ROBUST04. We show some basic statistics of them in Table 3.2.

3.3.3 *Adjudication Methods*

To allocate simulated assessor judgements, we used the following adjudication methods to select the documents from the pools that merit human judgements[3]:

[3]We have used the official qrels from TREC to get the relevance of the documents.

- **top-k pooling.** We adapt the standard method used in TREC to limited-budget situations. When limiting the budget of assessments, we choose a k deep enough to fill that budget. Then, pooled documents are sorted by their document identifier (Voorhees and Harman 2005).

- **DocPoolFreq.** A voting-based method which scores each document with the total number of runs that retrieved it. In other works it is called PPTake@N (Lipani, Losada et al. 2019), or PRI on NTCIR tasks (Sakai, Tao et al. 2022).

- **MTF.** MTF is a dynamic adjudication method proposed by Cormack and colleagues (Cormack, Palmer et al. 1998) that has been acknowledged as a robust adjudication method (Altun and Kutlu 2020).

- **MM, MMNS, TS and TSNS.** These are bandit-based methods for document adjudication, which apply Bayesian principles to formalize the uncertainty associated with the probabilities of pulling a positive reward (a relevant document) from playing a bandit (Losada, Parapar et al. 2016).

3.3.4 *Results*

3.3.4.1 *Recall*

We have summarized the recall figures in four different tables. In Table 3.3, we report the recall results obtained using the *title* pools. Then, Tables 3.4 to 3.6 report the corresponding values when using the *title+description* pools, the *manual* pools, and the *IDF* pools, respectively.

Overall, we can observe that the *IDF* and *manual* pools yield better results than the other two. In particular, we see that both the *IDF* pools and the *manual* pools obtain better recall figures than the other two in every collection, with this difference being particularly higher for larger budgets.

Although it is not the main focus of this chapter, it is also interesting to have a look at the differences between the adjudication methods. From these results, it appears that bandit-based methods and, in particular, MM and MMNS, are the best performing methods, while top-k and DocPoolFreq lag behind the rest. This supports the claim that bandit-based methods are able to find more relevant documents with the same assessor effort, and thus are better tools at allocating this effort, allowing the building of larger test collections.

Table 3.3: Averaged Recall's AUC at different budgets per topic. Results obtained with the *title* pool. Statistically significant improvements w.r.t. top-k, DocPoolFreq, MTF, MM, MMNS, TS and TSNS are superscripted with a, b, c, d, e, f and g, respectively. These are also added at the beginning of each line to ease the comparison. For each collection and budget, best figures are **bolded** and worst ones are underlined.

	Judgements per topic				Judgements per topic			
	100	300	750	1000	100	300	750	1000
	TREC6				**TREC7**			
(a) top-k	0.13	0.20	0.23	0.25	0.12	0.18	0.22	0.24
(b) DocPoolFreq	0.15	0.30^{a}	0.31^{a}	0.31^{a}	0.13	0.26^{a}	0.27^{a}	0.27
(c) MTF	0.22^{ab}	0.33^{a}	0.34^{a}	0.34^{a}	0.20^{ab}	$\mathbf{0.30}^{a}$	$\mathbf{0.31}^{a}$	$\mathbf{0.31}^{a}$
(d) MM	$\mathbf{0.24}^{ab}$	$\mathbf{0.34}^{a}$	$\mathbf{0.35}^{a}$	$\mathbf{0.35}^{a}$	0.20^{ab}	0.29^{a}	$\mathbf{0.31}^{a}$	$\mathbf{0.31}^{a}$
(e) MMNS	0.23^{ab}	0.33^{a}	$\mathbf{0.35}^{a}$	$\mathbf{0.35}^{a}$	$\mathbf{0.21}^{ab}$	$\mathbf{0.30}^{a}$	$\mathbf{0.31}^{a}$	$\mathbf{0.31}^{a}$
(f) TS	0.23^{ab}	0.33^{a}	$\mathbf{0.35}^{a}$	$\mathbf{0.35}^{a}$	0.20^{ab}	$\mathbf{0.30}^{a}$	$\mathbf{0.31}^{a}$	$\mathbf{0.31}^{a}$
(g) TSNS	0.23^{ab}	0.33^{a}	$\mathbf{0.35}^{a}$	$\mathbf{0.35}^{a}$	$\mathbf{0.21}^{ab}$	$\mathbf{0.30}^{a}$	$\mathbf{0.31}^{a}$	$\mathbf{0.31}^{a}$
	TREC8				**ROBUST04**			
(a) top-k	0.17	0.24	0.29	0.31	0.18	0.24	0.29	0.30
(b) DocPoolFreq	0.16	0.32^{a}	0.34	0.34	0.19	0.42^{a}	0.45^{a}	0.45^{a}
(c) MTF	$\mathbf{0.26}^{ab}$	0.37^{a}	0.38^{a}	0.38^{a}	0.32^{ab}	0.47^{a}	0.48^{a}	0.48^{a}
(d) MM	0.25^{ab}	0.36^{a}	0.37^{a}	0.30^{a}	0.31^{ab}	0.46^{a}	0.47^{a}	0.47^{a}
(e) MMNS	$\mathbf{0.26}^{ab}$	$\mathbf{0.38}^{ab}$	$\mathbf{0.39}^{a}$	$\mathbf{0.39}^{a}$	$\mathbf{0.33}^{ab}$	$\mathbf{0.48}^{a}$	$\mathbf{0.50}^{a}$	$\mathbf{0.50}^{a}$
(f) TS	0.25^{ab}	0.36^{a}	0.37^{a}	0.37^{a}	0.31^{ab}	0.45^{a}	0.47^{a}	0.47^{a}
(g) TSNS	$\mathbf{0.26}^{ab}$	$\mathbf{0.38}^{ab}$	$\mathbf{0.39}^{a}$	$\mathbf{0.39}^{a}$	$\mathbf{0.33}^{ab}$	$\mathbf{0.48}^{a}$	$\mathbf{0.50}^{a}$	$\mathbf{0.50}^{a}$

Table 3.4: Averaged Recall's AUC at different budgets per topic. Results obtained with the *title+description* pool. Statistically significant improvements w.r.t. top-k, DocPoolFreq, MTF, MM, MMNS, TS and TSNS are superscripted with a, b, c, d, e, f and g, respectively. These are also added at the beginning of each line to ease the comparison. For each collection and budget, best figures are **bolded** and worst ones are underlined.

	Judgements per topic				Judgements per topic			
	100	300	750	1000	100	300	750	1000
	TREC6				**TREC7**			
(*a*) top-k	0.11	0.15	0.20	0.22	0.10	0.15	0.21	0.23
(*b*) DocPoolFreq	0.24^a	0.39^a	0.48^a	0.48^a	0.18^a	**0.33**^{a}	0.40^a	0.40^a
(*c*) MTF	0.26^a	0.38^a	0.46^a	0.47^a	0.20^a	0.32^a	0.39^a	0.39^a
(*d*) MM	**0.27**^{a}	**0.40**^{a}	**0.49**^{a}	**0.49**^{a}	**0.21**^{a}	**0.33**^{a}	**0.41**^{a}	**0.41**^{a}
(*e*) MMNS	**0.27**^{a}	0.39^a	0.47^a	0.48^a	0.20^a	**0.33**^{a}	0.40^a	0.40^a
(*f*) TS	0.26^a	**0.40**^{a}	0.48^a	**0.49**^{a}	0.19^a	0.32^a	0.40^a	0.40^a
(*g*) TSNS	**0.27**^{a}	0.39^a	0.47^a	0.48^a	0.20^a	0.32^a	0.39^a	0.39^a
	TREC8				**ROBUST04**			
(*a*) top-k	0.15	0.22	0.28	0.30	0.16	0.23	0.29	0.30
(*b*) DocPoolFreq	0.22^a	0.38^a	**0.46**^{a}	**0.46**^{a}	0.32^a	**0.53**^{a}	**0.62**^{a}	**0.62**^{a}
(*c*) MTF	0.25^a	0.38^a	0.45^a	0.45^a	0.35^a	0.51^a	0.60^a	0.60^a
(*d*) MM	0.25^a	0.38^a	**0.46**^{a}	**0.46**^{a}	0.34^a	0.50^a	0.59^a	0.59^a
(*e*) MMNS	**0.26**^{a}	**0.39**^{a}	**0.46**^{a}	**0.46**^{a}	**0.36**^{a}	**0.53**^{a}	**0.62**^{a}	**0.62**^{a}
(*f*) TS	0.25^a	0.38^a	**0.46**^{a}	**0.46**^{a}	0.34^a	0.50^a	0.59^a	0.59^a
(*g*) TSNS	**0.26**^{a}	**0.39**^{a}	**0.46**^{a}	**0.46**^{a}	**0.36**^{a}	**0.53**^{a}	**0.62**^{a}	**0.62**^{a}

Table 3.5: Averaged Recall's AUC at different budgets per topic. Results obtained with the *manual* pool. Statistically significant improvements w.r.t. top-k, DocPoolFreq, MTF, MM, MMNS, TS and TSNS are superscripted with a, b, c, d, e, f and g, respectively. These are also added at the beginning of each line to ease the comparison. For each collection and budget, best figures are **bolded** and worst ones are underlined.

	Judgements per topic				Judgements per topic			
	100	300	750	1000	100	300	750	1000
	TREC6				**TREC7**			
(a) top-k	0.12	0.18	0.23	0.25	0.12	0.19	0.25	0.27
(b) DocPoolFreq	**0.29**^{a}	0.43^a	0.56^a	0.60^a	0.24^a	0.38^a	0.52^a	0.56^a
(c) MTF	**0.29**^{a}	0.42^a	0.54^a	0.58^a	**0.25**^{a}	**0.39**^{a}	0.51^a	0.55^a
(d) MM	**0.29**^{a}	**0.44**^{a}	**0.57**^{a}	**0.61**^{a}	0.24^a	**0.39**^{a}	**0.53**^{a}	**0.57**^{a}
(e) MMNS	0.28^a	0.43^a	0.55^a	0.59^a	**0.25**^{a}	**0.39**^{a}	0.52^a	0.56^a
(f) TS	0.28^a	**0.44**^{a}	**0.57**^{a}	**0.61**^{a}	0.24^a	**0.39**^{a}	**0.53**^{a}	**0.57**^{a}
(g) TSNS	0.24^a	0.39^a	0.52^a	0.56^a	**0.25**^{a}	**0.39**^{a}	0.52^a	0.56^a
	TREC8				**ROBUST04**			
(a) top-k	0.15	0.23	0.29	0.31	0.17	0.24	0.30	0.31
(b) DocPoolFreq	0.26^a	**0.42**^{a}	**0.55**^{a}	**0.59**^{a}	**0.39**^{a}	**0.58**^{a}	**0.72**^{a}	**0.75**^{a}
(c) MTF	**0.27**^{a}	0.41^a	0.53^a	0.57^a	0.36^a	0.53^a	0.67^a	0.70^a
(d) MM	0.26^a	**0.42**^{a}	**0.55**^{a}	**0.59**^{a}	0.35^a	0.52^a	0.67^a	0.71^a
(e) MMNS	**0.27**^{a}	**0.42**^{a}	**0.55**^{a}	0.58^a	0.38^a	0.55^a	0.69^a	0.73^a
(f) TS	0.26^a	**0.42**^{a}	**0.55**^{a}	**0.59**^{a}	0.35^a	0.52^a	0.67^a	0.71^a
(g) TSNS	**0.27**^{a}	**0.42**^{a}	**0.55**^{a}	0.58^a	0.35^a	0.53^a	0.67^a	0.71^a

Table 3.6: Averaged Recall's AUC at different budgets per topic. Results obtained with the *IDF* pool. Statistically significant improvements w.r.t. top-k, DocPoolFreq, MTF, MM, MMNS, TS and TSNS are superscripted with a, b, c, d, e, f and g, respectively. These are also added at the beginning of each line to ease the comparison. For each collection and budget, best figures are **bolded** and worst ones are underlined.

	Judgements per topic				Judgements per topic			
	100	300	750	1000	100	300	750	1000
	TREC6				**TREC7**			
(a) top-k	0.11	0.16	0.20	0.21	0.11	0.18	0.22	0.24
(b) DocPoolFreq	0.23^a	0.37^a	**0.49**^{a}	**0.51**^{a}	0.19^a	0.32^a	0.44^a	0.47^a
(c) MTF	0.26^a	0.38^a	0.47^a	0.50	0.21^a	0.34^a	0.45^a	0.47^a
(d) MM	0.26^a	0.38^a	**0.49**^{a}	**0.51**^{a}	**0.22**^{a}	**0.35**^{a}	0.46^a	**0.49**^{a}
(e) MMNS	**0.27**^{a}	**0.39**^{a}	**0.49**^{a}	**0.51**^{a}	**0.22**^{a}	0.34^a	0.46^a	0.48^a
(f) TS	0.26^a	0.38^a	**0.49**^{a}	**0.51**^{a}	**0.22**^{a}	**0.35**^{a}	**0.47**^{a}	**0.49**^{a}
(g) TSNS	0.26^a	0.38^a	0.48^a	**0.51**^{a}	**0.22**^{a}	0.34^a	0.46^a	0.48^a
	TREC8				**ROBUST04**			
(a) top-k	0.15	0.21	0.27	0.29	0.15	0.22	0.27	0.29
(b) DocPoolFreq	0.23^a	0.38^a	**0.51**^{a}	**0.54**^{a}	**0.33**^{a}	**0.51**^{a}	**0.65**^{a}	**0.68**^{a}
(c) MTF	**0.26**^{a}	0.39^a	0.49^a	0.52^a	0.32^a	0.48^a	0.61^a	0.64^a
(d) MM	0.25^a	0.39^a	**0.51**^{a}	**0.54**^{a}	0.31^a	0.46^a	0.61^a	0.65^a
(e) MMNS	**0.26**^{a}	**0.40**^{a}	**0.51**^{a}	**0.54**^{a}	**0.33**^{a}	0.50^a	0.64^a	0.67^a
(f) TS	0.25^a	0.38^a	0.50^a	0.53^a	0.31^a	0.46^a	0.61^a	0.65^a
(g) TSNS	**0.26**^{a}	**0.40**^{a}	**0.51**^{a}	**0.54**^{a}	**0.33**^{a}	0.50^a	0.64^a	0.67^a

Table 3.7: Minimum number of judgements per topic needed to obtain values of 0.90 for Kendall's τ correlation with respect to the official ranking of TREC submissions. An * means that no algorithm achieved a 0.90 correlation. For each collection and correlation, best values (lowest) are **bolded** and worst ones are underlined.

	manual	*IDF*	*manual*	*IDF*
	TREC6		**TREC7**	
top-k	719	*	308	740
DocPoolFreq	541	*	385	700
MTF	465	*	226	704
MM	489	*	207	589
MMNS	522	*	225	619
TS	**459**	*	**164**	**571**
TSNS	751	*	225	619
	TREC8		**ROBUST04**	
top-k	311	774	315	569
DocPoolFreq	376	650	**141**	453
MTF	**220**	748	229	469
MM	224	**565**	195	364
MMNS	223	804	203	**357**
TS	243	600	199	378
TSNS	223	804	255	**357**

3.3.4.2 *Reusability*

Since in the previous section we have seen that the *IDF* and *manual* pools are the ones that yield the best results, we focus this analysis only on these two.

Table 3.7 presents the reusability results. In particular, it shows the minimum number of judgement per topic that each method needs to reach a Kendall's τ correlation of 0.90 with the official ranking of systems. Again, *manual* pools are the best performing ones. This is something expected, since human-crafted query variations will be better reflecting the different aspects of the same topic than automatically-generated variations. However, we still see very competitive results obtained with the *IDF* pools. In fact, we obtained very strong correlations above 0.90, with every algorithm for collections TREC7, TREC8 and ROBUST04, while greatly reducing the number of judgements per topic. If we compare the figures we see in this table with the statistics reported in Table 3.2, we see that the average number of judgements per topic employed originally for building the collections is much higher than the one we need here to reach a 0.90 correlation.

3.4 RELATED WORK

In this chapter, we have proposed a simple way of generating new pools without the need of organizing an expensive evaluation campaign with the aim of facilitating the construction of new benchmarks. These benchmarks are the cornerstone element of offline retrieval evaluation, but they are very expensive to build from many perspectives. Ours is not the first attempt of reducing this cost. However, to the best of our knowledge, there is little work in trying to overcome, or at least to alleviate, the handicap of relying on high-quality diverse pools to gather the judgements. As we said before, this is useful for practitioners and researchers that have enough resources to allow for this, but not everyone can. One attempt that is tightly related to our approach is the work by Sanderson and Joho (2004). In this work, they put into question the assumption that some sort of pooling is needed to build high-quality benchmarks. They used just a single system in various feedback rounds to gather new relevant documents, these relevant documents where then used to recompute the relevance model of the system, and perform another search. This process repeated for five rounds. Authors concluded that "a single system can generate a usable set of qrels, and that the process building qrels using a single system can be facilitated by the relevance feedback".

3.5 CONCLUSIONS

We have proposed a new method for creating new document pools in an easy, automatic way. Our aim was to allow for a way of building high-quality benchmarks when organizing a TREC-like initiative is not an option. By running off-the-shelf ranking systems combined with automatically generated query variations, we created a set of sufficiently diverse runs to gather pooled judgements. Our results showed that our simulated qrels obtained strong correlations with the official ranking of TREC participants, acknowledging the utility of our proposed approach. Additionally, we also demonstrated that it is possible to find a great deal of the relevant documents with many fewer judgements than were required with traditional top-k pooling, showing that it is possible to reduce the effort needed to build new collections even further.

4

CONTENT-BASED DOCUMENT ADJUDICATION

As we have seen, constructing new datasets for IR evaluation has, among others, two difficult challenges. First, the need of having enough and diverse submissions from which to pool documents. Second, judging those documents without incurring in tremendous unaffordable costs and ensuring their reusability. In the previous chapter, we proposed an approach to tackle the former. By simulating diverse participant systems in combination with automatically generated query variations, we created synthetic runs from which to pool the documents to create reusable judgements. In this chapter, we aim to propose an adjudication method that guides the pooling process efficiently to tackle the second challenge. Our objective is to better employ the assessor's limited budget for judgement and thus allow the building of new collections with comparable quality but lower costs.

We formulate a new adjudication method that exploits the textual content of the documents that are pooled (Otero, Parapar and Barreiro 2023). Past adjudication methods have ignored this textual information, focusing on different factors such as the position of the documents in the submissions or the number of submissions that retrieved the same document, among others. Here, we show that this textual content is a helpful resource that helps estimate the relevance of the documents for selecting which ones merit human judgement. Our experiments show that our proposal is a cost-effective alternative to existing methods. In this chapter, we extend our work (Otero, Parapar and Barreiro 2023) by bringing more compared methods into the experimentation.

Hereafter, in Section 4.1 we give a brief introduction of our proposal; Section 4.2 comments on specific background not introduced in previous chapters; in Section 4.3 we provide a detailed definition of our proposed method, which is then thoroughly compared against state-of-the-art models

in Section 4.4. Finally, Section 4.5 includes a description of past work closely related to ours, and Section 4.6 explains this chapter's main conclusions and insights.

4.1 INTRODUCTION

As we have seen through this thesis, reproducible offline evaluation of retrieval systems relies on test collections (Sanderson 2010; Voorhees 2002). With these collections, which include topics, documents and relevance judgements, researchers can perform reproducible evaluations to compare the performance of existing and new IR systems. However, how to build *reusable* test collections that represent the always-increasing size of the web—and other kind of collections— and the particularities of new tasks at an affordable cost is still an open question. As we already explained, *reusability* refers to the ability of a collection to correctly evaluate systems that did not contribute to the building of that collection or even did not exist at that moment. Driven by the importance of test collections for IR evaluation and the high costs of building new ones, much research has been devoted to developing techniques to build new collections cost-effectively.

One of these approaches aims to improve the allocation of the number of judgements that can be made. That is, given a limited budget for documents to review, prioritize those that may have a higher chance of being relevant. As we described in Chapter 2, a method used to prioritize the pooled documents is called an adjudication method. We do not intend to make an extensive review of the different adjudication methods that exist here. We refer the reader to Chapter 2 for an extensive and detailed explanation of these approaches. The important thing to mention here is that most of these techniques focus on factors such as the position of the documents in the submissions or the number of submissions that retrieved the documents—as a voting mechanism—to estimate their relevance and prioritize the documents based on this information. In most of them, the textual content of the prioritized documents is just ignored. We argue that this textual content could be a valuable signal for estimating the document's relevance and thus use this content to prioritize the pooled documents. There are several ways in which we could take advantage of this textual information to estimate the relevance of the documents. We employ a seriesof relevance feedback methods that work well in the ad hoc retrieval task for estimating the document's relevance. Our proposed adjudication method is a cost-effective alternative that improves the appearance ratio of relevant documents without harming the reliability, fairness or the reusability of the judgements.

In the following section, we explain how traditional relevance feedback methods work and then explain how we apply them for document adjudication.

4.2 RELEVANCE FEEDBACK FOR AD HOC RETRIEVAL

IR systems help users in finding relevant pieces of information within repositories of tremendous size. Nowadays, IR systems are essential components to other neighbouring areas, such as web search or recommender systems. Due to its pervasiveness, there exists great interest in constantly improving them.

The most natural approach would be trying to improve the retrieval model that supports the system. However, this is not the only way. For instance, query expansion, a technique that consists in expanding the original query issued by the user with new terms, is a useful way of improving the retrieval results without changing the underlying retrieval model (Carpineto and Romano 2012). Relevance feedback (RF) is a query expansion technique. Under this approach, users indicate which documents from the ranking presented to them are relevant with respect to their information need. Then, this feedback is combined with the information of the original query to create an expanded query that is reissued to the system to obtain a second ranking. Despite its usefulness, manual relevance feedback is difficult to obtain in many situations. For this reason, most RF research has focused on developing techniques that could improve retrieval results without user interaction. Pseudo-relevance feedback (PRF) is an example. Also known as blind relevance feedback, is an automatic query expansion technique. Under this approach, the top-r documents retrieved by a search engine for a given query are assumed to be relevant. Instead of asking the user for this feedback, we just assume it. Using this set of documents, which is called pseudo-relevant set, and the text of the original query, these methods obtain new and reweighted terms to expand the original query. This expanded query is then reissued to the search engine to obtain a second ranking, just as in real[1] RF. Both techniques, RF and PRF, serve to refine the original decision of relevance made by the search engine. The difference is that in RF the information is explicitly stated by the user, while in PRF that relevance information is assumed. This makes PRF an appealing approach, since it allows improving the effectiveness of a retrieval system without changing its model and without needing the interaction of the user. Several PRF methods have been proposed in the past. We will centre

[1] The word real is used to distinguish it from pseudo-relevance feedback, where the feedback is assumed, i.e. it is not "real".

this work on those based on the statistical language modelling framework, since they perform empirically the best (Lv and Zhai 2009).

4.2.1 *Relevance Feedback Methods in the Language Modelling Framework*

Within this framework, the basic retrieval model is the Kullback-Leibler Divergence (KLD) (Lafferty and Zhai 2001) denoted by $D(\cdot\|\cdot)$, between the query language model θ_q and the document language model θ_d. This is rank equivalent to the negative cross entropy between both distributions:

$$Score(d,q) = -D(\theta_q\|\theta_d) \overset{\text{rank}}{=} \sum_{w\in V} p(w|\theta_q)\log p(w|\theta_d) \tag{4.1}$$

where V is the set of words in the vocabulary of the collection. Using this model, we can obtain a score for each document given a particular query, and obtain a ranking of documents in decreasing order of score. Now, we have to define how the query model ($p(w|\theta_q)$) and the document model ($p(w|\theta_d)$) are estimated.

In the most basic alternative, we could define both via the Maximum Likelihood Estimate (MLE) of a multinomial distribution: $p(w|\theta_q) = \frac{tf(w,q)}{|q|}$, and $p(w|\theta_d) = \frac{tf(w,d)}{|d|}$, where $tf(w,q)$ is the frequency of w in q, $tf(w,d)$ is the frequency of w in d, and $|q|$ and $|d|$ are the lengths of the query and the document, respectively. However, this would assign zero probabilities to query words that are not present in the documents, harming the final results. This problem is tackled by the process of *smoothing*, which consists in adjusting the MLE to compute more accurate estimation. In this work, we will use Dirichlet smoothing to compute the document model:

$$p(w|\theta_d) = \frac{tf(w,d) + \mu \cdot p(w|\theta_C)}{|d| + \mu} \tag{4.2}$$

where $tf(w,d)$ is the frequency of w in d, $p(w|\theta_C)$ is the MLE of w in the collection, and μ is a parameter of the smoothing (commonly set to $\mu = 1000$).

Without any extra information, query models (i.e., $p(w|\theta_q)$) are usually estimated using only the text of the query. However, we can exploit feedback information, whether real or pseudo, to estimate a more accurate query language model θ_F. Besides, this new feedback model can be interpolated with the model of the original query to improve the retrieval effectiveness:

$$p(w|\theta'_q) = \alpha\, p(w|\theta_q) + (1-\alpha)\, p(w|\theta_F) \tag{4.3}$$

where $\alpha \in [0,1]$ controls the importance of the relevance feedback. Thus, the goal of a feedback model is to provide an estimation of θ_F. With this model,

the terms of the original query are expanded and reweighed (Equation 4.3) to obtain a second retrieval. In particular, using the probability of each term computed with Equation 4.3, the top-e terms are selected to build a new query that is reissued to the retrieval system and the new ranking is computed as in Equation 4.1.

4.2.1.1 *Relevance-Based Language Models*

Relevance-based language models or, for short, relevance models (RM), are a technique to compute a new query model that were devised to explicitly introduce the concept of relevance in statistical language models. Lavrenko and Croft (2001) proposed RM1, a model that is a weighted average of the probability of the word w given by each document in the pseudo-relevant set, where the weights are the query likelihood scores for those documents. When the weight of the feedback model is interpolated with the original query (Equation 4.3), it is coined as RM3 (Abdul-Jaleel et al. 2004). In this work, we focus only on RM3, since it has been shown to perform better than RM1 (Lv and Zhai 2009). More formally, let F be the set of pseudo-relevant documents, then RM1 is estimated as follows:

$$p(w|\theta_F) \propto \sum_{d \in F} p(\theta_d)\, p(w|\theta_d) \prod_{q \in Q} p(q|\theta_d) \tag{4.4}$$

where $p(w|\theta_d)$ is the smoothed language model of each feedback document, computed using Dirichlet priors (Zhai and Lafferty 2001a, 2004). Usually, $p(\theta_d)$ is assumed to be uniform among all documents, thus the estimation reduces to:

$$p(w|\theta_F) \propto \sum_{d \in F} p(w|\theta_d) \prod_{q \in Q} p(q|\theta_d) \tag{4.5}$$

4.2.1.2 *Divergence Minimization Model*

Divergence Minimization Model (DMM) (Zhai and Lafferty 2001b) is a RF technique which assumes that the feedback model θ_F should be close to the language model of the pseudo-relevant documents F but far away from the background model. The model is computed as follows:

$$p(w|\theta_F) \propto \exp\left(\frac{1}{1-\lambda}\frac{1}{|F|}\sum_{d \in F} \log p(w|\theta_d) - \frac{\lambda}{1-\lambda}\log p(w|\theta_C)\right) \tag{4.6}$$

where $p(w|\theta_d)$, in this case, is computed using additive smoothing as recommended (Hazimeh and Zhai 2015):

$$p(w|\theta_d) = \frac{tf(w,d) + \gamma}{|d| + \gamma \cdot |V|}$$

This model has a parameter λ that controls the influence of the collection language model and a parameter γ that controls the smoothed document model.

4.2.1.3 *Maximum-Entropy Divergence Minimization Model*

Maximum-Entropy Divergence Minimization Model (MEDMM) (Lv and Zhai 2014) is a RF technique that stems from DMM (Zhai and Lafferty 2001b). Based on the same idea that DMM, this new estimation aims to overcome some of the DMM problems (Lv and Zhai 2014).

This model is proposed as an optimization problem, on which applying Lagrange Multiplier leads to the following analytic solution (Lv and Zhai 2014):

$$p(w|\theta_F) \propto \exp\left(\frac{1}{\beta}\sum_{d\in F}\alpha_d \log p(w|\theta_d) - \frac{\lambda}{\beta}\log p(w|\theta_C)\right) \tag{4.7}$$

where $p(w|\theta_d)$ and $p(w|\theta_C)$ are computed as in Equation 4.6. This model has two parameters: λ controls the IDF effect, giving more importance to terms that appear less in the collection, i.e. terms with a higher IDF (Robertson 2004); and β controls the entropy of the feedback language model. In contrast with DMM, where each feedback document is weighted equally (setting $\alpha_d = \frac{1}{|F|}$), MEDMM gives each feedback document a different weight, based on the posterior of the document language model:

$$\alpha_d = p(\theta_d|q) = \frac{p(q|\theta_d)}{\sum_{d'\in F} p(q|\theta_{d'})} = \frac{\prod_{w\in q} p(w|\theta_d)}{\sum_{d'\in F}\prod_{w'\in q} p(w'|\theta_{d'})} \tag{4.8}$$

4.3 RELEVANCE FEEDBACK FOR DOCUMENT ADJUDICATION

At the beginning of this chapter, we stated that our aim was to employ the textual content of the pooled documents to improve the adjudication process when building new test collections. We have just seen how RF methods employ this textual content in ad hoc retrieval for estimating a probabilistic relevance model of the documents. Now marrying these two ideas is simple: we could use the above relevance feedback models to prioritize the pooled documents when building new test collections.

In Section 4.2, we explained the motivation of using pseudo feedback for estimating the probabilistic model: gathering real feedback from the user is difficult as it burdens the task of the user of finding relevant pieces of information to their information need. Although this real relevance feedback might seem more appealing, if the underlying retrieval model provides decent results, the expanded query might provide better results than the original one. Thus, for the ad hoc retrieval task, assuming the relevance of the top-r documents without user interaction is not a strong assumption.

The key idea here is that, for the task of document adjudication for test collections, real relevance feedback is, actually, much easier to obtain. During the adjudication process, we can use the documents marked as relevant by the assessor to estimate a probabilistic relevance model, and then use this method to improve the ranking in which assessors navigate the pool. We now explain our method in more detail.

4.3.1 *Reranking the Pool*

As we said, the main idea of this approach is to use the estimated feedback model to prioritize all pooled documents as a whole. We gradually enlarge the relevance feedback set F with the relevant documents. After adding a new relevant document to the set, we update our feedback model estimate and rank the documents in the pool. Assessors inspect the documents from the pool according to this ranking. When a new relevant document appears, we again add the document to the relevance set, update the model estimate and reorder de pool again, repeating the loop until the assessment budget is consumed.

We show the pseudocode of this method in Algorithm 4.1. First, we construct the pool by taking the union of the top-k documents of each submission (Line 8). Then, we obtain a ranking of the pooled documents (Line 9). At this moment, the relevance set is empty. Thus there is no relevance feedback going on here. Then, the adjudication process begins. The assessors inspect the pool in the order set by the ranking (Line 11). The first document of this rank is judged (Lines 12 and 13), and removed from the pool (Line 14), so that we do not see it again later. If this document is relevant, we add it to the relevance set F (Line 16), and rerank the unjudged documents that are left in the pool (Line 17). If it is not relevant, the assessor just keeps judging the documents in the order set by the last time the pool was reranked. This process continues until the budget for judgements is exhausted (Line 10), which we assume is lower than the size of the pool.

Now we have left to discuss the implementation of the rerank (Lines 9 and 17) function. Here is where we use the relevance feedback models we described in the previous section: RM3, DMM, and MEDMM. These different models estimate the relevance model in different ways, and thus will provide different rankings of the pool given the same documents in the relevance set F. In our experiments, we only report the results of DMM (named just like this in tables and figures) since it was the one that performed the best among the relevance feedback models.

Algorithm 4.1 Document Adjudication with Relevance Feedback

Input:

1: $\mathcal{P}_q$ set of runs for a topic.
2: q a topic query (*title* + *description* of TREC topics).
3: b budget size.
4: k depth.

Output:

5: $\mathcal{R}$ set of judgements for a topic.

6: $\mathcal{R} \leftarrow \varnothing$ ▷ *Set of judgements.*
7: $\mathcal{F} \leftarrow \varnothing$ ▷ *Relevance set.*
8: $\mathcal{P} \leftarrow \textsc{get\_pool}(\mathcal{P}_q, k)$ ▷ *Union of top-k documents.*
9: $r \leftarrow \textsc{rerank}(\mathcal{P}, \mathcal{F}, q)$ ▷ *Get initial ranking of the pool*
10: **while** $|\mathcal{R}| < b$ **do** ▷ *Judge documents until the budget is exhausted.*
11: $d \leftarrow \textsc{pop\_top\_ranked\_doc}(r)$
12: $j \leftarrow \textsc{judge}(d, q)$ ▷ $j \in \mathbb{N}_0$
13: $\mathcal{R} \leftarrow \mathcal{R} \cup \{(d, j)\}$
14: $\mathcal{P} \leftarrow \mathcal{P} \setminus \{d\}$
15: **if** $j > 0$ **then** ▷ *If the document is relevant*
16: $\mathcal{F} \leftarrow \mathcal{F} \cup \{d\}$ ▷ *we add it to the relevance set,*
17: $r \leftarrow \textsc{rerank}(\mathcal{P}, \mathcal{F}, q)$ ▷ *and rerank the pool with RF.*

4.3.2 *Reranking Each Submission*

With this algorithm we have just presented, we are reranking the entire set of pooled documents as a whole, discarding the fact that different submissions might provide a higher number of relevant documents than others. Thus, allocating more judgements for those and avoiding sampling documents from poor runs could be useful.

As before, the idea is to use relevance feedback for prioritizing the adjudicated documents. However, instead of reranking the whole set of pooled

documents, we will rerank each participant submission separately. We want to better employ the judgements budget by focusing more on promising runs, while refining their ranking with relevance feedback. We will implement a policy to jump among submissions to choose from which run we will sample, and then use a relevance feedback model for reranking this submission every time a new relevant document appears.

We provide a detailed pseudocode of this algorithm in Algorithm 4.2. Here, we have a policy that maintains a priority for each submission (Line 9), and that is responsible for choosing, at each time, from which submission we are going to sample the next document to judge (Line 12). Once we pick the next document from the top of the chosen run (Line 14), it gets judged (Line 23). The priority of the current run is updated depending on the document's relevance. We also update the priority of every run that contains this same document, even if it is not in their top position. This reward propagation helps to better estimate the probabilities the policy assigns to each run. Then, if the document was relevant, we add it to the relevance set (Line 28), and, with a new estimation of the relevance feedback model, we rerank the documents of each run (Line 29). Then, the process is repeated: the policy chooses which run to sample, we judge its first document, and so on.

There are several options to implement the policy that updates each submission probability and chooses the next one based on these. Given the recent success demonstrated by dynamic adjudication methods (Altun and Kutlu 2020; Losada, Parapar et al. 2017; Otero, Parapar and Ferro 2023), we have chosen to use them for this purpose. In particular, we have used bandit-based methods, namely, MM, MMNS; we have also used MTF. As before, we used RM3, DMM, and MEDMM as reranking methods also in this case. Again, the best performing one was always DMM, and thus is the one we report in the results section. In our experiments, the instantiations of these algorithms are thus referred to as MTF+DMM, MM+DMM, and MMNS+DMM.

Now that we have thoroughly explained our proposed algorithms for document adjudication, we proceed to evaluate them.

4.4 EXPERIMENTS

To evaluate our proposals, we conducted a series of standard experiments under different experimental setups, which we explain now.

Algorithm 4.2 Document Adjudication by Reranking Each Submission

Input:

$\mathcal{P}_q$ set of submissions for a topic.
q a topic query (*title* + *description* of TREC topics).
b budget size.

Output:

$\mathcal{R}$ set of judgements for a topic.

$\mathcal{R} \leftarrow \varnothing$ ▷ *Set of judgements.*
$\mathcal{F} \leftarrow \varnothing$ ▷ *Relevance set.*
$p[]$ ▷ *Array of length $|\mathcal{P}_q|$ to save a probability for each pooled submission.*
for $i \leftarrow 0, |\mathcal{P}_q|$ **do**
    $p[i] \leftarrow 1/|\mathcal{P}_q|$ ▷ *Each submission has a priority that is initialized uniformly.*
r ▷ *Variable to save a submission as a list of ranked documents.*
▷ *Choose next submission according to their probabilities.*
$r \leftarrow$ CHOOSE_SUBMISSION$(\mathcal{P}_q, p)$
while $|\mathcal{R}| < b$ and $|\mathcal{P}_q| \neq 0$ **do**
    $d \leftarrow$ POP_TOP_RANKED_DOC(r) ▷ *Remove top-ranked document from r and save it in d.*
    ▷ *If the submission is empty after removing the document, we remove it from the pooled runs.*
    if r is empty **then**
        $\mathcal{P}_q \leftarrow \mathcal{P}_q \setminus \{r\}$
    ▷ *If we have already judged the document and the submission is empty, choose a new submission.*
    if $d \in \mathcal{R}$ and r is empty **then**
        $r \leftarrow$ CHOOSE_SUBMISSION$(\mathcal{P}_q, p)$
    ▷ *If the document is not judged*
    if $d \notin \mathcal{R}$ **then**
        $j \leftarrow$ JUDGE(d, q) ▷ $j \in \mathbb{N}_0$
        ▷ *Reward propagation: update the priority of submissions that retrieved the current document.*
        for all $s \in \{s \in \mathcal{P}_q : d \in s\}$ **do**
            $p[s] \leftarrow$ UPDATE_PRIORITY$(p[s], j)$
        if $j > 0$ **then** ▷ *if d is relevant*
            $\mathcal{F} \leftarrow \mathcal{F} \cup \{d\}$
            for all $s \in \mathcal{P}_q$ **do** ▷ *Rerank every submission.*
                $s \leftarrow$ RERANK$(s, \mathcal{F}, q)$
            if r is empty **then**
                $r \leftarrow$ CHOOSE_SUBMISSION$(\mathcal{P}_q, p)$
        else
            $r \leftarrow$ CHOOSE_SUBMISSION$(\mathcal{P}_q, p)$
        $\mathcal{R} \leftarrow \mathcal{R} \cup \{(d, j)\}$

Table 4.1: Statistics of the collections used for experimentation.

	TREC5	TREC6	TREC7	TREC8	TREC9	CDS14	CDS15	CDS16
Train/Test	Train	Test	Test	Test	Test	Train	Test	Test
Topics	50	50	50	50	50	30	30	30
Runs	101	46	84	71	59	102	102	115
Teams	30	33	41	38	21	26	36	26
Avg. pool size	2692	1445	1611	1786	1404	770	591	666
Max. pool size	4472	1902	2585	2646	2978	987	859	906
Min. pool size	1623	914	1025	1114	710	554	351	452
Pool depth (k)	100	100	100	100	100	20	20	15
Avg. # of relevants	110	92	93	94	52	82	95	112
Max. # of relevants	593	471	354	346	517	304	390	479
Min. # of relevants	1	3	7	6	1	10	2	3

4.4.1 *Collections*

We have performed experiments on eight different collections. In Table 4.1, we include some statistics about them. TREC5–8 are classic testbeds associated with the TREC ad-hoc retrieval task, while TREC9 comes from the web track. CDS14–16[2] are newer collections created in the TREC Clinical Decision Support (CDS) track.

4.4.2 *Compared Methods*

We have compared the following adjudication methods:

- **top-k**: it is the most basic baseline and the method used traditionally in TREC (Voorhees and Harman 2005). With this method, every document present in the first k documents of every submission gets a judgement. To limit the budget of judgements for each topic, we select a value of k deep enough to fill that budget, and then sort the pooled documents by their document identifier. For this reason, this method is also known as DocID in the literature.
- **MTF** (Cormack, Palmer et al. 1998): MTF is a dynamic adjudication method known to be robust for creating new judgements for IR evaluation (Altun and Kutlu 2020; Losada, Parapar et al. 2017; Otero, Parapar and Ferro 2023).

[2] According to the overviews of these tracks, the official qrels include some documents sampled outside the top-k pool. To have a fair comparison in every collection, we just use the top-k pools.

- **MM** (Losada, Parapar et al. 2017): a bandit-based method, which is also a very robust baseline (Altun and Kutlu 2020; Losada, Parapar et al. 2017; Otero, Parapar and Ferro 2023).
- **DMM**: the instantiation of Algorithm 4.1 with DMM as the model for reranking the pool.
- **MTF+DMM**: the instantiation of Algorithm 4.2 with MTF as the policy of choosing the submissions, and DMM as the model for reranking these submissions.
- **MM+DMM**: the instantiation of Algorithm 4.2 with MM as the policy of choosing the submissions, and DMM as the model for reranking these submissions.
- **MMNS+DMM**: the instantiation of Algorithm 4.2 with MMNS as the policy of choosing the submissions, and DMM as the model for reranking these submissions.

4.4.3 *Metrics*

We evaluated the proposals from four different perspectives: recall, reliability, fairness and reusability.

4.4.3.1 *Recall*

We study the pooling strategies regarding their ability to identify relevant documents early, i.e., the sooner they obtain high recall values, the better. We do this as follows: for each topic, an adjudicating method creates a sequence of judgements of the pooled documents. We can compute recall@n at any point in this sequence, where n is the number of judgments. The most productive use of assessors' time is when they judge relevant documents. Also, at any point in this sequence, we can obtain the accumulated area under the curve (AUC) of this Recall curve. Our main metric is the Recall's AUC averaged over the set of topics in each collection.

4.4.3.2 *Reliability*

We also study the reliability of the methods to induce the same ranking of systems as the official qrels. To evaluate it, we compute two different ranking correlations, Kendall's τ (Kendall 1938, 1948) and τ_{AP} (Yilmaz et al. 2008),

between the official rankings of systems and the ranking obtained with each adjudicating method. Each ranking is constructed using AP as the measure to score the runs. From now on, we always use AP for scoring the runs, unless stated otherwise.

4.4.3.3 *Fairness*

We must consider if the collection can provide a fair comparison between the runs that participated in the pool. Following the same approach as Voorhees (2018) to evaluate the fairness, we compute the maximum drop (negative change) suffered by a run when ranking it with the evaluated method compared with the official ranking. In particular, we build a new qrels file using each of the adjudicating methods proposed. Then, we rank the runs (using AP scores) using this reduced qrels file. We compute the difference between a run's position in the official ranking and the ranking obtained with the test qrels. We do this for every run. The maximum negative drop suffered by a run is what we call MaxDrop. A high MaxDrop means that a system is treated differently in both rankings.

4.4.3.4 *Reusability*

A test collection is reusable if it is able to correctly evaluate runs that did not contribute to the pool. We performed a *leave-one-group-out* (LOGO) experiment to measure the reusability (Voorhees 2002). This type of tests are a common way of evaluating the reusability of retrieval test collections (Craswell, Mitra, Yilmaz, Campos, Voorhees et al. 2021; Zobel 1998). In these tests, the ground-truth rankings of a team's runs are compared to the rankings that those runs would have obtained if the team had not participated in the construction. When these rankings are similar, we can conclude that the collection is reusable. In this work, in particular, we perform the experiment in this way. We create a reduced qrels file for each team that participated in the competition without that team's runs and the corresponding adjudicating method. Next, we rank the participant runs using both the ground-truth qrels and the reduced qrels. Finally, we compute the Kendall's τ correlation between both rankings. Our evaluation measure is Kendall's τ averaged over all teams.

Table 4.2: Tuned parameters after optimization.

Algorithm	TREC5		CDS14	
	e	α	e	α
DMM	75	0.1	75	0
MTF+DMM	75	0	75	0
MM+DMM	50	0.20	50	0.30
MMNS+DMM	100	0	75	0

4.4.4 *Training and Testing*

We performed a training and test strategy optimising for Recall. In this regard, there are several parameters to train. The feedback model itself is also a parameter, which we tune among the following values: RM3, DMM or MEDMM. The number of expansion terms, e, and the degree of interpolation between the expanded query and the original one, α. The number of expansion terms was tuned among $e \in \{5, 10, 25, 50, 75, 100\}$. The interpolation parameter was tuned among $\alpha \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$.

We used two collections for tuning these parameters, TREC5 and CDS14. After optimisation, the best performing feedback model for document adjudication was DMM. The best parameters obtained with TREC5 were then used in TREC6, TREC7, TREC8, and TREC9. The best parameters obtained with CDS14 were then used in CDS15 and CDS16. The values of the parameters after optimisation are shown in Table 4.2. It is interesting to note the low values of α after the optimisation. This shows that feedback models are able to come up with a strong reweighting of the query without relying on its original text.

Finally, for the parameters we have not tuned, we used the following values. For reranking, we used the retrieval algorithm based on the KLD divergence (see Equation 4.1) with Dirichlet priors smoothing with the parameter μ set to 1000. For DMM, we set the parameter λ to 0.03 and the additive smoothing parameter γ for the estimation of $p(w|\theta_d)$ to 1, as recommended (Hazimeh and Zhai 2015).

4.4.5 *Results and Discussion*

In this section, we report and then discuss the results we have obtained. Overall, these results acknowledge our proposal as a cost-effective alternative to existing methods. In terms of recall, we were able to improve the rate at

which relevant documents appear in the adjudication. In terms of reliability, fairness, and reusability, our proposals obtain competitive results, similar to very strong baselines. Thus, our methods are able to improve the adjudication process in terms of assessor's effort, since relevant documents appear earlier, without sacrificing the reliability, the fairness and the reusability of the created relevance judgements.

4.4.5.1 *Recall*

We have summarised the evaluation of Recall in Table 4.3. In this table, we report the averaged AUC at different levels of judgements per topic for each evaluated method. We also flag the statistically significant improvements according to the randomised version of Tukey HSD test (Carterette 2012; Sakai 2018).

We are going to analyse the results obtained in the TREC5-9 collections and in the CDS14-15 collections separately, since we observe different trends in both. TREC5-9 collections are deep pooled collections, while CDS14-15 used shallow pools.

Regarding TREC5-9 collections, we first observe that top-k pooling, the method used traditionally in TREC workshops, yields the worst figures, acknowledging that there is room for improvement in developing adjudication methods, at least in terms of recall. We observe that real relevance feedback is a useful tool for improving the rate of relevant documents when building new relevance judgements. In particular, we see that the combination of relevance feedback with a dynamic adjudication policy (Algorithm 4.2) always gets best results than these policies alone. In other words, MTF+DMM improves over MTF, and MM+DMM also improves over MM in terms of recall. In fact, MM+DMM yields the best recall figures in every case (for every budget in TREC5-9 collections). Conversely, we see that using relevance feedback for reranking the entire pool, instead of reranking per-submission, is not a very competitive approach, since it rarely improves over the baselines.

Results for CDS14-15 collections have some common trends. Here, as before, MTF+DMM improves over MTF, and MM+DMM also improves over MM, further acknowledging the usefulness of real relevance feedback for document adjudication.

The most relevant result here is that, differently from TREC5-9 datasets, DMM yields very competitive results. In fact, it yields the best figures in every case, being significantly better than every baseline. We lucubrate that this is due to the fact that these collections used shallow pools (depth-10). With deeper pools, MTF and bandit based methods are able to perform more ex-

Table 4.3: Averaged Recall's AUC at different budgets per topic. Statistically significant improvements w.r.t. top-k, MTF, MM, DMM, MTF+DMM, MM+DMM and MMNS+DMM are superscripted with a, b, c, d, e, f and g, respectively. These are also added at the beginning of each line to ease the comparison. For each collection and budget, best figures are **bolded** and worst ones are underlined.

	Judgements per topic				Judgements per topic			
	100	300	750	1000	100	300	750	1000
	TREC5 (train)				**TREC6** (test)			
(a) top-k	0.23	0.35	0.49	0.54	0.17	0.26	0.32	0.35
(b) MTF	0.27^{d}	0.42^{a}	0.58^{a}	0.64^{a}	0.39^{d}	0.56^{a}	0.71^{a}	0.76^{a}
(c) MM	0.27^{d}	0.45^{ad}	0.64^{a}	0.69^{a}	$\mathbf{0.42}^{ad}$	$\mathbf{0.62}^{ad}$	0.77^{a}	$\mathbf{0.82}^{a}$
(d) **DMM**	0.21	0.40^{a}	0.61^{a}	0.67^{a}	0.31^{a}	0.51^{a}	0.70^{a}	0.77^{a}
(e) **MTF+DMM**	0.25	0.41^{a}	0.59^{a}	0.65^{a}	0.40^{ad}	0.58^{a}	0.75^{a}	0.80^{a}
(f) **MM+DMM**	$\mathbf{0.29}^{ad}$	$\mathbf{0.48}^{abde}$	$\mathbf{0.66}^{abe}$	$\mathbf{0.71}^{abe}$	$\mathbf{0.42}^{ad}$	$\mathbf{0.62}^{ad}$	$\mathbf{0.78}^{a}$	$\mathbf{0.82}^{a}$
(g) **MMNS+DMM**	0.27^{d}	0.46^{ad}	0.65^{abe}	$\mathbf{0.71}^{abe}$	0.39^{ad}	0.59^{a}	0.76^{a}	0.81^{a}
	TREC7 (test)				**TREC8** (test)			
(a) top-k	0.18	0.28	0.37	0.40	0.16	0.25	0.34	0.40
(b) MTF	0.34^{ad}	0.54^{a}	0.71^{a}	0.77^{a}	0.35^{ad}	0.54^{a}	0.71^{a}	0.76^{a}
(c) MM	0.39^{ad}	0.60^{ad}	0.77^{a}	0.82^{a}	0.42^{adeg}	0.64^{abdg}	0.79^{ad}	0.83^{a}
(d) **DMM**	0.27^{a}	0.49^{a}	0.71^{a}	0.77^{a}	0.26^{a}	0.48^{a}	0.69^{a}	0.75^{a}
(e) **MTF+DMM**	0.35^{ad}	0.57^{a}	0.76^{a}	0.81^{a}	0.33^{a}	0.55^{a}	0.74^{a}	0.79^{a}
(f) **MM+DMM**	$\mathbf{0.41}^{abdg}$	$\mathbf{0.63}^{abd}$	$\mathbf{0.79}^{a}$	$\mathbf{0.83}^{a}$	$\mathbf{0.43}^{abdeg}$	$\mathbf{0.65}^{abdeg}$	$\mathbf{0.81}^{ad}$	$\mathbf{0.84}^{a}$
(g) **MMNS+DMM**	0.34^{ad}	0.57^{a}	0.76^{a}	0.81^{a}	0.32^{a}	0.54^{a}	0.74^{a}	0.8^{a}
	TREC9 (test)				**CDS14** (train)			
(a) top-k	0.22	0.36	0.47	0.50	0.12	0.27	0.51	0.54
(b) MTF	0.35^{a}	0.54^{a}	0.71^{a}	0.76^{a}	0.17	0.37	0.64^{a}	0.66^{a}
(c) MM	0.34^{a}	0.57^{a}	0.76^{a}	0.81^{a}	0.19	0.45^{a}	0.71^{a}	0.73^{a}
(d) **DMM**	0.33^{a}	0.56^{a}	0.76^{a}	0.81^{a}	$\mathbf{0.33}^{abcfg}$	$\mathbf{0.63}^{abcf}$	$\mathbf{0.81}^{abc}$	$\mathbf{0.83}^{abc}$
(e) **MTF+DMM**	0.36^{a}	0.57^{a}	0.75^{a}	0.80^{a}	0.27^{abc}	0.53^{ab}	0.76^{ab}	0.78^{ab}
(f) **MM+DMM**	$\mathbf{0.38}^{a}$	$\mathbf{0.61}^{a}$	$\mathbf{0.79}^{a}$	$\mathbf{0.83}^{a}$	0.25^{ab}	0.50^{ab}	0.74^{ab}	0.76^{ab}
(g) **MMNS+DMM**	0.36^{a}	0.60^{a}	0.78^{a}	$\mathbf{0.83}^{a}$	0.25^{a}	0.53^{ab}	0.76^{ab}	0.78^{ab}
	CDS15 (test)				**CDS16** (test)			
(a) top-k	0.15	0.35	0.56	0.56	0.12	0.28	0.52	0.53
(b) MTF	0.20	0.45	0.66^{a}	0.66^{a}	0.15	0.35	0.61^{a}	0.62^{a}
(c) MM	0.22	0.48^{a}	0.69^{a}	0.69^{a}	0.17	0.42^{a}	0.67^{a}	0.68^{a}
(d) **DMM**	$\mathbf{0.30}^{abc}$	$\mathbf{0.59}^{abc}$	$\mathbf{0.76}^{abc}$	$\mathbf{0.76}^{abc}$	$\mathbf{0.23}^{abc}$	$\mathbf{0.51}^{abc}$	$\mathbf{0.73}^{ab}$	$\mathbf{0.74}^{ab}$
(e) **MTF+DMM**	0.25^{a}	0.54^{a}	0.73^{ab}	0.73^{ab}	0.20^{a}	0.46^{ab}	0.70^{ab}	0.70^{ab}
(f) **MM+DMM**	0.26^{a}	0.52^{a}	0.71^{a}	0.72^{a}	0.19^{a}	0.45^{ab}	0.68^{ab}	0.69^{ab}
(g) **MMNS+DMM**	0.25^{a}	0.54^{a}	0.72^{a}	0.72^{a}	0.20^{a}	0.46^{ab}	0.69^{ab}	0.70^{ab}

ploration and thus the estimations of the performance of each submission are more refined, providing, at the end, better results. Conversely, with shallower pools, submissions are exhausted before these methods can elaborate finer distributions, and thus their performance is superseded by DMM.

4.4.5.2 *Reliability*

Reliability results are summarised in Table 4.4. This table shows the number of judgements per topic needed to reach a value equal or higher than 0.90 of Kendall's τ and τ_{AP} correlations. The second correlation, τ_{AP}, was designed with the idea that errors in high positions are worse when comparing two rankings of search systems than errors in deeper positions. Kendall's τ penalises both errors the same. Typically, when evaluating search systems, we aim to find the best ones, those a top positions. Thus, we think that τ_{AP} correlation is a good measure to evaluate the adjudicating strategies we develop here. Note that we do not perform a statistical test in this case since it is a global metric. Additionally, although we report the results on the training datasets here, we only include those on the test ones in the following comments.

Overall, these results acknowledge that we need very few judgements to obtain strong correlations with the official ranking of runs. In particular, if our focus is Kendall's τ, the minimum number of judgements varies between 40 and 102. On the other hand, if we focus on τ_{AP} correlation, this number lies between 90 and 269. These figures are tiny in comparison with the average size of the pool in the original qrels (see Table 4.1). This means we can greatly reduce the assessor effort and still produce a reliable benchmark.

When comparing our methods against the baselines, we observe the following results. The method using relevance feedback solely (DMM) or its combination with MTF (MTF+DMM), MM (MM+DMM) or MMNS (MMNS+DMM) achieve the best results on TREC5, TREC7, TREC9, CDS14, CDS15 and CDS16 (that is 6 out of 8 test collections). On the other hand, on TREC6 and TREC8, MM and MTF are the best performing methods.

The recall and reliability results we have presented above acknowledge that these algorithms are a good choice if we aim to find relevant documents early, and we can build reliable benchmarks with lower assessor costs. Nonetheless, we do not know if we are harming the fairness and reusability of the constructed collection. In the following sections, we analyse to what extent this effect exists.

Table 4.4: Minimum number of judgements per topic needed to obtain values of 0.90 for Kendall's τ correlation and τ_{AP} correlation. For each collection and correlation, best values (the lowest) are **bolded** and worst ones are underlined.

	$\tau \geq 0.90$	$\tau_{AP} \geq 0.90$	$\tau \geq 0.90$	$\tau_{AP} \geq 0.90$
	TREC5 (train)		**TREC6** (test)	
top-k	85	393	250	287
MTF	48	196	138	138
MM	52	**96**	**102**	**115**
DMM	148	399	288	376
MTF+DMM	120	382	174	214
MM+DMM	**44**	115	123	154
MMNS+DMM	100	327	166	190
	TREC7 (test)		**TREC8** (test)	
top-k	137	238	70	246
MTF	80	166	48	**103**
MM	78	126	**40**	208
DMM	134	274	201	439
MTF+DMM	74	122	75	252
MM+DMM	**59**	**90**	41	104
MMNS+DMM	92	157	117	250
	TREC9 (test)		**CDS14** (train)	
top-k	78	258	233	368
MTF	48	**146**	127	193
MM	162	274	151	203
DMM	131	289	**53**	217
MTF+DMM	48	168	63	**130**
MM+DMM	**41**	220	162	191
MMNS+DMM	65	187	77	146
	CDS15 (test)		**CDS16** (test)	
top-k	204	280	219	347
MTF	124	**192**	216	282
MM	180	215	234	**269**
DMM	151	242	182	287
MTF+DMM	**85**	197	191	294
MM+DMM	136	212	215	**269**
MMNS+DMM	108	210	**138**	271

4.4.5.3 *Fairness*

We report fairness results in Figure 4.1. This figure shows, for each collection, the maximum drop in the position that a run suffers when comparing it with the official ranking. A high drop means that a submission is treated differently in the reduced pools, and thus those judgements are not entirely fair. Although this metric has some limitations (Otero, Parapar and Ferro 2023), it is common in the literature (Voorhees 2018; Voorhees, Craswell et al. 2022).

We observe diverse performances among collections and algorithms. Overall, it seems that MM+DMM is the best performing algorithm in most of the datasets, at leats from 300 judgements on. Also, in most cases, this algorithm performs better than its counterpart MM, further acknowledging the utility of relevance feedback for document adjudication.

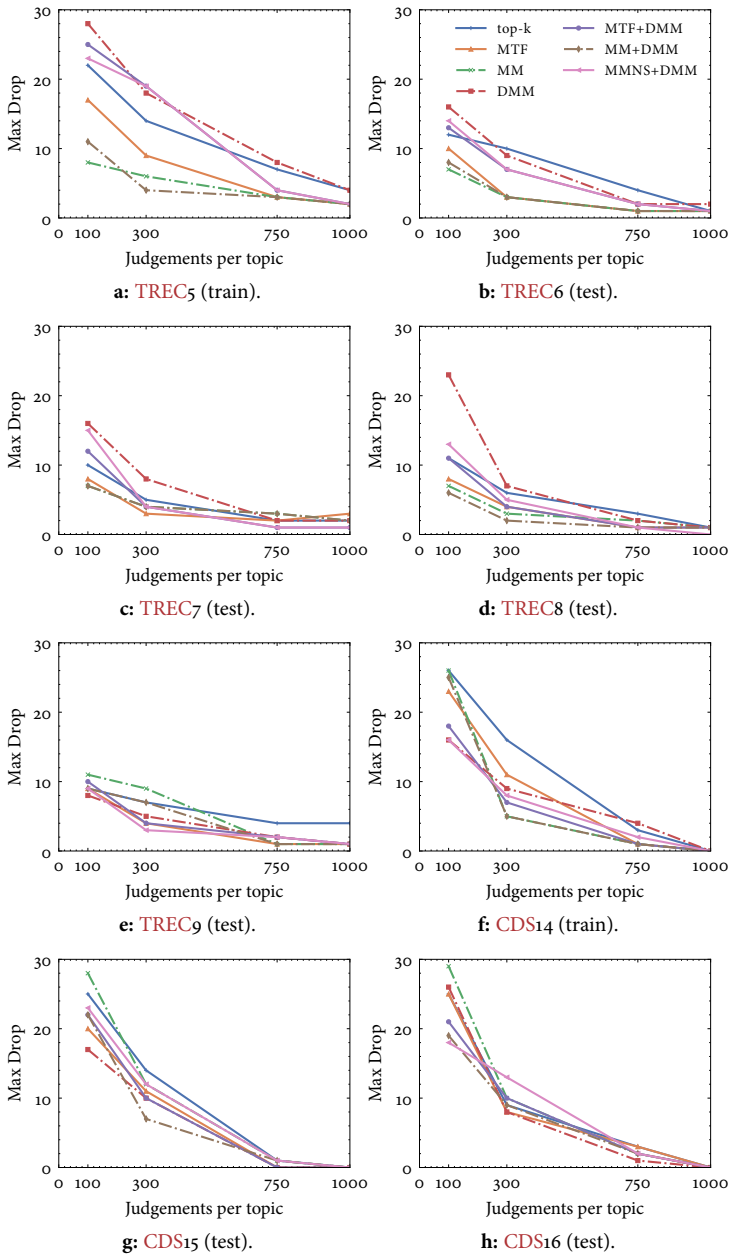
Following the same trends as before, DMM performs better in CDS collections than in others, although it is not better than MM+DMM for lower budgets.

4.4.5.4 *Reusability*

For evaluating the reusability, we performed a LOGO experiment. In this experiment, the unique, relevant documents retrieved by each team are removed from the pool. Then this pool is used to evaluate these runs that have been left out of the process. Then, we compute the correlation between the official ranking of runs and the ranking obtained with these reduced pool. If these correlations are high, we thus could conclude that the pools are reusable. We report the results of this experiment in Table 4.5. This table depicts the average of Kendall's τ correlation between ground-truth qrels—the official TREC judgements— and the qrels built with each adjudicating method when limiting the number of relevant documents per topic.

We observe overall high correlations, supporting the idea that there are no strong biases against non-pooled runs. We also continue to see similar trends as in previous experiments. Bandit-based methods, along with MTF, are benefited from the use of relevance feedback reranking. Also, DMM shows better performance on the shallow pooled datasets.

If we analyse these results along with the recall figures, we can argue that adjudication methods using relevance feedback are a well-performing alternative when gathering new assessments.



a: TREC5 (train). **b:** TREC6 (test).

c: TREC7 (test). **d:** TREC8 (test).

e: TREC9 (test). **f:** CDS14 (train).

g: CDS15 (test). **h:** CDS16 (test).

Figure 4.1: Values of MaxDrop for a varying number of judgements per topic.

Table 4.5: Average of LOGO Kendall's τ correlations. For each collection and budget, best figures are **bolded** and worst ones are underlined.

	Judgements per topic				Judgements per topic			
	100	300	750	1000	100	300	750	1000
	TREC5 (train)				**TREC6** (test)			
top-k	0.91	0.95	0.97	0.98	0.80	0.91	0.98	0.99
MTF	**0.93**	0.96	**0.99**	**0.99**	0.87	0.96	**0.99**	0.99
MM	**0.93**	0.96	**0.99**	**0.99**	**0.89**	**0.97**	**0.99**	**1.00**
DMM	0.87	0.94	0.98	**0.99**	0.81	0.91	0.98	0.99
MTF+DMM	0.90	0.95	0.98	**0.99**	0.84	0.94	**0.99**	0.99
MM+DMM	**0.93**	**0.97**	**0.99**	**0.99**	0.87	0.96	**0.99**	**1.00**
MMNS+DMM	0.90	0.94	**0.99**	**0.99**	0.85	0.94	**0.99**	**1.00**
	TREC7 (test)				**TREC8** (test)			
top-k	0.88	0.95	0.99	0.99	0.90	0.95	0.98	0.99
MTF	0.91	0.97	0.99	**1.00**	0.94	0.97	0.99	**1.00**
MM	0.92	0.97	0.99	**1.00**	**0.95**	0.98	0.99	**1.00**
DMM	0.87	0.95	0.98	0.99	0.85	0.94	0.99	**1.00**
MTF+DMM	0.92	0.97	**1.00**	**1.00**	0.91	0.96	**1.00**	**1.00**
MM+DMM	**0.94**	**0.98**	0.99	**1.00**	**0.95**	**0.99**	**1.00**	**1.00**
MMNS+DMM	0.91	**0.98**	**1.00**	**1.00**	0.89	0.96	**1.00**	**1.00**
	TREC9 (test)				**CDS14** (train)			
top-k	0.90	0.95	0.98	**0.99**	0.80	0.92	**0.99**	**0.99**
MTF	0.92	**0.97**	**0.99**	**0.99**	0.87	0.95	**0.99**	**0.99**
MM	0.89	0.94	**0.99**	**0.99**	0.85	0.96	**0.99**	**0.99**
DMM	0.89	0.95	**0.99**	**0.99**	**0.92**	0.97	**0.99**	**0.99**
MTF+DMM	0.92	0.96	0.98	**0.99**	0.91	**0.98**	**0.99**	**0.99**
MM+DMM	0.92	0.95	**0.99**	**0.99**	0.86	0.97	**0.99**	**0.99**
MMNS+DMM	**0.93**	**0.97**	**0.99**	**0.99**	0.90	0.97	**0.99**	**0.99**
	CDS15 (test)				**CDS16** (test)			
top-k	0.82	0.94	**0.99**	**1.00**	0.80	0.93	**0.99**	**0.99**
MTF	0.88	0.96	**0.99**	0.99	0.81	0.93	**0.99**	**0.99**
MM	0.82	0.96	**0.99**	0.99	0.76	0.94	**0.99**	**0.99**
DMM	0.88	0.95	**0.99**	0.99	**0.87**	**0.95**	**0.99**	**0.99**
MTF+DMM	**0.91**	**0.97**	**0.99**	0.99	0.84	0.94	**0.99**	**0.99**
MM+DMM	0.84	0.96	**0.99**	0.99	0.80	0.94	**0.99**	**0.99**
MMNS+DMM	0.89	0.96	**0.99**	0.99	**0.87**	**0.95**	**0.99**	**0.99**

4.5 RELATED WORK

In this chapter, we have formulated a new adjudication method that uses relevance feedback to prioritize pooled documents when building new test collections.

Ours is not the first attempt to use relevance feedback for this task. Sanderson and Joho (2004) used a single system to rank the entire collection, not just the pooled documents, and judged documents in sets of 100, to then refine the ranking of the collection using DMM as the method for updating the query. They concluded that "a single system can generate a usable set of qrels, and that the process building qrels using a single system can be facilitated by the relevance feedback". Previously, Cormack, Palmer et al. (1998) implemented what they called Interactive Searching and Judging (ISJ). The process consisted of leaving the assessor with no objective but trying to find as many relevant documents as possible for each topic, with reasonable effort. From the original article: "The usual strategy was to formulate a query and to judge the results of the query until the frequency of relevant judgements dropped to a level where continuing seemed fruitless. At this point, another query was formulated or the topic was abandoned". Assessors spent, on average, 2.1 hours per topic. Then, the authors compared these qrels with the official qrels from the TREC-6 track, obtaining a Kendall's τ correlation of 0.89 between both ranking of systems. Soboroff and Robertson (2003) employed a similar approach for building a collection for the TREC Filtering track of 2002 (Robertson and Soboroff 2002). The main difference is that, instead of manually updating the query, this update is made by a feedback process implemented in the pooled systems. In particular, they used four different systems to make a pool formed by the top 100 documents. Then, they used CombMNZ (Fox and Shaw 1993) to rank the documents in the pool, and the top 100 documents of this ranking were given to the assessor. Then, each system used these documents to rewrite the query and generate a second ranking, which generated a new pool to be assessed. This process was repeated five times. After the track was held and all the submissions were made, they examined the top 100 documents of each submission and judged the documents that were not judged in the previous round. A comparison between both sets of qrels yielded a correlation over 0.90, which led the authors to conclude that rankings of systems were "virtually identical."

4.6 CONCLUSIONS

RF is a successful technique to improve the performance of an existing IR system without modifying the underlying model. In this chapter, we have successfully introduced relevance feedback for prioritizing the pooled documents when building new test collections. There are several conclusions we can devise from the results we have seen. First, that real relevance feedback for document adjudication is able to improve the ratio of relevant documents, thus making a better use of resources for assessor's judgements. Second, that this improvement comes without lose on other very crucial aspects of the judgements: reliability, fairness and reusability.

Among all the relevance feedback models we have evaluated, namely, RM3, DMM, MEDMM, DMM clearly stayed above the rest. Also, we have seen that these models were able to provide strong query expansions with little information of the original query (i.e. the optimal interpolation parameter was near 0), and that long queries worked better than short queries.

Overall, these results acknowledge our proposal as cost-effective algorithms to build new test collections in scenarios where resources are scarce.

Part III

SIGNIFICANCE MATTERS

5

STATISTICAL SIGNIFICANCE OF DOCUMENT ADJUDICATION METHODS

We proposed a method for simulating diverse participant systems from which to pool the documents in Chapter 3. Then, in Chapter 4, we proposed a cost-effective adjudication method based on statistical relevance feedback models. In these two chapters, we evaluated our proposals following long-standing and well-established evaluation procedures that acknowledged their advantages.

In this chapter, we propose a new methodology for evaluating adjudication methods. Our proposal is focused on looking at how a given method preserves the real statistically significant differences between systems—instead of just measuring ranking swaps—. Our results show that our proposal is a more reliable evaluation method than existing ones, allowing us to gain new insights into adjudication methods that were not explored before. The results we are presented here have already been published (Otero, Parapar and Ferro 2023).

In Section 5.1, we explain the shortcomings of current evaluation procedures for adjudication methods. Then, in Section 5.2, we define a new method that aims to tackle some of these problems. In Section 5.3, we perform a throughout evaluation of our method, and then we extensively discuss the implications of the presented results, some of which were never devised before. Finally, in Sections 5.4 and 5.5, we describe related work and a final discussion on the conclusions of this chapter.

5.1 INTRODUCTION

IR is a field with a strong focus on evaluation (Harman 2011; Voorhees 2002). The main purpose is to empirically measure the effectiveness of retrieval systems using test collections under controlled conditions. These collections consist of a corpus of documents, topics, and relevance judgements (Sanderson 2010; Voorhees and Harman 2005). As we have seen in previous chapters,

acquiring the assessments for creating these collections is costly since human experts have to judge the documents' content and decide which ones are relevant for each topic.

Consequently, when larger collections arose, there was the need to implement some kind of *sampling* so that assessors would not have to judge the relevance of each document for each topic. However, simple random sampling, the most immediate approach, would not work since the number of relevant documents for a topic is extremely small compared to the corpus of documents. Thus, a random sample would end up consisting of (almost all) non-relevant documents. The first solution to this problem was the *pooling* technique implemented by Text Retrieval Conference (TREC) (Spärck Jones and Rijsbergen 1975; Voorhees and Harman 2005). Top-k pooling builds on the assumption that IR systems try to push relevant documents towards the top of the ranking, and thus, there is a good chance to pool most of the relevant documents for a topic, provided that k is deep enough and the pooled systems are diverse enough. The number of judgements that an assessor can perform, i.e. the budget, is limited and, therefore, there is a trade-off with the depth k of the pool and the number of pooled systems, since the more they grow, the higher the number of documents in the pool.

However, collections kept growing in size, and just judging deep pools over a diverse set of systems stopped being practicable as well (Voorhees, Craswell et al. 2022). Therefore, much work has focused on developing alternative methods to select better which documents to pool and judge by performing some sort of *focused sampling*, aimed at picking documents with higher chances of being relevant and better employing the assessor budget or allowing for lower budgets at a comparable quality (Losada, Parapar et al. 2017; Rahman et al. 2019). A method that *actively* decides which document to judge next is called an adjudication method. However, alternative prioritisation models may introduce biases or incompleteness in the judgements, hampering the future reusability of a test collection (Voorhees 2018).

Pooling does not guarantee finding all the relevant documents for a topic but, as said, it strives to find a very good share of them. Most of the time, we are not interested in the absolute value that a metric may give to a system—for most popular metrics, this would demand finding all the relevant documents—but in a good estimate of system performance that allows us to reliably distinguish between systems. Therefore, the quality of a pool is *traditionally* measured on its ability to *fairly rank* systems, i.e. to fairly compare them. This is not limited to the systems that were actually pooled, but it should also hold for systems that were not pooled (Zobel 1998), to ensure the future reusability of a test collection with new systems.

For the same reasons, the quality of new adjudication methods is *traditionally* assessed by checking that they rank systems as closely as possible to the full set of judgements of a (good quality) top-k pool, ensuring that they can still properly answer the question "is system A better than system B?". This is quantified by computing the correlation, e.g. Kendall's τ (Kendall 1938, 1948), between the ranking of systems produced under the qrels gathered by an adjudication method and the ranking of systems produced under the qrels of the full top-k pool. The rationale is that if this correlation is high, one may assume the validity of the new method and aim to use it in the future for building new test collections at a comparable quality but with a lower assessment cost.

However, the question researchers are really interested in is "is system A *statistically significantly* better than system B?", since this ensures that observed differences are not due just to the randomness present in the construction process of a collection and, especially, that the found differences would *generalise* better and still hold in operational settings (Fuhr 2018; Sakai 2021). The problem is that the above correlation measures ignore whether the evaluated systems' statistical significance is preserved.

Let us better explain this problem with an example. Let us assume we have three different IR systems, `Sys1`, `Sys2` and `Sys3`, and that their true ranking, given by the full top-k pool, is (`Sys1`, `Sys2`, `Sys3`). We perform a significance test between all possible pairwise comparisons and we obtain that `Sys1` is significantly better than `Sys2` and `Sys3`, and `Sys2` is also significantly better than `Sys3`. Then, we create a new set of judgements using some adjudication method and repeat the above procedure. Using this new pool, we find the same ranking of systems as when using the full top-k pool, leading to a perfect correlation and concluding that the adjudication method is fully equivalent, but less costly, than the full top-k pool. However, we do not know anything about the significance between systems. If we repeat the same significance test using the new pool instead, we may not find any significant difference between any pair. We may thus conclude that there is no evidence of any system being different from the rest. This would be the opposite conclusion to the one drawn on the full top-k pool, where all the system pairs were significantly different.

In this chapter, our contributions are two-fold. First, we propose a new approach to evaluate the validity of low-cost adjudication methods, focusing on how they preserve the statistically significant differences between systems. Second, we analyse some state-of-the-art adjudication methods using our new approach to gain new insights about them.

5.2 STATISTICAL SIGNIFICANCE OF ADJUDICATION METHODS

Now that we have outlined the problems of traditional evaluation of adjudication methods, we proceed to define and formalize our proposed method.

Let $S = \{s_i\}$ ($|S| = n$) be the set of systems under experimentation, and let G be the *gold assessments* of the full top-k pool. Using an effectiveness measure of choice, we compute the per-topic scores for each of the n systems, and we perform a statistical test for each pairwise comparison between systems. From this test, we obtain, for each pair of systems s_i and s_j ($i < j \leq n$), a triplet $\langle s_i, s_j, c \rangle$, where $c \in \{>, \gg, <, \ll\}$, denoting the four outcomes we are interested in: s_i is better than s_j ($s_i > s_j$), s_i is *significantly* better than s_j ($s_i \gg s_j$), s_j is better than s_i ($s_i < s_j$), or s_j is *significantly* better than s_i ($s_i \ll s_j$).

We use R_G to denote the set of triplets that result from the statistical test performed using the gold qrels. Similarly, we use L to denote the qrels obtained with a low-cost adjudication method ($L \subseteq G$) andR_L to denote the set of triplets that result from the statistical test performed with them. Note that $|R_G| = |R_L| = \frac{n(n-1)}{2}$. Finally, we use T_G to denote the set of comparisons from R_G that are significantly different,that is, the set triplets for which $c \in \{\ll, \gg\}$, and T_L for the significantly different comparisons obtained with the low-cost assessments.

As we already explained, we are interested in studying to what extent the judgements produced by different low-cost adjudication methods preserve the statistically significant differences between systems we observe when using the gold qrels. The idea here is that if the low-cost method is able to preserve such differences, we could confidently use it to build new collections in the future with fewer assessment costs. Thus, we compare how T_G and T_L agree with each other using the measures described hereunder.

5.2.1 *Kendall's τ*

Kendall's τ is the measure traditionally used to evaluate adjudication methods. It computes the correlation between the ranking of systems under the gold qrels setting and the one under the qrels produced with the different adjudication methods.

Given two rankings over the same set of items, Kendall's τ computes how many items are swapped as:

$$\tau = \frac{(P - Q)}{\binom{n}{2}} \tag{5.1}$$

where P is the number of concordant pairs (pairs of systems ranked in the same relative order in both lists), Q is the number of discordant pairs (swapped pairs of systems), and $\binom{n}{2} = \frac{n(n-1)}{2}$ is the number of total pairs, given that we have n items.

5.2.2 *Precision and Recall*

We consider the *Precision* (P) and *Recall* (R) of the significantly different pairs detected by the low-cost adjudication methods, defined as follows:

$$P = \frac{|T_G \cap T_L|}{|T_L|} \quad R = \frac{|T_G \cap T_L|}{|T_G|} \tag{5.2}$$

where $|T_G \cap T_L|$ is the number of significantly different pairs common to both the gold and adjudication qrels, i.e. the correct ones when assuming the gold qrels detect the true differences. Precision indicates how much *noise* is introduced by an adjudication method, meant as additional significant differences not detected by gold qrels; Recall indicates how many of the total possible significant differences are not detected by an adjudication method.

5.2.3 *Agreements*

We consider an adaptation of a series of agreement measures that have been used in past work (Faggioli and Ferro 2021; Ferro and Sanderson 2022; Moffat, Scholer et al. 2012; Urbano, Marrero et al. 2013). Note that, while Kendall's τ and Precision/Recall focus on ranking of systems (the former) or on matching significantly different pairs (the latter) in isolation, the following agreement measures consider them jointly.

- **Active Agreements (AA)**: the set of consistent outcomes between both methods. This is, $\langle s_i, s_j, \gg \rangle \in T_G$ and $\langle s_i, s_j, \gg \rangle \in T_L$ or $\langle s_i, s_j, \ll \rangle \in T_G$ and $\langle s_i, s_j, \ll \rangle \in T_L$. This is the best possible case, and thus, the larger AA are, the better.

- **Active Disagreements (AD)**: the set of opposite outputs between both methods. This is, $\langle s_i, s_j, \gg \rangle \in T_G$ and $\langle s_i, s_j, \ll \rangle \in T_L$, or $\langle s_i, s_j, \ll \rangle \in T_G$ and $\langle s_i, s_j, \gg \rangle \in T_L$. This is the worst possible case, since it means that both methods reach complete opposite conclusions for a given pair. Thus, the lesser, the better.

- **Mixed Agreements (MA)**: we have four possible options: **(i)** $\langle s_i, s_j, \ll \rangle \in T_G$ and $\langle s_i, s_j, < \rangle \in T_L$; **(ii)** $\langle s_i, s_j, \gg \rangle \in T_G$ and $\langle s_i, s_j, > \rangle \in T_L$;

(iii) $\langle s_i, s_j, < \rangle \in T_G$ and $\langle s_i, s_j, \ll \rangle \in T_L$, and **(iv)** $\langle s_i, s_j, > \rangle \in T_G$ and $\langle s_i, s_j, \gg \rangle \in T_L$ We distinguish between MA_G (**(i)** and **(ii)**), which counts the cases where the adjudication method was not able to see a true significant difference, and MA_L (**(iii)** and **(iv)**) counts the cases where a low-cost method sees a significant difference that is not in the gold qrels. Note that $\text{MA}_\text{G} + \text{MA}_\text{L} = \text{MA}$ (MA_total).

- **Mixed Disagreements (MD)**: we also have four possible cases here: **(i)** $\langle s_i, s_j, \ll \rangle \in T_G$ and $\langle s_i, s_j, > \rangle \in T_L$; **(ii)** $\langle s_i, s_j, \gg \rangle \in T_G$ and $\langle s_i, s_j, < \rangle \in T_L$; **(iii)** $\langle s_i, s_j, > \rangle \in T_G$ and $\langle s_i, s_j, \ll \rangle \in T_L$, and **(iv)** $\langle s_i, s_j, < \rangle \in T_G$ and $\langle s_i, s_j, \gg \rangle \in T_L$ Here, as with MA, we also distinguish between MD_G (**(i)** and **(ii)**) and MD_L (**(iii)** and **(iv)**). As before, note that $\text{MD}_\text{G} + \text{MD}_\text{L} = \text{MD}$ (MD_total).

5.2.4 *Bias*

Analogously to Ferro and Sanderson (2022), we also consider the *publication bias*, i.e. the likelihood of a researcher publishing a significant result using an adjudication method when in fact a significance test on the gold qrels would have produced either no significance (MA, MD) or a significant result in the opposite direction (AD). We define it as follows:

$$Bias = 1 - \frac{\text{AA}}{\text{AA} + \text{AD} + \text{MA}_L + \text{MD}_L} \tag{5.3}$$

Here, a value of 0% means that every significance detected by an adjudication method leads to the same conclusions (and publication) as those of the gold qrels. Conversely, a value of 100% means that every significance detected by an adjudication method leads to opposite conclusions (and publication) to those of the gold qrels. Thus, the lower the bias, the better. Note that, differently from Ferro and Sanderson (2022), we do not consider the whole MA and MD but just MA_L and MD_L, since we are interested only in the publication bias induced by the adjudication method. This metric tries to measure the situations where a researcher sees a significant outcome under the reduced pools when, in reality, it would be a different conclusion under the gold qrels.

5.2.5 *Family-Wise Error Rate*

Performing *multiple comparisons*—in our case between each pair of systems—leads to an increase of the *Type I error*, i.e. incorrectly rejecting the null hypothesis, and inflates the number of significant differences found (Hochberg and Tamhane 1987; Hsu 1996; Sakai 2018). The Type I error probability is

equal to the significance level α and, as the number of comparisons increases, this probability also does. If we perform k different system comparisons, the probability of correctly accepting the null hypothesis for all of them is equal to $(1-\alpha)^k$. Thus, the probability of committing at least one Type I error is $1-(1-\alpha)^k$. This is the family-wise error rate (FWER). If we have, for example, $\alpha = 0.05$ and $k = 6$ comparisons (4 systems, $\frac{4(4-1)}{2} = 6$), this probability would rise to 0.264, which is not acceptable. For this reason, when we perform multiple comparisons, we should employ a technique to adjust the p-values, so that the FWER stays below α. Obviously, this has the side effect of reducing the *power* of the statistical test and increasing the number of *Type II errors*, i.e. not detecting an actual significant difference.

There are several options to control the FWER in a multiple comparison situation. The Bonferroni correction, for example, is a post-hoc correction where, if we have k different comparisons, we should use $p < \frac{\alpha}{k}$ as our significance level in each pairwise comparison. However, the Bonferroni correction is known to be too conservative and to reduce the power of a test too much, especially when the number of comparisons increases as in our case. Therefore, we employ the randomized version of the Tukey Honestly Significant Difference (HSD) test (Carterette 2012; Sakai 2018). This is a non-parametric computer-based generalization of the common permutation test for handling more than 2 systems. At each permutation, the test permutes the array of system scores of each topic independently. After this perturbation, it computes the difference between the maximum and minimum average system scores. Then, the test counts how many times the actual differences between system average performance is lower than this permuted mean difference (d' in Algorithm 5.1). The Tukey HSD test produces a p-value for each pairwise comparison, which is the ratio of times the permuted mean is higher than the actual difference between systems with respect to the number of permutations performed. These p-values can be compared to the significance level α to decide whether that pair of systems is significantly different or not. Algorithm 5.1, adapted from prior work (Carterette 2012; Sakai 2018), shows the details of our implementation.

5.3 EXPERIMENTS

To evaluate the validity of our proposal, we perform a series of experiments in different TREC collections. We now give the details of our experimental settings and then discuss our results.

Algorithm 5.1 Paired Randomized Tukey Honestly Significant Difference test

Input:

X $m \times n$ topic-system scores matrix.
B number of permutations.

Output:

P $n \times n$ matrix holding a p-value for each pairwise system comparison.

for $k \leftarrow 1$ to B **do**
  create $m \times n$ matrix X'
  for each topic t **do**
    row t of $X' \leftarrow$ permutation of values in row t of X
  $d' \leftarrow \max_i \bar{X}'_i - \min_j \bar{X}'_j$ ▷ $\bar{X}'_i$ *is the mean of column* i
  for each pair of systems i, j **do**
    if $d' > |\bar{X}_i - \bar{X}_j|$ **then**
      $P_{i,j} \leftarrow P_{i,j} + \frac{1}{B}$

5.3.1 *Collections*

We employ the TREC8 ad hoc collection, known to have a very high-quality pool (Voorhees and Harman 2000; Voorhees, Soboroff et al. 2022). It includes 129 runs (system submissions), retrieving 1000 documents for each topic, and 50 topics. Official relevance judgements are based on a pool of depth 100 over 71 out of 129 submitted runs, resulting in 86 830 assessments across all 50 topics. The average pool size per topic is 1736, while the maximum and the minimum are 2992 and 1046, respectively. Additionally, we use the collection from the document ranking task of TREC 2021 Deep Learning track (Craswell, Mitra, Yilmaz, Campos and Lin 2021), which adopted a shallow pooling approach at depth 10, then enlarged with a method based on active learning. With the DL21 dataset, we used only the documents in the top-10 pools as our gold qrels to provide a fairer comparison to the case of TREC8. It includes 66 runs, retrieving 100 documents for each topic, and 13 058 judgements made by NIST assessors over 57 different topics. The depth-10 pools we used include 6510 judgements, with an average pool size of 114. The maximum pool size is 226 and the minimum is 50.

5.3.2 *Compared Methods*

We consider a series of state-of-the-art adjudication methods.

- **top-k pooling.** We adapt the standard method used in TREC to limited-budget situations. When limiting the budget of assessments, we choose a k deep enough to fill that budget. Then, pooled documents are sorted by their document identifier (Voorhees and Harman 2005).
- **MTF.** It is a dynamic adjudication method proposed by Cormack and colleagues (Cormack, Palmer et al. 1998) that has been acknowledged as a robust adjudication method (Altun and Kutlu 2020).
- **MM, MMNS, TS and TSNS.** Bandit-based methods for document adjudication apply Bayesian principles to formalize the uncertainty associated with the probabilities of pulling a positive reward (a relevant document) from playing a bandit (Losada, Parapar et al. 2016).
- **Hedge.** Hedge is an online learning algorithm adapted for pooling (Aslam et al. 2003). A more detailed explanation of applying Hedge for document adjudication can be found in this article (Losada, Parapar et al. 2017).
- **NTCIR top-k prioritization.** It is the method used in NTCIR workshops (Sakai, Kando et al. 2008). Documents in the pool are sorted by the number of runs that contain the document at or above the depth k (the higher, the better), ties are solved with the sum of the ranks of that document within the runs (the smaller, the better) (Sakai, Kando et al. 2008).

5.3.3 *Other Settings*

We used AP (Buckley and Voorhees 2005) and NDCG (Järvelin and Kekäläinen 2002) as performance measures to score runs. We used $\alpha = 0.05$ as significance level and $B = 1\,000\,000$ permutations in Tukey HSD test (see Algorithm 5.1). Finally, since MTF, MM, MMNS, TS, and TSNS have a stochastic nature, the reported results for those methods are averaged over 50 executions of each.

5.3.4 *Preservation of Significant Differences*

In Tables 5.1 and 5.2, we report the Kendall's τ, Precision and Recall, as defined in Section 5.2, that each adjudication method achieves, while varying the

Table 5.1: Values of Kendall's τ, Precision and Recall (see Section 5.2) of each adjudication method for a varying number of judgements per topic. Parentheses indicate the size of this budget with respect to the full pool. We used the 71 pooled systems of TREC8, and AP for computing performance scores.

Method	**Budget: 100** (6%)			**Budget: 300** (17%)		
	τ	**P**	**R**	τ	**P**	**R**
top-k	0.91	0.932	0.888	0.95	0.955	0.955
MTF	0.94	0.946	**0.961**	0.97	0.962	0.980
MM	**0.95**	0.948	0.958	**0.98**	**0.969**	**0.992**
MMNS	0.93	0.942	0.957	0.97	0.967	0.987
TS	**0.95**	0.947	0.954	**0.98**	**0.969**	0.991
TSNS	0.93	0.945	0.949	0.97	0.966	0.983
Hedge	0.94	**0.955**	0.947	**0.98**	0.968	0.980
NTCIR	0.83	0.900	0.876	0.96	0.942	0.925

Table 5.2: Values of Kendall's τ, Precision and Recall (see Section 5.2) of each adjudication method for a varying number of judgements per topic. Parentheses indicate the size of this budget with respect to the full pool. We used the 71 pooled systems of TREC8, and NDCG for computing performance scores.

Method	**Budget: 100** (6%)			**Budget: 300** (17%)		
	τ	**P**	**R**	τ	**P**	**R**
top-k	0.90	**0.975**	0.929	0.94	0.985	0.970
MTF	0.91	**0.975**	0.953	**0.96**	0.982	0.985
MM	**0.92**	0.942	0.973	**0.96**	0.976	**0.991**
MMNS	0.90	0.970	0.962	**0.96**	**0.986**	**0.991**
TS	**0.92**	0.940	0.970	**0.96**	0.975	0.990
TSNS	0.90	0.971	0.960	**0.96**	0.985	**0.991**
Hedge	**0.91**	0.959	**0.978**	0.95	0.972	0.989
NTCIR	0.81	0.961	0.942	0.93	0.977	0.988

number of assessments per topic. We report the scores for 100 judgements per topic (which is a 6% budget of the original pool), and 300 (17%). All these values were obtained using the pooled systems of the TREC8 collection, which includes 71 different systems.

Regarding Kendall's τ and consistently with previous findings in the literature, we see almost every method achieves a very high correlation ($\tau > 0.90$) already at a 6% of the original budget. While this means that every method obtains a ranking of systems very similar to the one of the gold qrels, it also makes it very difficult to distinguish among methods. Moreover, we can observe that top-k and NTCIR methods stay behind the rest, leaving room for improvement in developing more efficient adjudication strategies for building new collections in evaluation workshops.

As we mentioned earlier, Kendall's τ does not allow us to know whether the compared algorithms preserve the same statistically significant differences as the gold qrels. Therefore, we study to which extent this effect might hold by using the Precision and Recall measures previously introduced.

We observe that every method obtains Precision and Recall values over 90% in almost all cases, which is a quite solid result. Moreover, every method is able to mostly preserve the same differences just having a 6% of the original budget. With 300 assessment per topic (17% of the budget), Recall is (almost) 1.00 for most of the methods, indicating that they are able to detect all the significant differences of the gold qrels at less than one third of the cost.

It is also interesting to observe that most of them detect some differences that were not detected in the gold qrels. Indeed, Precision is lower than 1.00 while Recall is almost 1.00 (all the differences in the gold qrels detected). In other terms, T_L (the set of significant differences detected by the adjudication method) is not a proper subset of T_G (the set of significant differences detected by the gold qrels). A possible explanation might be that, since reduced pools lack some relevant documents, the performance difference of some pair of systems (delta AP/NDCG between the two systems in our case) turns out to be increased with respect to the gold qrels and this makes the pair significantly different on the reduced pool but not on the gold qrels. Since more evaluation on this issue would need more experimentation, we leave this investigation for future work.

To support a more detailed analysis, in Tables 5.3 and 5.4, we report the raw agreements of each method. Table 5.3 includes the results obtained when using AP for evaluating the runs. In this case, there are a total of 966 gold significant differences ($|T_G| = 966$). Table 5.4 includes the results when using NDCG for evaluating the runs. In this case, there are a total of 917 gold significant differences ($|T_G| = 917$).

Table 5.3: Values of relevants, agreements and bias of each adjudication method for a varying number of judgements per topic. Parentheses indicate the size of this budget with respect to the full pool. We used the 71 pooled systems of TREC8. The top-100 full pool includes 4728 relevant documents. There are 2485 pairwise comparisons, of which 966 are significant under the gold qrels when using AP to compute performance scores.

	Metric	Adjudication method							
		top-k	MTF	MM	MMNS	TS	TSNS	Hedge	NTCIR
AP (966 gold significantly different pairs) — Budget per topic: 100 (6%)	# rels.	1077	1685	2148	1553	2102	1514	**2170**	1481
	AA	858	**929**	926	925	922	917	915	846
	MA_{total}	170	**90**	**90**	98	94	102	94	185
	MA_G	108	**37**	40	41	44	49	51	91
	MA_L	62	52	50	57	50	53	**43**	94
	MD_{total}	**0**	**0**	1	**0**	1	**0**	**0**	29
	MD_G	**0**	**0**	**0**	**0**	**0**	**0**	**0**	29
	MD_L	**0**	**0**	1	**0**	1	**0**	**0**	**0**
	AD	0	0	0	0	0	0	0	0
	Bias	7%	5%	5%	6%	5%	5%	**4%**	10%
Budget per topic: 300 (17%)	# rels.	2042	2923	**3628**	2913	3607	2868	3609	2723
	AA	923	**961**	959	954	958	950	947	894
	MA_{total}	86	43	**38**	44	39	50	50	127
	MA_G	43	**5**	7	12	8	16	19	72
	MA_L	43	38	**30**	32	**30**	33	31	55
	MD_{total}	0	0	0	0	0	0	0	0
	MD_G	0	0	0	0	0	0	0	0
	MD_L	0	0	0	0	0	0	0	0
	AD	0	0	0	0	0	0	0	0
	Bias	4%	4%	**3%**	**3%**	**3%**	**3%**	**3%**	6%

Table 5.4: Values of relevants, agreements and bias of each adjudication method for a varying number of judgements per topic. Parentheses indicate the size of this budget with respect to the full pool. We used the 71 pooled systems of TREC8. The top-100 full pool includes 4728 relevant documents. There are 2485 pairwise comparisons, of which 917 are significant under the gold qrels when using NDCG to compute performance scores.

		Metric	Adjudication method							
			top-k	MTF	MM	MMNS	TS	TSNS	Hedge	NTCIR
NDCG (917 gold significantly different pairs)	Budget per topic: 100 (6%)	# rels.	1077	1685	2148	1553	2102	1514	**2170**	1481
		AA	852	874	893	883	890	881	**897**	864
		MA_{total}	86	65	79	61	83	62	**58**	88
		MA_G	65	43	24	34	27	36	**20**	53
		MA_L	**21**	22	55	27	56	26	38	35
		MD_{total}	0	0	0	0	0	0	0	0
		MD_G	0	0	0	0	0	0	0	0
		MD_L	0	0	0	0	0	0	0	0
		AD	0	0	0	0	0	0	0	0
		Bias	**2%**	**2%**	6%	3%	6%	3%	4%	4%
	Budget per topic: 300 (17%)	# rels.	2042	2923	**3628**	2913	3607	2868	3609	2723
		AA	890	904	**909**	**909**	**909**	**909**	907	906
		MA_{total}	40	29	30	**20**	31	21	36	32
		MA_G	27	13	**8**	**8**	**8**	**8**	10	11
		MA_L	13	16	22	**12**	22	13	26	21
		MD_{total}	0	0	0	0	0	0	0	0
		MD_G	0	0	0	0	0	0	0	0
		MD_L	0	0	0	0	0	0	0	0
		AD	0	0	0	0	0	0	0	0
		Bias	**1%**	2%	2%	**1%**	2%	**1%**	3%	2%

The AA counts confirm that adjudication methods are more effective than top-k and NTCIR pooling methods in detecting significant pairs in the correct order, especially at lower budgets. They provide further insights about the (almost) 1.00 Recall (see Tables 5.1 and 5.2) we observed for most adjudication methods. Indeed, with AP, the gold qrels detect 966 significantly different pairs and the AA counts is (almost) 966, indicating that the 1.00 Recall is due to significant pairs in the correct order. The same happens for NDCG, where we observe that most methods obtain AA values near 917. In other terms, the slight drop in Kendall's τ observed in the previous experiment is not caused by wrongly ordered pairs that were originally significantly different. When it comes to the specific methods, MTF achieves the best AA figures for budgets of 100, 300 when using AP, while under NDCG Hedge works slightly better with lower budgets and bandit-based methods perform the best with a budget of 300.

If we compare the AA counts with the number of relevant documents found by a method (the # rels. row), we observe a somehow unexpected behaviour. One might think that the more relevant documents found, the more AA increases. However, for a budget of 100 judgements per topic, Hedge adjudicated 2170 relevant documents, 485 more than MTF, but the latter one achieves the highest AA with AP; the same happens again for a budget of 300: MTF is not the best one in terms of relevant documents but it is the best in terms of AA. We can observe something similar with NDCG: finding more relevant documents does not necessarily mean more AA. Obviously, having more relevant documents in the pool helps in increasing the number of AA, but these results showcase that it is not the only factor. Overall, these observations suggest that not all the relevant documents are equally discriminative in finding significantly different pairs. Indeed, relevant documents appear at different ranks in the results lists and the same (or even higher) number of relevant documents may contribute differently to the performance score of a run and, in turn, to the significant differences found. So far, research has mostly focused on determining the number of topics needed (Buckley and Voorhees 2000; Sakai 2016b; Sanderson and Zobel 2005; Voorhees 2009; Voorhees and Buckley 2002) or on identifying the most discriminative subset of topics (Hauff et al. 2009; Hosseini et al. 2012; Mizzaro and Robertson 2007; Roitero et al. 2020). These findings open up the possibility of future research on which are the best relevant documents to more reliably discriminate among systems, an area not well explored yet, to the best of our knowledge.

Almost in every case, no method fails in a mixed or active disagreement, i.e. detecting significant differences when there is a swap. This represents a very

Table 5.5: Values of Kendall's τ, Precision and Recall (see Section 5.2) of each adjudication method for a varying number of judgements per topic. Parentheses indicate the size of this budget with respect to the full pool. We used the 66 pooled systems from DL21, and AP for computing performance scores.

Method	**Budget: 10** (9%)			**Budget: 30** (26%)		
	τ	**P**	**R**	τ	**P**	**R**
top-k	0.46	0.448	0.445	0.69	0.668	0.833
MTF	0.49	**0.611**	0.414	0.69	0.687	0.798
MM	**0.53**	0.566	0.477	**0.73**	**0.764**	0.778
MMNS	0.50	0.517	0.505	0.70	0.654	0.841
TS	0.52	0.554	0.489	**0.73**	0.761	0.777
TSNS	0.50	0.509	0.502	0.69	0.642	0.839
Hedge	0.42	0.430	0.419	0.50	0.558	0.603
NTCIR	0.47	0.423	**0.560**	0.69	0.594	**0.871**

important insight from this experiment, since it shows that no method causes a ranking swap between a pair of systems that were originally significantly different. In other terms, the drop in Kendall's τ is not due to swaps between systems that are significantly different on the gold qrels but swaps only happen among not significantly different systems, having a much lower impact.

Let us now consider MA_G and MA_L. The former accounts for significant pairs in the gold qrels which are missed by reduced pools; thus, it helps mainly to explain drops in Recall. The latter accounts for significant pairs in a reduced pool that are not present in the gold qrels; thus, it mainly helps explain drops in Precision. We can observe that MA_G gets reduced as the budget size increases up to almost 0, except for top-k pooling, Hedge and NTCIR method, consistently with the previous findings in Tables 5.1 and 5.2. Moreover, MA_L is consistently higher than MA_G, explaining the loss in Precision even at very high Recall levels.

When it comes to publication bias, we observe moderate values, from 7% and below, suggesting that all the methods would not lead to drawing conclusions severely different from the gold qrels. We can observe that bias quickly decreases as the budget increases and that adjudication methods are more effective than top-k pooling, achieving a bias up to 2-3 times lower than it.

Finally, we can observe that there are no different trends between the two evaluation metrics employed, AP and NDCG. This shows that the results presented here are not an artefact of the metric used, but of the adjudication methods being evaluated.

Table 5.6: Values of Kendall's τ, Precision and Recall (see Section 5.2) of each adjudication method for a varying number of judgements per topic. Parentheses indicate the size of this budget with respect to the full pool. We used the 66 pooled systems from DL21, and NDCG for computing performance scores.

Method	**Budget: 10** (9%)			**Budget: 30** (26%)		
	τ	**P**	**R**	τ	**P**	**R**
top-k	0.61	0.531	0.554	**0.82**	0.723	0.832
MTF	0.61	**0.632**	0.534	0.79	0.734	0.808
MM	**0.66**	0.628	0.598	0.81	0.772	0.808
MMNS	0.64	0.593	0.607	**0.82**	0.725	**0.844**
TS	**0.66**	0.624	0.605	**0.82**	**0.780**	0.809
TSNS	0.63	0.589	0.603	0.81	0.715	0.839
Hedge	0.51	0.521	0.484	0.61	0.657	0.674
NTCIR	0.59	0.522	**0.621**	0.76	0.669	0.827

Additionally, we run experiments on the TREC DL21. We selected this collection as having opposing characteristics to TREC8. The DL21 collection adopts a very shallow pooling at just depth 10, representing a quite challenging setting for adjudication methods. We believe that using these two collections helps in supporting the generalizability of the results presented here. Tables 5.5 and 5.6 report the Kendall's τ, Precision, and Recall, similarly to Tables 5.1 and 5.2 for TREC8; Tables 5.7 and 5.8 report the agreement counts, similarly to Tables 5.3 and 5.4 for TREC8. In general, we observe quite lower and much more varied performance on DL21 than on TREC8.

Kendall's τ is generally low for all the methods with both metrics. In TREC8, adjudication methods were able to obtain very strong results only with a 17% of the original budget, while in this case no method is able to reach that performance even with a 26%. One important difference is that, while in TREC8 top-k and NTCIR method were clearly underperforming with respect to the other methods, in DL21 Hedge clearly achieves the worst performance.

When it comes to the agreements (Tables 5.7 and 5.8), a notable difference is that, at low budgets (9%), MD appear while they go to (almost) zero for higher budgets. The MD at 9% budget indicate that the drop in Kendall's τ are also due to swaps in the significantly different pairs. The problem concerns more MD_L, i.e. swaps in significant pairs detected by a reduced pool but not the gold qrels, than MD_G, i.e. swaps in significant pairs detected by the gold qrels but not a reduced pool. As a consequence, part of the loss of Precision is due to swaps in the significant pairs a more severe condition than the one causing the loss of Precision in TREC8. This issue impacts more top-k and

Table 5.7: Values of relevants, agreements and bias of each adjudication method for a varying number of judgements per topic. Parentheses indicate the size of this budget with respect to the full pool. We used the 66 pooled systems from DL21. The top-10 pool includes 3541 relevant documents. There are a total of 2145 pairwise comparisons, of which 418 are significant under the gold qrels when using AP to compute performance scores.

	Metric	Adjudication method							
		top-k	MTF	MM	MMNS	TS	TSNS	Hedge	NTCIR
AP (418 gold significantly different pairs) — Budget per topic: 10 (9%)	# rels.	441	488	489	474	483	469	504	**513**
	AA	186	173	199	211	204	210	175	**234**
	MA_{total}	413	**345**	358	386	361	392	442	464
	MA_G	214	237	212	201	206	203	235	176
	MA_L	199	**108**	146	185	155	189	207	288
	MD_{total}	48	**13**	15	19	18	19	33	39
	MD_G	18	8	7	**6**	8	**6**	8	8
	MD_L	30	**5**	9	13	10	14	25	31
	AD	0	0	0	0	0	0	0	0
	Bias	55%	**39%**	43%	48%	45%	49%	57%	58%
AP (418 gold significantly different pairs) — Budget per topic: 30 (26%)	# rels.	1186	1327	**1359**	1289	1345	1267	1352	1337
	AA	348	334	325	352	325	351	252	**364**
	MA_{total}	243	237	**194**	251	196	262	355	299
	MA_G	70	84	93	66	93	67	161	**54**
	MA_L	173	152	**101**	185	103	194	194	245
	MD_{total}	**0**	**0**	**0**	1	**0**	1	11	4
	MD_G	**0**	**0**	**0**	**0**	**0**	**0**	5	**0**
	MD_L	**0**	**0**	**0**	1	**0**	1	6	4
	AD	0	0	0	0	0	0	0	0
	Bias	33%	31%	**24%**	35%	**24%**	36%	44%	41%

Table 5.8: Values of relevants, agreements and bias of each adjudication method for a varying number of judgements per topic. Parentheses indicate the size of this budget with respect to the full pool. We used the 66 pooled systems from DL21. The top-10 pool includes 3541 relevant documents. There are a total of 2145 pairwise comparisons, of which 417 are significant under the gold qrels when using NDCG to compute performance scores.

		Metric	Adjudication method							
			top-k	MTF	MM	MMNS	TS	TSNS	Hedge	NTCIR
NDCG (417 gold significantly different pairs)	Budget per topic: 10 (9%)	# rels.	441	488	489	474	483	469	504	**513**
		AA	231	223	249	253	252	252	202	**259**
		MA_{total}	376	322	**314**	333	315	337	388	381
		MA_G	184	193	167	164	165	165	215	**158**
		MA_L	192	**129**	146	170	151	172	173	223
		MD_{total}	14	**3**	**3**	4	**3**	5	13	14
		MD_G	2	1	**0**	**0**	**0**	**0**	**0**	0
		MD_L	12	**2**	**2**	4	**2**	5	13	14
		AD	0	0	0	0	0	0	0	0
		Bias	47%	**37%**	**37%**	41%	38%	41%	48%	48%
	Budget per topic: 30 (26%)	# rels.	1186	1327	**1359**	1289	1345	1267	1352	1337
		AA	347	337	337	**352**	338	350	281	345
		MA_{total}	203	203	180	199	**175**	207	283	243
		MA_G	70	80	80	**65**	79	67	136	72
		MA_L	133	122	101	134	**96**	140	147	171
		MD_{total}	0	0	0	0	0	0	0	0
		MD_G	0	0	0	0	0	0	0	0
		MD_L	0	0	0	0	0	0	0	0
		AD	0	0	0	0	0	0	0	0
		Bias	28%	27%	23%	27%	**22%**	29%	34%	33%

NTCIR than the adjudication methods but, overall, low budgets and shallow pools do not lead to reliable enough results.

When it comes to AA, differently from TREC8, they struggle to get close to the total number of significantly different pairs on the gold qrels. As in the TREC8 case, an increase in the number of relevant documents found does not necessarily lead to an increase in the AA counts.

On a positive side, AD is always 0, also for DL21.

When it comes to MA, we observe two different patterns. Differently from TREC8, MA_G is always quite high, motivating the general lack of Recall. In addition, MA_L does not substantially decrease as the budget increases, explaining the general lack of Precision.

Publication bias is exceedingly high, especially at low budgets, ranging between 25% and 50%. Overall, these high values shed a negative light on the reliability of the conclusions you would draw when using these methods under shallow pool conditions.

5.3.5 *How and Where the Methods Fail*

We study how and where, in terms of rank positions, the different methods fail in detecting significant differences.

We focus our analysis on the cases of MA, which have shown to be the main factor for the loss of Precision and Recall. Figure 5.1 shows the distribution of the score differences in system pairs which belong to MA with respect to their position in the gold ranking of systems for a budget of 100 assessments (6%). For each MA pair, we compute the difference between the score of the best and the worst system in the pair (under the adjudicated qrels, not the gold ones), recording it with a positive sign for the best system and a negative one for the worst system.[1] Figure 5.1 tries to convey information about the distribution of such differences as a series of boxplots would do, but in a more compact and readable way. The x-axis is the position of each system in the ranking of systems under the gold qrels, and we consider bins of three rank positions to make the figure more readable. For example, the first point in the figure represents the distribution of the mentioned differences for the first three systems in the gold ranking of systems. The solid line represents the median of the bin; the shaded area is limited by the first and third quartiles of

[1] For example, if we have the pair of system1 and system2 in mixed agreement, and system1 has the highest score, and their score difference is 0.15 (with the reduced pool). Then, for system1 we record 0.15 and for system2 we save −0.15. The mentioned figure plots the distribution of these differences for each system, according to their position in the ranking induced with the gold qrels.

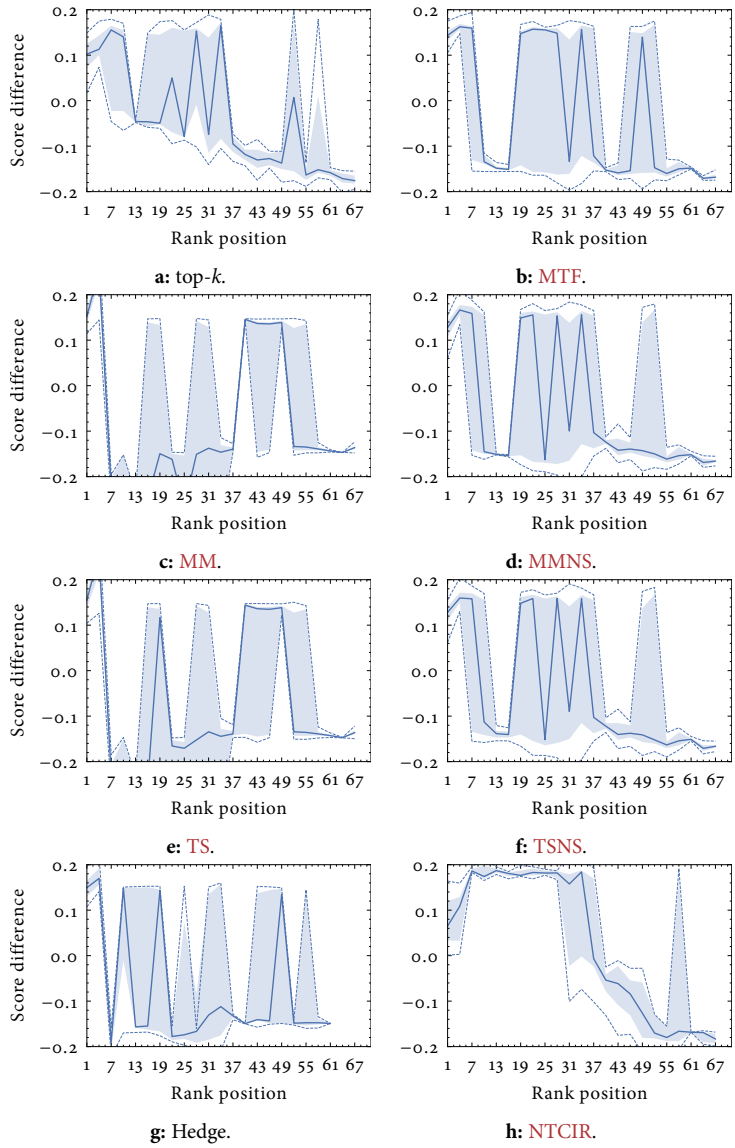


Figure 5.1: Distribution of MAP differences between systems in MA for a budget of 100 assessments (6%). The x-axis represents the systems sorted by their position in the official ranking. Each data point holds the distribution of 3 systems. The solid line represents the median of the bin. The shaded area is limited by the first and third quartiles of the distribution, i.e. it represents the inter-quartile range. Finally, the dashed lines are the maximum and the minimum. Breaks in the lines mean that there was not any mixed agreement for those systems. We used the 71 pooled systems of TREC8.

the distribution, i.e. it represents the inter-quartile range; finally, the dashed lines are the maximum and the minimum. A break in the lines means that no pair of systems in that range of rank positions is a MA.

We can see some clear trends among all the evaluated methods. As a general trend for most adjudication methods, the biggest differences occur between MA systems in the middle of the ranking (we see wider areas in the middle of the ranking), whereas we see more narrow distributions in the top-ranked and lowest-ranked methods. This suggests that the MA, and the consequent loss of Precision, happen in a region of moderate impact, since mid-rank systems may receive less interest in any case. Top-k and NTCIR method represent two notable exceptions. Indeed, top-k concentrates most of the score differences in the top ranks; therefore, top-k is not only the less performing method, but it also fails in the most impactful region of the ranking. This is even worse for NTCIR, where the biggest differences (of 0.2 points), are all clustered in the top positions of the ranking.

On the other hand, Hedge performs well compared to the other methods for the 6% budget, e.g. it achieves top Precision (0.955) and minimum bias (4%). However, Figure 5.1 shows that it spreads sizeable differences all over the ranking, affecting also quite impactful regions of it.

5.3.6 *Evaluation of Unseen Systems*

We investigate the *reusability* of the judgements produced by a low-cost method, i.e. their ability to fairly evaluate unseen systems. Usually, reusability is evaluated by following a *leave-one-group-out* approach. This consists in forming pools leaving one participating group each time and using those pools to evaluate the submissions of the group that was left out. We follow a different approach using the non-pooled systems of TREC8 (we do not perform these experiments on the DL21 collection since it does not include non-pooled runs). To this aim, we performed the same experiments as in the previous sections, but using the non-pooled systems of TREC8. In this way, we are evaluating systems that did not participate in the constructions of the pools. As commented in Section 5.3, this collection has been repeatedly acknowledged in the community as a high-quality one to evaluate unseen systems. Thus, we assume that the TREC8 gold judgements are reusable and, if a low-cost method provides the same significant differences as them, we conclude that it is reusable as well.

Tables 5.9 and 5.10 report the Kendall's τ, Precision and Recall values of every method, for a varying number of assessments per topic, using the non-pooled systems. On a positive side, Tables 5.9 and 5.10 show similar trends

Table 5.9: Values of Kendall's τ, Precision and Recall (see Section 5.2) of each adjudication method for a varying number of judgements per topic. Parentheses indicate the size of this budget with respect to the full pool. We used the 58 non-pooled systems from TREC8, and AP for computing performance scores.

Method	**Budget: 100** (6%)			**Budget: 300** (17%)		
	τ	**P**	**R**	τ	**P**	**R**
top-k	0.82	0.931	0.903	0.91	0.948	0.966
MTF	0.88	0.934	0.933	0.95	0.968	0.988
MM	**0.91**	0.967	**0.942**	**0.97**	0.976	**0.997**
MMNS	0.88	0.948	0.936	0.96	0.966	0.989
TS	**0.91**	0.969	0.940	**0.97**	0.973	0.996
TSNS	0.87	0.945	0.933	0.95	0.966	0.986
Hedge	**0.91**	**0.973**	0.929	0.96	**0.980**	0.982
NTCIR	0.89	0.898	0.931	0.95	0.962	0.984

Table 5.10: Values of Kendall's τ, Precision and Recall (see Section 5.2) of each adjudication method for a varying number of judgements per topic. Parentheses indicate the size of this budget with respect to the full pool. We used the 58 non-pooled systems from TREC8, and NDCG for computing performance scores.

Method	**Budget: 100** (6%)			**Budget: 300** (17%)		
	τ	**P**	**R**	τ	**P**	**R**
top-k	0.83	0.941	0.880	0.90	0.966	0.943
MTF	0.89	0.941	0.916	0.94	0.980	0.968
MM	0.92	0.955	**0.946**	**0.97**	**0.983**	**0.979**
MMNS	0.88	0.952	0.921	0.95	0.978	0.976
TS	0.92	0.956	0.944	**0.97**	0.979	0.977
TSNS	0.88	0.952	0.918	0.94	0.979	0.974
Hedge	**0.93**	**0.974**	**0.946**	0.96	0.977	0.977
NTCIR	0.86	0.938	0.911	0.94	0.974	0.977

Table 5.11: Values of relevants, agreements and bias of each adjudication method for a varying number of judgements per topic. Parentheses indicate the size of this budget with respect to the full pool. We used the 58 non-pooled systems from TREC8. The top-100 full pool includes 4728 relevant documents. There are 1653 pairwise comparisons, of which 509 are significant under the gold qrels when using AP to compute performance scores.

		Metric	Adjudication method							
			top-k	MTF	MM	MMNS	TS	TSNS	Hedge	NTCIR
AP (509 gold significantly different pairs)	Budget per topic: 100 (6%)	# rels.	1077	1685	2148	1553	2102	1514	**2170**	1481
		AA	460	475	**480**	477	479	475	473	474
		MA_{total}	83	68	**45**	58	**45**	62	49	89
		MA_G	49	34	**29**	32	30	34	36	35
		MA_L	34	34	16	26	15	28	**13**	54
		MD_{total}	0	0	0	0	0	0	0	0
		MD_G	0	0	0	0	0	0	0	0
		MD_L	0	0	0	0	0	0	0	0
		AD	0	0	0	0	0	0	0	0
		Bias	**7%**	**7%**	3%	5%	3%	5%	3%	4%
	Budget per topic: 300 (17%)	# rels.	2042	2923	**3628**	2913	3607	2868	3609	2723
		AA	492	503	**508**	504	507	502	500	501
		MA_{total}	44	23	**13**	23	16	25	19	28
		MA_G	17	6	**1**	5	2	7	9	8
		MA_L	27	17	12	18	14	18	**10**	20
		MD_{total}	0	0	0	0	0	0	0	0
		MD_G	0	0	0	0	0	0	0	0
		MD_L	0	0	0	0	0	0	0	0
		AD	0	0	0	0	0	0	0	0
		Bias	5%	3%	**2%**	3%	3%	3%	**2%**	4%

Table 5.12: Values of relevants, agreements and bias of each adjudication method for a varying number of judgements per topic. Parentheses indicate the size of this budget with respect to the full pool. We used the 58 non-pooled systems from TREC8. The top-100 full pool includes 4728 relevant documents. There are 1653 pairwise comparisons, of which 527 are significant under the gold qrels when using NDCG to compute performance scores.

		Metric	Adjudication method							
			top-k	MTF	MM	MMNS	TS	TSNS	Hedge	NTCIR
NDCG (527 gold significantly different pairs)	Budget per topic: 100 (6%)	# rels.	1077	1685	2148	1553	2102	1514	**2170**	1481
		AA	464	483	**499**	486	498	484	**499**	480
		MA_{total}	92	74	52	65	52	67	**41**	79
		MA_G	63	44	**28**	41	29	43	**28**	47
		MA_L	29	30	23	24	23	24	**13**	32
		MD_{total}	0	0	0	0	0	0	0	0
		MD_G	0	0	0	0	0	0	0	0
		MD_L	0	0	0	0	0	0	0	0
		AD	0	0	0	0	0	0	0	0
		Bias	6%	6%	4%	5%	4%	5%	**3%**	6%
	Budget per topic: 300 (17%)	# rels.	2042	2923	**3628**	2913	3607	2868	3609	2723
		AA	497	510	**516**	514	515	514	515	515
		MA_{total}	47	27	**19**	24	23	24	24	26
		MA_G	30	17	**11**	13	12	13	12	12
		MA_L	17	10	**9**	11	11	11	12	14
		MD_{total}	0	0	0	0	0	0	0	0
		MD_G	0	0	0	0	0	0	0	0
		MD_L	0	0	0	0	0	0	0	0
		AD	0	0	0	0	0	0	0	0
		Bias	3%	**2%**	**2%**	**2%**	**2%**	**2%**	**2%**	3%

as Tables 5.1 and 5.2, suggesting that there is not a specific bias against non-pooled systems. On a slightly negative side, we observe that performance with the non-pooled are generally slightly lower than those with the pooled ones, especially at the lowest budget, indicating a bit more loss and some more swaps due to not being pooled.

More in detail, TS, MM and Hedge always have the highest correlation scores and while MM achieves always the best Recall, independently of the budget and the metric. This means that if we were to gather the judgements of a new collection, MM would be the best option in terms of reusability of the collected assessments. As before, top-k and NTCIR method lag behind the other methods in all the cases and for every considered measure. This finding suggests that other alternative methods might be a better option to gather assessments when constructing new experimental collections.

Tables 5.11 and 5.12 report the agreements for the non-pooled systems. The results follow the same trends as with the pooled systems, further supporting the lack of strong biases against non-pooled systems. These scores confirm that alternative adjudication methods are more effective than top-k, which, contrary to what we observed before, now is clearly the worst method. As before, the more relevant documents found does not necessarily mean the more AA; therefore, not all the relevant documents are equally discriminative also for non-pooled systems.

No method fails in a mixed or active disagreement when evaluating the non-pooled systems. This further supports the fact that most drops in Kendall's τ are due to swaps between systems that are not significantly different under the gold qrels.

When it comes to the publication bias, we observe similar trends as in the case of the pooled systems, even with lower values, indicating that published conclusions would not change also in the case of non-pooled systems.

Finally, we can observe similar trends between the results obtained with AP and those obtained with NDCG, supporting the fact that the results presented here are generalizable in terms of the evaluation of unseen systems, and that they are not an artefact of the evaluation metric used.

Analogously to Figure 5.1, Figure 5.2, shows the distribution of the differences at different rank positions. On a quite positive side, we observe that, for all the considered methods, the MA are more clustered to the middle positions of the ranking than they were for the pooled systems (Figure 5.1). This suggests that, even if the performance (Precision and Recall) for non-pooled systems may be a bit lower than for the pooled ones, these drops actually affect less impactful areas of the ranking. In particular, in this case, Hedge

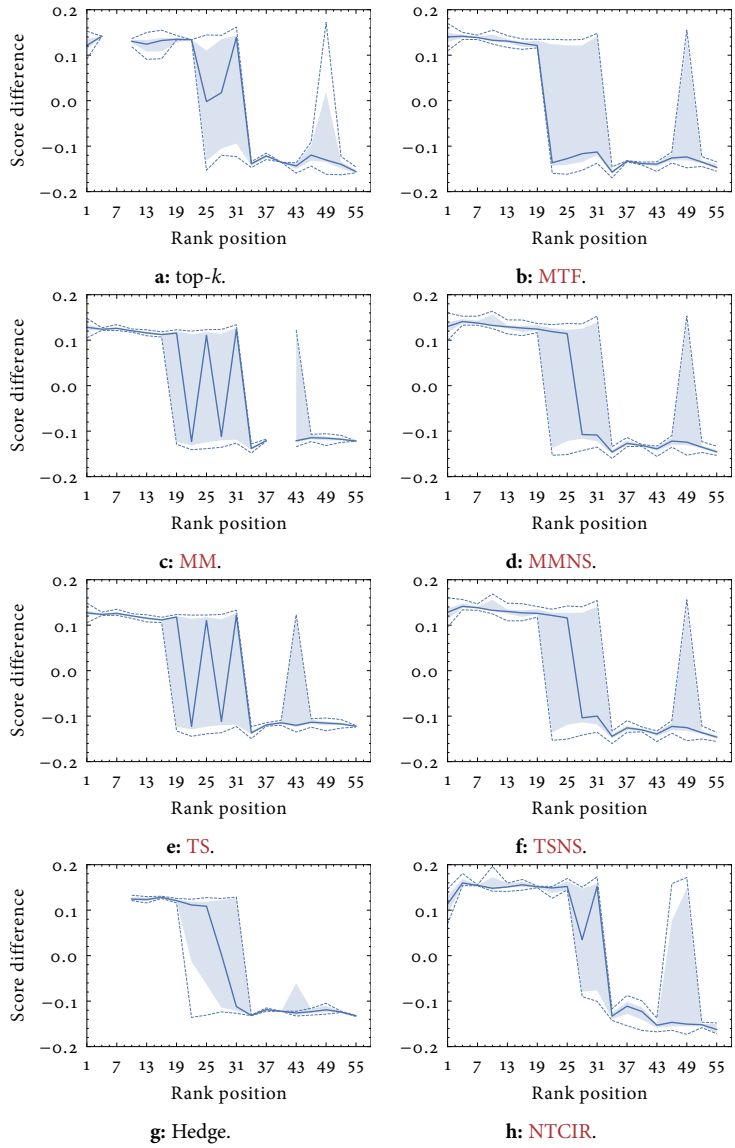


Figure 5.2: Distribution of MAP differences between systems in MA for a budget of 100 assessments (6%). The x-axis represents the systems sorted by their position in the official ranking. Each data point holds the distribution of 3 systems. The solid line represents the median of the bin. The shaded area is limited by the first and third quartiles of the distribution, i.e. it represents the inter-quartile range. Finally, the dashed lines are the maximum and the minimum. Breaks in the lines mean that there was not any mixed agreement for those systems. We used the 58 non-pooled systems of TREC8.

turns out to be the best algorithm in evaluating the top-ranked non-pooled systems, since it does not produce any MA in the top positions.

5.4 RELATED WORK

How to build high-quality test collections for retrieval evaluation is still an open research question (Craswell, Mitra, Yilmaz, Campos, Voorhees et al. 2021; Voorhees, Craswell et al. 2022; Voorhees, Soboroff et al. 2022). Research in adjudication methods looks for ways of prioritizing the pooled documents so that the assessors expend their effort in judging relevant documents. In this way, we may only need to judge some of the pooled documents while maintaining the quality of the judgements, thus making more efficient use of the resources. Here, we are briefly reviewing some works on adjudication methods. We refer the reader to Chapter 2 for a more extensive enumeration and description.

Losada, Parapar et al. (2017) proposed a series of sampling methods based on the multi-armed bandit problem. The multi-armed bandit problem (Sutton and Barto 2018, Chapter 2) has been a subject of research for decades in Reinforcement Learning (RL), statistics and other fields. These methods bring ideas from RL to the task of document adjudication for building test collections. They apply Bayesian principles to this problem, formalizing the uncertainty associated with reviewing a document from a pooled system. Other works have also explored the development of adjudication methods (Cormack, Palmer et al. 1998; Li and Kanoulas 2017; Moffat, Webber et al. 2007; Rahman et al. 2019). Adjudication methods have shown remarkable improvements in bringing relevant documents earlier in the pooling process, and indeed they were used to build the collection of the TREC Common Core Track of 2017 (Allan et al. 2017). However, the quality of the judgements produced with a limited budget is still an open question (Voorhees 2018).

Previous work on adjudicating methods used a series of metrics to evaluate the quality of these algorithms. The commonest is Kendall's τ (Kendall 1938, 1948) correlation, which researchers use to measure how well a new adjudication method can induce the gold ranking of systems, i.e. the one on the full top-k pool. Another top-weighted correlation, τ_{AP} (Yilmaz et al. 2008), is also common. This correlation penalizes swaps in higher positions more. In some works (Voorhees 2018; Voorhees, Craswell et al. 2022), they also measure the change in the ranking position of the system that suffers the highest drop as a measure of the reusability of an experimental collection. The problem with all these measures, as we already introduced earlier, is that

they ignore the significance between the scores of the systems. If we ignore this, it is meaningless to account for ranking swaps.

Statistical significance testing is of paramount importance in IR, and studying the properties of significance tests is an active area of research (Banks et al. 1999; Carterette 2012, 2017; Cormack and Lynam 2006, 2007b; Ferro and Sanderson 2019, 2022; Hull 1993; Parapar, Losada and Barreiro 2021; Parapar, Losada, Presedo-Quindimil et al. 2020; Sakai 2016a; Sanderson and Zobel 2005; Savoy 1997; Urbano, Lima et al. 2019; Urbano, Marrero et al. 2013; Webber et al. 2008). However, this is out-of-scope for the present chapter which, instead, focused on considering the output of a statistical significance test as a way to assess the quality of an adjudication method.

5.5 CONCLUSIONS

We argued for the need of a more powerful way of evaluating adjudication methods. In particular, while the current approach just focuses on how close two alternative methods rank systems, quantified by Kendall's τ, we think that we should focus our attention also on how different methods behave with respect to the significantly different pairs of systems detected. Indeed, while the current approach looks for stability in answering the question "is system A better than B?", our proposed method looks for stability in answering the question "is system A significantly better than B?", which is the ultimate question researchers are interested in to ensure generalizability of results.

To this end, we considered two measures—namely Precision and Recall—which consider significantly different pairs in isolation, as well as measures—the agreement/disagreement counts—which relate them to swaps in the ranking of systems. We also considered the problem of publication bias, i.e. the chance of publishing results/conclusions that would not hold or be the opposite when using the full pool instead of a reduced one.

To validate and showcase our proposed approach, we conducted thorough experimentation on TREC8, a collection renowned for its high-quality deep pool, and TREC Deep Learning 2021, a collection adopting a very shallow pool. In this way, we have shown that our methodology allows us to obtain insights that are not possible simply by using Kendall's τ.

For example, we found that no AD and (almost) no MD happen. This means that observed drops in Kendall's τ are mostly due to swaps between not significantly different systems. Therefore, those drops concern not very interesting system pairs, and it might not be worth striving for (or to judge a method just by) 1.00 Kendall's τ.

We also found that the number of relevant documents detected by a method does not necessarily increase the number of significantly different pairs detected, suggesting that not all the relevant documents in a pool are equally discriminative. This opens up interesting future investigations on which (relevant) documents would be optimal for a pool, while the current focus has been more on determining how many and which topics to sample.

We have shown that drops in Precision and Recall are mainly caused by MA, which distribute unevenly at different rank positions and, therefore, they have a quite different impact: those happening at mid-to-bottom rank positions are less serious than those happening at the top positions of the ranking.

Finally, we also found that no adjudication method induces strong biases against non-pooled systems, thus further supporting the use of these methods to construct new test collections for IR evaluation. Previous work evaluated the reusability of bandit-based methods using Kendall's τ and other swap-based measures, and concluded that the collections built with them were less reusable than desirable (Voorhees 2018). With the new evaluation approach we have presented in this paper, we shed more light on this issue and show that bandit-based methods are indeed reusable when focusing on significance between systems.

Overall, our approach shows that existing methods for human assessment adjudication in IR evaluation could preserve most of the true statistical differences between the pairwise comparisons of systems. Besides this, as discussed in detail, our approach allowed us to pinpoint which adjudication method works better in specific conditions, why, and how it differs from other methods. This will thus be a helpful tool and guidance for researchers when they have to decide which method to choose in their settings.

Part IV

COLLECTIONS FOR NOVEL TASKS

6

BUILDING COLLECTIONS FOR NOVEL TASKS

So far, we have focused on solving two of the main challenges of building new IR test collections: the need to have enough diverse participant systems and the effort needed to obtain enough reusable relevance judgements. We also proposed a novel way of evaluating adjudication methods, focusing on their ability to preserve the real significant differences between systems. In this closing chapter, we use our contributions in a real-world scenario and apply them in a real task, where we build a new annotated dataset in a reproducible and shareable way. In particular, we propose a standardized methodology for building new annotated datasets for novel tasks. We also create—and release—a dataset about the *patrimonialization* of cultural heritage reflected in social networks (Otero, Martin-Rodilla et al. 2021). As we will see later in this chapter, this collection constitutes a valuable resource for researchers to study social processes about the patrimonialization of cultural heritage entities and their relation to racial tensions. The work presented in this chapter has already been published (Otero 2019; Otero, Martin-Rodilla et al. 2021; Otero, Parapar and Barreiro 2020).

The content of this chapter is organized as follows: Section 6.1 presents the new task and the main problems we have identified in this new domain. Then, in Section 6.2, we present and explain our approach for building this new annotated collection. Finally, in Section 6.3, we conclude the chapter by presenting our main conclusions.

6.1 INTRODUCTION

Patrimonialization is the process by which a material or immaterial element becomes a constitutive part of a community's identity (Rivero et al. 2020). The community imbues the said element with meaning and significance,

thus becoming a constitutive part of its identity. This patrimonialization can include any material or immaterial element. In this case, we are particularly interested in the patrimonialization of cultural heritage, which includes the construction, identification, rejection, or destruction of cultural heritage. This cultural heritage includes, for example, tangible elements such as statues of public and important historic figures, plaques, monuments, or books. These patrimonialization processes, which are inherently social, are reflected in different areas of our social reality and, in particular, in social media.

Recently, the study of these social patrimonialization processes, particularly those related to cultural heritage, has focused on analysing the information generated in social networks (Kirschenbaum et al. 2010). In this kind of media, people create content in various and different formats, such as textual comments, photos and videos. Research in this area is interested in studying all this content from an archival perspective, where the objective is to construct snapshots of social media content about the processes that are being studied (Pybus 2013; Ries and Palko 2019; Schefbech et al. 2012).

These collections, that result from archiving content directly from social networks, are called *born-digital* archives. These born-digital archives are essential for cultural heritage researchers since they constitute a primary source in the empirical study of patrimonialization processes. These archives act as a snapshot of the patrimonialization processes reflected in social networks at a given point in time, and are very useful for researchers who want to study them. Thus, it would be useful to have a well-established methodology for building new born-digital archives in a standardised way.

Computational Archival Science (CAS) is a transdisciplinary field that integrates computational and archival theories, methods, and resources, both to support the creation and preservation of reliable and authentic records/archives and to address large-scale records/archives processing, analysis, storage, and access, with the aim of improving efficiency, productivity and precision, in support of recordkeeping, appraisal, arrangement and description, preservation and access decisions, and engaging and undertaking research with archival material (Marciano et al. 2018).

A large volume of work and initiatives have been developed in recent years around the concept of CAS from a transdisciplinary perspective, combining techniques, tools, and methodologies that expand disciplines and improve the treatment of large-scale records. Good examples are CAS Workshops (*IEEE Big Data CAS Workshop* 2022; *IEEE Big Data CAS Workshop* 2023). Also, CAS as a field has required analysis and methodological and applied contributions by the disciplines that compose it, as can be seen in examples of projects with different sources: textual (Stančić 2018), cartographic-based (Lee et al. 2017;

Stančić 2018) or even from multimedia archives (Hamouda et al. 2019), as well as in recent compendia and theoretical studies on the discipline (Lee et al. 2017; Payne 2018).

Additionally, this kind of work supports in some way the *records continuum theory* (Upward 2005). Upward's theory establishes that any record has to be managed at an archival level without strict phases or protocols since the creation of each record. This vision of archiving as a continuum flow is beneficial when working with social network information. Retrieval from social networks to create reference collections from them as born-digital archives, that constitute a snapshot of the social processes we are studying, is an example of the application of this theory.

However, all these works present some problems in the retrieval techniques applied for building reference collections. We have detected the following:

- Records (from social media) are retrieved manually or semi-manually (Schefbech et al. 2012).
- Records are retrieved using computer techniques not adapted to the informational domain or the final purpose of the information (Mordell 2019; Winters and Prescott 2019).
- Most works do not deal with privacy and security issues, such as anonymisation processes in the information retrieval workflow (McNealy 2011; Obodoruku 2016), with also implications for cultural heritage archives (Kirschenbaum et al. 2010).
- Most of these works are one-case studies, in which the information retrieval techniques and software tools applied (Blanke et al. 2013) and, in some cases, the resultant collections are not available for further use and applications (Schefbech et al. 2012).

In this chapter, we propose a methodology for building new digital archives by importing ideas from common practices in IR and employing the previous contributions we have made in this thesis: automatically generating runs from which to pool the documents and using an adjudication method to reduce the budget of judgements. This methodology allows the creation of born-digital archives in cultural heritage from social media in a customisable, reproducible and shareable way, overcoming some of the problems we have just enumerated. In addition, the methodology is evaluated through a real case study: the creation of a reference collection on the recent attacks on patrimonial entities motivated by anti-racist protests. This reference collection and the results obtained from its preliminary study are freely available for use and already

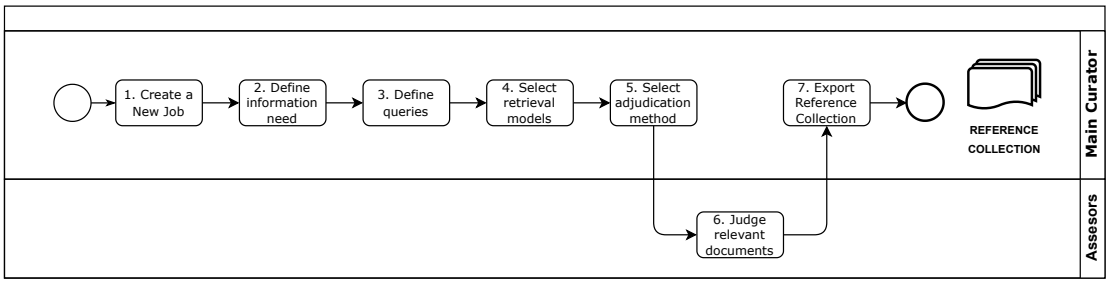


Figure 6.1: Workflow of our proposed methodology for building new reference collections from social media.

constitute a born-digital archive that will allow future researchers to have real information about citizen's motivations and comments, the different social sensitivities concerning these attacks on heritage entities, as well as references to places or heritage entities under attack throughout the protests.

6.2 CULTURAL HERITAGE REFERENCE COLLECTIONS FROM SOCIAL MEDIA

As we have mentioned, in this chapter, we present a methodology we developed to build annotated collections from social media, taking advantage of the methods we have proposed in previous chapters. In this section, we first explain the details of this methodology, and then we go through a real case of how we built a new collection about attacks and riots on cultural heritage entities and its reflection in social media.

6.2.1 *Methodology*

Figure 6.1 depicts a summary of our methodology. This figure shows that the functionality is divided into roles: Main Curator and Assessor. We have centred our methodology around the concept of *Job*. Each job represents the process of creating a different reference collection. We now explain the tasks to build a new collection corresponding to each role.

6.2.1.1 *Main Curator*

The main curator is the person responsible for the whole experiment. This curator has to establish the main topic and scope of an experiment (step number 2), representing the collection's information need. This would mean that we are interested in retrieving, from social media, information that is related, in one way or another, to this information need. Once the topic is

set, the first step to build the collection is to derive some queries (step 3) in the form of short sentences consisting of a few keywords that will then be used to search for this information on social media. This task is assigned to the main curator. The next step in the configuration of an experiment is to select the retrieval models (step 4). These retrieval models will be used to generate the runs that will serve as the input to the adjudication method to pool the documents. This adjudication method is chosen in the next step (step 5). These retrieval methods and the query variations introduced by the curator are then going to be used to generate the synthetic runs, exactly as we did in Chapter 3. Once we have downloaded the posts from social media, the runs created, and the adjudication method selected, it is now the assessors' turn to start judging each document's relevance.

6.2.1.2 *Assessors*

The assessors may be one person or a group that should know the domain and topic of the collection so they can judge the relevance of the downloaded documents that will be part of the final reference collection. In our case, assessors should be experts who work on cultural heritage attacks, for example. The primary and only work of the assessors is to judge the relevance of each downloaded document (step 6). This assessment should be according to the main topic of the experiment.

Once the assessors have finished making the relevance judgements of the selected posts, the curator is responsible for exporting the final reference collection (step 7). This reference collection will include the posts' textual content and the relevance judgements made by the assessors.

We now proceed to show how we applied this methodology to build a new reference collection on attacks against patrimonial entities.

6.2.2 *The Case of 2020's Tensions over Race and Heritage*

During the year 2020, there were many protests and attacks on heritage elements such as statues and commemorative plaques worldwide. All these revolts began due to the death of George Floyd in the United States. This event gave rise to a series of protests in which various cultural elements commemorating important figures in history were attacked under the motivation that the figures represented were racists and genocides.

This connection between heritage and racism has been widely studied previously, as well as its connection with the events that occurred in 2020, from different disciplinary perspectives, such as philosophy (Arday 2021; Davidson-

Harden 2021; Samayeen et al. 2022), history and anthropology (Gibson and Reich 2017; Ince et al. 2017), sociological (Schwartz 2020) or archaeological and heritage studies (Meskell 2002). It is also possible to find analysis of the phenomenon within the so-called heritage in social conflict or social fractures in heritage, as detailed by anthropological studies in the area (Cortés-Vázquez et al. 2017; Sanchez 2013).

Therefore, the patrimonial attacks that occurred in 2020 reflect social and patrimonial processes of a very diverse nature that, in our opinion, require the existence of reference collections that allow the subsequent research on these issues from different methodologies, fields, and contributions.

2020's protests and heritage attacks events also encouraged the appearance of people on social networks who comment and discuss their opinions on these issues. Thus, we want to build a new reference collection with these publications to serve as a social thermometer to study this *patrimonialization* process.

As we have previously shown, this topic is probably one of the current phenomena related to patrimonial entities that receive the most interest in public opinion and media and researchers in different disciplines (Arday 2021; Davidson-Harden 2021; Ince et al. 2017; Samayeen et al. 2022; Schwartz 2020). It has also been shown the critical role of social networks in influencing public opinion (and electoral processes while in progress in the USA), as a call for action instrument (both to attack cultural heritage entities and to manifest against attacks) and in the organisation of related platforms and collectives. The preservation of consistent information on social networks on 2020's racial tensions and heritage attacks and the possibility of his later study from different points of view seems to us a great motivation to validate and illustrate our methodology to a real case study.

In this case, we have used Reddit as the social network from which we are building the collection. Thus, a document is a Reddit post. Reddit is a social network very suitable for our use case: to build a reference collection on the recent anti-racist protests and patrimonial attacks because of the threaded nature of its posts. However, the document entity can also represent other types of sources.

6.2.2.1 *Queries definition*

To find posts relevant to this topic, we wanted to curate queries that included references to a patrimonial entity and the citizen's opinions about it. For this reason, we created two lists of terms. The first list included terms related to patrimonial entities, such as *monument*, *memorial*, *plaque*, *bust*, *statue*,

Table 6.1: Lists of terms for both groups. Terms in the same row does not mean we have use them together.

Conflicts related terms	Entity related terms
anti-black	monument
blacklives	removed
blacklivesmatter	removal
police brutality	statue
polive violence	tribute
abuse of authority	memorial
racism	plaque
racial bias	bust
anti-racism	take down
george floyd	beheaded
slavery	desecrated
slave	vandalized
-	vandalism
-	vandals
-	protests
-	protesters

tribute, among others. The second list comprised useful terms to find people's opinions on the anti-racist protests and conflicts, such as *racism*, *slavery*, *slave*, *black lives*, *violence*, among others. The queries we used contained terms from both lists, so the posts that match the queries are related to our information needs. In Table 6.1, we show each group's entire set of terms.

We did initial research to find the terms and queries that retrieved the more precise and accurate comments on this topic. This initial research consisted of creating jobs, one for each query, and seeing if the retrieved posts were relevant. Finally, the set of queries that retrieved the more relevant documents were the following: "statue racism", "monument racism", "police violence statue", "black lives monument", and "slave plaque". Thus, these are the five queries we have used to build this collection.

It is important to note that our methodology allows us to replicate the entire collection building process and expand the set of terms that constitutes queries in a new experiment. Thus, it could be possible to use existing thesauri or similar linguistic resources to expand the queries done in the future.

Post ID	Date	Body	URL
6549117	01-11-2020	I'm an alum, and I would be upset if the statue were removed. I doubt that the push to remove the statue represents the views of a majority of the over 45,000 students enrolled at UW-Madison. Back in the 80s I was a student activist who participated in protests on Bascom Hill (where the statue of Lincoln is) to pressure the university to divest in companies doing business with the government of South Africa (the point of this effort was to help end apartheid). I would encourage students who want to take action against racism to do something that would benefit irl people. My experience when I was a student was that even politically active people like me viewed the student government as irrelevant. They have no power and no real function. So people don't pay attention to them. Exception: Back in the 70s there was a joke party called [Pail and Shovel](https://madison.com/wsj/news/local/pink-flamingos-statue-of-liberty-boombox-parade-the-legacy-of-madison-prankster-leon-varjian/collection_687ab699-2e01-592f-82cf-4ecae2dfd6af.html) that did whimsical pranks that people still talk about today.	www.reddit.com/r/centrist/comments/jm0w1o/wisconsin_student_gov_votes_to_remove_lincoln/gat833v/

Relevant Non relevant

Figure 6.2: Example of a judged post.

6.2.2.2 *Relevance judgment*

These selected queries resulted in 522 different posts to judge. A group of three assessors made the judgments, where each one judged approximately the third part of the posts. Figure 6.2 shows an example of one of the posts that were judged by the assessors and how they saw it. As we said before, the assessors should be familiar with the topic of the collection to provide standardised and high-quality judgments. In this case, all assessors have worked before in cultural heritage contexts and entities in conflict (Gonzalez-Perez et al. 2012; Martín-Rodilla et al. 2019). Also, the criteria to decide the relevance of the posts were made a priori before the assessment process. This was to mitigate any personal bias that could be introduced in the collection and provide the assessors with a simple guide on deciding the relevance of the documents. Annotators were presented with one post at a time. They only annotated as relevant posts that contained references to patrimonial entities, such as statues or plaques, which also contained the writer's opinion about the attacks and protests. This process took one week.

6.2.2.3 *2020's Tensions over Race and Heritage Collection*

The final reference collection consists of 522 Reddit posts judged as positive. Additionally, the collection includes the content of 296 threads extracted from

the posts judged as relevant. In total, the collection has 260 578 posts (that gives us an average of 880.3 posts incorporated in the reference collection per thread). The posts were written between 25th May, 2020 (George Floyd's death date) and 31st October, 2020. Note that it is necessary to judge only 522 posts during the judgment phase, avoiding the complete annotation of the reference collection, for obtaining almost six months of Reddit activity and information about Lloyd's movement and anti-racism related protests. The resultant reference collection is available at this link[1], constituting the first reference collection, as far as we know, that preserves and gives access to the study of the recent attacks on patrimonial entities motivated by anti-racist protests.

It is important to note that this high number of different threads ensures that the collection includes posts that deal with the issue of protests and their relationship with attacks on patrimonial entities, not only from explicit threads with that topic but also from many threads with different central topics. This is important to ensure coverage of the topic within the social network, not only recovering those conversational threads focused on the topic (which may reflect points of view that are excessively polarised or directed by associations or people directly implicated in the conflicts). In this way, we also reach threads with topics different from the main topic where the conversation has turned, dealing with protests at some point. These conversations may be less polarised and include citizens with more diverse profiles and responsibilities.

2020's Tensions over Race and Heritage Collection, as a real application, has allowed us to validate the methodology and the pooling strategies for creating reference collections in heritage studies from social networks.

In addition to this reference collection, we release the history published by a sample of users participating in those threads. From the 296 threads comprising the main collection, we found more than 90 000 users participating in them. From these users, we sampled 1400 of them and retrieved the whole history of posts published by each one. This resulted in a collection with 6 455 258 (including the content of the threads) different posts. We hope that this will help to research the sensitivities of those users concerning patrimonial attacks and a deeper investigation into those users' profiles. In Table 6.2, we present a summary of both collections.

[1] https://www.dc.fi.udc.es/~david/heritage

Table 6.2: Summary of released collections.

Collection	# posts
Only threads	260 578
Full collection	6 455 258

6.2.2.4 *Privacy issues*

To avoid revealing private or personal information about the users, we have anonymised them by substituting their original Reddit username with a randomly generated identifier. We know that just hiding its original username in Reddit may not be sufficient to cloak their identification. However, we must note one important aspect here. All the data we have crawled to create this collection is public data made available by Internet users. We have not gathered any private or personal information of the users. We must point out that if the users' personal information is retrieved from this collection, it must be treated following practices that ensure their anonymity.

6.3 CONCLUSIONS

We have analysed the intersection between IR and CAS in the specific case of reference collections (as born-digital archives) from social networks. Social networks have become a real-time reflection of social processes. Researchers use social network information for studying cultural heritage processes. From our particular interest in creating born-digital archives from social networks as a dynamic, continuum, and transdisciplinary archive, we have developed a methodology that enables innovative pooling-based judging to create reference collections from social media. Besides, the platform is evaluated in a real case study on cultural heritage, with the creation of a born-digital reference collection from Reddit that retrieves, monitors, documents, preserves, and allows the evolutionary study of the phenomenon of attacks on heritage entities in the anti-racism protest of 2020 around the world. The resulting collection consists of more than 260 000 relevant posts with all kinds of opinions, visions, and attitudes towards the racial conflict and the heritage involved in different areas of the world and by very different people's profiles. This constitutes a born-digital archive about the attacks suffered by patrimonial entities and the activity in social networks generated around these attacks and the anti-racist riots in 2020. Moreover, this chapter goes a step further and demonstrates,

through a real and successful application, the usefulness of all the proposals we have presented in this doctoral thesis in previous chapters.

Part V

EPILOGUE

7

CONCLUSIONS AND NEW RESEARCH OPPORTUNITIES

In this last chapter, we summarize the main conclusions of this doctoral thesis and suggest some ideas for future research directions.

7.1 CONCLUSIONS

In this thesis, our aim was to explore new approaches to facilitate the construction of new collections for retrieval evaluation. We proposed a way of generating synthetic runs when no participant teams are available. We also tried to improve the adjudication of the pooled documents with the aim of making a more efficient use of the judgements budget. Then, we provided a new perspective in the evaluation of adjudication methods, to finally put into practice our proposals under a real-world scenario.

We now provide detailed explanations of all our contributions:

- In Chapter 3, we studied how to create synthetic runs from which to pool the documents. One of the assumptions of why pooling usually works, even though we are leaving a great deal of documents unjudged when building pooled test collections, is that judged documents are pooled from a set of a relatively high number of submissions which, at the same time, are diverse in terms of the documents that they are ranking in the top positions. Having enough and diverse participant systems is not always feasible and not everyone has enough resources to organize a TREC-like workshop. With our basic and simple strategy for automatically creating runs from a corpus of documents—using automatically generated query variations and well-known retrieval models—, we simulated the making of judgements that were comparable in quality to those created from TREC workshops. In particular, with a strong adjudication method like MTF, we

obtained ranking correlations above 0.9 with respect to the official ranking of systems.

- In Chapter 4, we examined the adaptation of several relevance feedback models to the task of document adjudication. We found that relevance feedback is indeed useful for improving the performance of existing adjudication methods. In particular, our experiments demonstrated that we were able to improve the ratio—with respect to existing strong baselines—at which relevant documents appear when making new judgements. Also, these algorithms were, at least, as competitive as these baselines in terms of reliability, fairness, and reusability. This concludes that our proposed methods are a cost-effective alternative for building new retrieval evaluation collections. Interestingly, we also found that the best feedback model, both for reranking the whole pool and for reranking per each submission, was Divergence Minimization Model (DMM) and that long queries (*title* + *description*) with low interpolation values worked better for this particular task of document adjudication.

- In Chapter 5, we proposed a methodology for evaluating adjudication methods. Our proposal was based on looking for stability in preserving the statistically significant differences between systems. Then, we applied this methodology to evaluate state-of-the-art adjudication methods. With this work, we found several interesting results. First is that losses on the traditional Kendall's τ are primarily due to swaps in pairs that are not truly significantly different. Thus, these drops concern a less interesting part of the ranking. We also found that an increase in the number of relevant documents does not always yield better detections of significantly different pairs. This suggests that not all the relevant documents in a pool are equally discriminative. Our methodology also allowed us to demonstrate that, when focusing on the significantly different pairs, bandit-based methods are able to build reusable judgements, something that was not clear from past experiments (Voorhees 2018). Overall, our method allowed us to show which adjudication method works better in specific conditions, why, and how it differs from other methods.

- Last, in Chapter 6, we put into practice some of the contributions of previous chapters. We employed our methodology for creating synthetic participant runs and strong adjudication methods for constructing a new collection for a specific task and domain. In particular, we built a new dataset about attacks and riots on patrimonial entities reflected in social networks. Our aim was to provide future researchers with a snapshot of the

content in social networks about the patrimonialization of these cultural heritage entities, and the opinions and visions of social networks users on these processes. As a result of our work, we released a collection that includes more than 260 000 relevant posts with all kinds of opinions, visions, and attitudes towards the racial conflict and the heritage involved, in different areas of the world and by very different people's profiles.

7.2 LOOKING AHEAD

Several research opportunities stem from the contributions presented in this doctoral thesis. We now take a look ahead and briefly propose some of them:

- Regarding the simulation of synthetic submissions for pooling, something we proposed and evaluated in Chapter 3, we envision several lines that could be investigated. The most natural path is to include more and better retrieval models to generate new runs. In recent years, advancements in the field of Natural Language Processing (NLP) about contextualized embeddings and large language models led to the development of new IR systems, such as those based on transformers, that behave very differently from traditional ones (Craswell, Mitra, Yilmaz, Campos and Lin 2021; Craswell, Mitra, Yilmaz, Campos, Voorhees et al. 2021; Lin et al. 2021; Voorhees, Soboroff et al. 2022). We are interested in evaluating the addition of this kind of models to the synthetic runs, since our work only included traditional models. Regarding the automatic generation of query variations, we want to explore more robust approaches to improve the quality of the generated runs, and thus the quality of the pools. To this aim, we will also explore NLP models based on transformers used in related areas for query generation (Alaofi et al. 2023; Penha et al. 2022).

- Regarding our work on statistical feedback models for document adjudication, it paves the way for further research on using relevance feedback for document adjudication. The good results achieved by employing popular relevance feedback models indicate that there may be room for improvement using other methods that explore different techniques. Thus, we plan to research other approaches here. In particular, we want to test the behaviour of relevance feedback with matrix factorization and linear methods (Valcarce et al. 2018a,b). We also plan to investigate a method based on Reinforcement Learning (Montazeralghaem et al. 2020) as well as variations of relevance models (Parapar and Barreiro 2011; Roy et al. 2019). Finally, in this scenario, there is explicit negative feedback (documents judged as non-relevant). Thus, we believe it would be interesting to study

the use of the negative relevance feedback in the estimation of the models used for reranking (Wang et al. 2008).

- In Chapter 5, we proposed a more reliable methodology for evaluating low-cost adjudication methods for building new IR benchmarks. Then, we applied this methodology to perform an extensive evaluation of several adjudication methods. We believe that from the conclusions of this work there are a lot of directions that can be explored in the future. In this thesis, we focused our analysis on two of the most common benchmarks used in IR evaluation and two of the most employed metrics, namely, AP and NDCG. We think extending it to more benchmarks and metrics would be interesting. We found that the number of relevant documents detected by a method does not necessarily increase the number of significantly different pairs detected. This suggests that not all the relevant documents in a pool are equally discriminative. This opens up at least two interesting investigations that might be worth exploring. The first one is studying which relevant documents would be optimal for a pool, while so far the focus in the field has been on how many and which topics to use in the evaluation (Buckley and Voorhees 2000; Sakai 2016b; Sanderson and Zobel 2005; Voorhees 2009; Voorhees and Buckley 2002). The second one is studying which factors are the ones that most influence the number of active agreements obtained with reduced pools, for example, the position of the most discriminative documents in the runs.

- Regarding our work on creating collections for novels tasks and, in particular, the dataset we released about patrimonialization of cultural heritage processes, we propose here several ideas that could be explored. We plan to use natural language processing and text mining techniques on the collection to identify and extract heritage sites or entities. This analysis will allow us to elaborate maps on where the attacks have been and which heritage entities or historical events are involved. We also plan to enrich the collection with information about the profiles of users of the social network whose posts are in our reference collection. To do this, we will add to the published reference collection all the posts in all the subreddits of each user who is involved in the current collection. This will allow us to have a personal history of activity on the social network of each of the users who expressed their opinion about racial tensions and attacks on heritage: what other topics interest them and what do they post, their interactions' registry, expertise in the social network. This personal information is of complementary value for studies focused on people: What kinds of people

have posted about tensions over race and heritage? What else can we know about them?

APPENDICES

A

PUBLICATIONS

In this appendix, we list all the publications made during the doctoral period. For the conferences, we provide their rank according to CORE 2023[1] and the acceptance rate of the full or short papers track. For each journal, we detail its Journal Citation Reports Impact Factor[2] and its quartile.

A.1 CONFERENCE ARTICLES

David Otero (2019).
Exploiting Pooling Methods for Building Datasets for Novel Tasks. In: *Proceedings of the 9th PhD Symposium on Future Directions in Information Access*. FDIA '19. Milan, Italy: CEUR-WS.org, pp. 96–102. URL: http://ceur-ws.org/Vol-2537/paper-16.pdf.

David Otero, Javier Parapar and Álvaro Barreiro (2020).
Beaver: Efficiently Building Test Collections for Novel Tasks. In: *Proceedings of the Joint Conference of the Information Retrieval Communities in Europe*. CIRCLE '20. Samatan, Gers, France: CEUR-WS.org. URL: http://ceur-ws.org/Vol-2621/CIRCLE20_23.pdf.

David Otero, Javier Parapar and Álvaro Barreiro (2021).
The Wisdom of the Rankers: A Cost-Effective Method for Building Pooled Test Collections without Participant Systems. In: *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. SAC '21. New York, NY, USA: Association for Computing Machinery, pp. 672–680. DOI: 10.1145/3412841.3441947. CORE 2023: B. Acceptance rate (full papers): 25%.

David Otero, Javier Parapar and Nicola Ferro (2023).

[1] The CORE 2023 Conference ranking is available at: http://portal.core.edu.au/conf-ranks.
[2] The JCR Impact Factor can be consulted at: https://jcr.clarivate.com/jcr/home.

How Discriminative Are Your Qrels? How To Study the Statistical Significance of Document Adjudication Methods. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. CIKM '23. New York, NY, USA: Association for Computing Machinery. DOI: 10.1145/3583780.3614916. CORE 2023: A. Acceptance rate (full papers): 24%.

David Otero, Daniel Valcarce, Javier Parapar and Álvaro Barreiro (2019). Building High-Quality Datasets for Information Retrieval Evaluation at a Reduced Cost. In: *Proceedings of The 2nd XoveTIC Conference (XoveTIC 2019)*. MDPI. DOI: 10.3390/proceedings2019021033.

A.2 JOURNAL ARTICLES

David Otero, Patricia Martin-Rodilla and Javier Parapar (2021). Building Cultural Heritage Reference Collections from Social Media through Pooling Strategies: The Case of 2020's Tensions Over Race and Heritage. In: *Journal on Computing and Cultural Heritage* 15.1, pp. 1–13. DOI: 10.1145/3477604. IF 2022: 2.4 (Q3).

David Otero, Javier Parapar and Álvaro Barreiro (2023). Relevance feedback for building pooled test collections. In: *Journal of Information Science*. DOI: 10.1177/01655515231171085. IF 2022: 2.4 (Q3).

A.3 BOOK'S CHAPTERS

Patricia Martín-Rodilla and David Otero (2023). Estrategias de recuperación de información mediante "pooling" para la construcción de colecciones de referencia desde redes sociales: caso de estudio durante las tensiones raciales de 2020. In: *Scire vias. Humanidades digitales y conocimiento*. Ed. by Universidade da Coruña, pp. 347–365.

B

EXTENDED SUMMARY IN SPANISH

In accordance with the current Regulations of the PhD Studies of the Universidade da Coruña, we present in this appendix an extended summary of this doctoral thesis in Spanish.

B.1 INTRODUCCIÓN

Los motores de búsqueda son herramientas esenciales para explorar vastas colecciones de contenidos. Localizar información relevante en estas colecciones es casi imposible sin ellos. Los usuarios confían en los motores de búsqueda para examinar colecciones de documentos, como correos electrónicos, productos de comercio electrónico, vídeos en *streaming*, páginas web y noticias. Detrás de cada motor de búsqueda hay un sistema de *ranking* responsable de identificar los elementos más relevantes en función de las necesidades de información del usuario. La recuperación de información es un campo de la informática dedicado a desarrollar sistemas que ayuden a los usuarios a encontrar información adaptada a sus necesidades. La importancia de estos sistemas ha crecido enormemente, sobre todo a medida que se convierten en parte integral de ámbitos como el comercio electrónico, los sistemas de recomendación y las redes sociales. La calidad del *ranking* influye profundamente en la experiencia del usuario, por lo que la eficacia del sistema es vital para su satisfacción. En consecuencia, el rendimiento de estos sistemas es primordial tanto para los usuarios como para los proveedores, lo que lleva a realizar grandes esfuerzos en su evaluación y mejora.

Tradicionalmente, la investigación y el desarrollo en este campo se han basado en datos etiquetados por humanos en forma de anotaciones de expertos: éstos evalúan la relevancia de los documentos con respecto a necesidades de información específicas, creando conjuntos de datos anotados que sirven para entrenar y evaluar sistemas de *ranking*. Construir nuevas colecciones

específicas para el número siempre creciente de tareas diferentes relacionadas con la recuperación de información resulta muy caro debido a este esfuerzo humano necesario. Este coste dificulta la creación de nuevas colecciones y, por tanto, dificulta también la investigación y el desarrollo de nuevas ideas en este campo.

Esta tesis introduce métodos novedosos para construir, de forma eficiente y fiable, colecciones anotadas en el campo de la recuperación de información.

B.2 MOTIVACIÓN

La recuperación de información tiene una larga y rica tradición en la experimentación con sistemas de *ranking*. Esta tradición se remonta a las décadas de 1950 y 1960, marcadas por el trabajo pionero de Cyril Cleverdon. Antes de sus experimentos, los debates sobre los distintos enfoques para la búsqueda de información eran en gran medida anecdóticos y filosóficos. Cleverdon fue el primero en realizar un experimento empírico, formalmente científico, para comparar distintos esquemas de indexación. Los experimentos de Cleverdon, controvertidos en su momento, fueron los primeros en utilizar *colecciones de prueba*: un conjunto común de documentos, un conjunto común de necesidades de información y un conjunto común de juicios de relevancia. Este enfoque permitió realizar experimentos más controlados y mejores comparaciones, reduciendo la variabilidad generalizada en intentos similares anteriores. Esta metodología, conocida ahora como el paradigma de Cranfield (llamado así por el pueblo británico donde trabajaba Cleverdon), sentó un precedente para la evaluación en recuperación de información. El uso de colecciones de prueba para evaluar la eficacia de los sistemas de recuperación ha sido la metodología *de facto* desde entonces.

A medida que el campo de la recuperación de información evoluciona, las colecciones de documentos se amplían y surgen nuevas tareas, los sistemas se enfrentan a nuevos retos. Las colecciones de pruebas deben adaptarse para reflejar los entornos de los sistemas operativos. Crear colecciones de prueba más amplias y representativas es crucial para evaluar con precisión los sistemas en condiciones reales, pero también es un proceso muy costoso. Avanzar en el desarrollo de sistemas de recuperación de información resulta cada vez más difícil si no se dispone de herramientas y metodologías adecuadas para crear nuevas colecciones de pruebas experimentales.

B.3 OBJETIVOS Y ALCANCE

Esta tesis se centra en el desarrollo de métodos innovadores para construir nuevas colecciones de prueba en el campo de recuperación de información. Nuestro objetivo es crear colecciones de una manera rentable y que sean de alta calidad, abordando una necesidad crucial en este campo. La construcción de colecciones específicas para diversas tareas de recuperación de información requiere muchos recursos, principalmente debido al gran esfuerzo humano necesario para las evaluaciones de relevancia. Este requisito encarece el proceso. Nuestro objetivo es introducir técnicas que reduzcan estos costes, apoyando así el desarrollo y la investigación más amplios en recuperación de información.

Nuestro enfoque combina conocimientos ya establecidos en el campo con métodos novedosos de selección de documentos de un corpus para su evaluación por expertos. Además, exploramos metodologías recientes para evaluar los métodos de creación de colecciones y proponemos un nuevo enfoque. Además de demostrar la eficacia de nuestros métodos mediante experimentos empíricos de laboratorio, hacemos hincapié en su aplicabilidad en el mundo real. Como conclusión de la tesis, demostramos cómo pueden aplicarse nuestras contribuciones para construir una colección práctica en el mundo real.

B.4 METODOLOGÍA DE EVALUACIÓN

A largo de toda la tesis seguiremos una metodología de evaluación estándar en el campo, usando colecciones establecidas y métricas de referencia que nos permitan evaluar de manera fehaciente si los resultados obtenidos suponen una mejora con respecto a los métodos de referencia. Más concretamente, centraremos la evaluación de los métodos de adjudicación en la fiabilidad, imparcialidad y reusabilidad de los juicios de relevancia creados con dichos métodos. A continuación introducimos someramente los aspectos de evaluación relevantes con respecto al trabajo realizado en todos los capítulos. Sin embargo, el protocolo de evaluación seguido en cada uno de ellos está explicado en más detalle en los propios capítulos, con el objetivo de hacerlos lo más autocontenidos posible.

En este sentido, la fiabilidad consiste en evaluar si los juicios creados por un método de adjudicación son capaces de mantener el *ranking* de sistemas dado por los juicios oficiales de la colección. La imparcialidad consiste en evaluar si, en ese mismo *ranking*, algún participante es tratado de manera significativamente diferente. Por último, la reusabilidad consiste en evaluar si

los juicios creados con los métodos de adjudicación son capaces de evaluar correctamente las *runs* que no fueron empleadas para construir esos propios juicios, de tal manera que podamos asegurar que estos juicios valdrán para evaluar nuevos sistemas que sean desarrollados en el futuro. La evaluación de estos aspectos es la metodología común en el campo (Sanderson and Joho 2004; Voorhees 2002, 2018; Zobel 1998).

Para evaluar todos estos aspectos emplearemos colecciones de evaluación establecidas en el campo. Concretamente, emplearemos las colecciones de TREC-5, TREC-6, TREC-7, TREC-8, TREC-9, DL21 (documentos), CDS14, CDS15, CDS16, y ROBUST04. Todas estas colecciones son el resultado de diferentes tracks de TREC a lo largo de los años. Además, también utilizaremos métricas estándar en el campo, como AP o NDCG.

B.5 ESTRUCTURA

Este manuscrito está dividido en cinco partes con siete capítulos. El capítulo actual presenta la introducción a este trabajo. El capítulo 2 introduce conceptos básicos de recuperación de información y trabajos relacionados. Los capítulos 3, 4, 5 y 6 contienen las aportaciones novedosas de esta tesis. Hemos intentado que estos capítulos sean lo más autocontenidos posible, de forma que sean fáciles de entender sólo con la información presentada en el capítulo 2, para facilitar su legibilidad. Por último, el capítulo 7 concluye recopilando las principales conclusiones de esta tesis y proponiendo algunas ideas para futuros trabajos. Más detalladamente, los contenidos de cada parte y de cada capítulo son:

PARTE I Esta primera parte incluye el capítulo 1, en el que se presenta esta tesis, y el capítulo 2, en el que se introducen los conceptos básicos pertinentes y se analizan los trabajos de referencia.

PARTE II En esta parte exploramos algunos métodos novedosos para construir nuevas colecciones de pruebas de recuperación en situaciones de escasez de recursos. Por un lado, en el capítulo 3, presentamos una novedosa metodología para crear ejecuciones sintéticas a partir de las cuales agrupar documentos para obtener nuevos juicios de relevancia cuando no se dispone de participantes reales. Por otro lado, en el capítulo 4, exploramos el uso de la retroalimentación de relevancia real para priorizar los documentos agrupados, con el objetivo de reducir el número de evaluaciones necesarias para construir un conjunto de juicios reutilizables.

PARTE III Proponemos, en el capítulo 5, una nueva forma de evaluar los métodos de adjudicación para construir nuevas colecciones de prueba. En particular, argumentamos que los métodos existentes pierden una parte de la imagen completa al mirar sólo cómo estos métodos son capaces de preservar la clasificación de los envíos agrupados. Para llenar este vacío, proponemos centrarnos en la preservación de las significaciones estadísticas entre los sistemas evaluados.

PARTE IV En esta parte, que incluye el capítulo 6, exploramos una novedosa aplicación de las aportaciones presentadas en las partes anteriores para construir una nueva colección. Partiendo de las aportaciones presentadas en los capítulos 3 y 5, creamos una nueva colección que incluye contenidos relevantes de las redes sociales sobre los procesos de patrimonialización que sufren las entidades del patrimonio cultural.

PARTE V En esta última parte, que incluye el capítulo 7, resumimos nuestras principales conclusiones y contribuciones, y sugerimos algunas ideas para futuros trabajos.

B.6 CONCLUSIONES

En esta tesis, nuestro objetivo era explorar nuevos enfoques para facilitar la construcción de nuevas colecciones para la evaluación de la recuperación. Propusimos una forma de generar *runs* sintéticas cuando no se dispone de equipos participantes. También intentamos mejorar la adjudicación de los documentos con el objetivo de hacer un uso más eficiente del presupuesto de juicios. A continuación, aportamos una nueva perspectiva en la evaluación de los métodos de adjudicación, para finalmente poner en práctica nuestras propuestas en un escenario del mundo real.

A continuación ofrecemos explicaciones detalladas de todas nuestras contribuciones:

- En el capítulo 3, estudiamos cómo crear *runs* sintéticas a partir de las cuales adjudicar los documentos. Uno de los supuestos por los que *pooling* suele funcionar, a pesar de que estamos dejando una gran cantidad de documentos sin juzgar cuando construimos colecciones de prueba, es que los documentos juzgados se ponen en común a partir de un conjunto de un número relativamente alto de presentaciones que, al mismo tiempo, son diversas en términos de los documentos que están clasificando en

las primeras posiciones. Contar con un número suficiente y diverso de sistemas participantes no siempre es factible y no todo el mundo dispone de recursos suficientes para organizar un taller como TREC. Con nuestra estrategia básica y sencilla para crear automáticamente ejecuciones a partir de un corpus de documentos—utilizando variaciones de consulta generadas automáticamente y modelos de recuperación conocidos—, simulamos la realización de adjudicaciones comparables en calidad a las creadas a partir de talleres TREC. En concreto, con un método de adjudicación sólido como MTF, obtuvimos correlaciones de clasificación superiores a 0.9 con respecto a la clasificación oficial de sistemas.

- En el capítulo 4 examinamos la adaptación de varios modelos de retroalimentación de relevancia a la tarea de adjudicación de documentos. Hemos comprobado que la retroalimentación de relevancia es útil para mejorar el rendimiento de los métodos de adjudicación existentes. En concreto, nuestros experimentos demostraron que éramos capaces de mejorar la proporción en la que aparecen los documentos relevantes a la hora de emitir nuevos juicios. Además, estos algoritmos eran, como mínimo, tan competitivos como los algoritmos de referencia en términos de fiabilidad, equidad y reutilización. Esto concluye que nuestros métodos propuestos son una alternativa rentable para construir nuevas colecciones de evaluación de la recuperación. Curiosamente, también descubrimos que el mejor modelo de retroalimentación, tanto para el *ranking* de todos los documentos adjudicados como el *ranking* de cada *run*, era el modelo de minimización de la divergencia (DMM) y que las consultas largas (*título* + *descripción*) con valores de interpolación bajos funcionaban mejor para esta tarea concreta de adjudicación de documentos.

- En el capítulo 5 propusimos una metodología para evaluar los métodos de adjudicación. Nuestra propuesta se basó en la búsqueda de la estabilidad en la preservación de las diferencias estadísticamente significativas entre los sistemas. A continuación, aplicamos esta metodología para evaluar los métodos de adjudicación más avanzados. Con este trabajo, encontramos varios resultados interesantes. El primero es que las pérdidas en el τ de Kendall tradicional se deben principalmente a intercambios en pares que no son realmente significativamente diferentes. Así pues, estas caídas afectan a una parte menos interesante de la clasificación. También hemos observado que un aumento del número de documentos relevantes no siempre produce mejores detecciones de pares significativamente diferentes. Esto sugiere que no todos los documentos relevantes de un conjunto son igualmente discriminatorios. Nuestra metodología también nos permitió demostrar

que, cuando se centran en los pares significativamente diferentes, los métodos basados en bandidos son capaces de construir juicios reutilizables, algo que no estaba claro en experimentos anteriores (Voorhees 2018). En general, nuestro método nos permitió demostrar qué método de adjudicación funciona mejor en condiciones específicas, por qué y en qué se diferencia de otros métodos.

- Por último, en el capítulo 6, pusimos en práctica algunas de las aportaciones de los capítulos anteriores. Empleamos nuestra metodología para crear *runs* de participantes sintéticas y métodos de adjudicación para construir una nueva colección para una tarea y un dominio específicos. En concreto, construimos un nuevo conjunto de datos sobre ataques y disturbios a entidades patrimoniales reflejados en redes sociales. Nuestro objetivo era proporcionar a futuros investigadores una instantánea del contenido en las redes sociales sobre la patrimonialización de estas entidades del patrimonio cultural, y las opiniones y visiones de los usuarios de las redes sociales sobre estos procesos. Como resultado de nuestro trabajo, liberamos una colección que incluye más de 260 000 textos relevantes con todo tipo de opiniones, visiones y actitudes sobre el conflicto racial y el patrimonio implicado, en distintas zonas del mundo y por perfiles de personas muy diferentes.

B.7 MIRANDO AL FUTURO

De las aportaciones presentadas en esta tesis doctoral se derivan varias oportunidades de investigación. A continuación echamos un vistazo al futuro y proponemos brevemente algunas de ellas:

- En cuanto a la simulación de *runs* sintéticas para la adjudicación de documentos, algo que propusimos y evaluamos en el capítulo 3, prevemos varias líneas que podrían investigarse. La vía más natural es incluir más y mejores modelos de recuperación para generar nuevas *runs*. En los últimos años, los avances en el campo del Procesamiento del Lenguaje Natural sobre *embeddings* contextualizados y grandes modelos de lenguaje han llevado al desarrollo de nuevos sistemas de recuperación de información, como los basados en *transformers*, que se comportan de forma muy diferente a los tradicionales (Craswell, Mitra, Yilmaz, Campos and Lin 2021; Craswell, Mitra, Yilmaz, Campos, Voorhees et al. 2021; Lin et al. 2021; Voorhees, Soboroff et al. 2022). Nos interesa evaluar la adición de este tipo de modelos a las ejecuciones sintéticas, ya que nuestro trabajo sólo incluía modelos tradicionales. En cuanto a la generación automática de variaciones de con-

sulta, queremos explorar enfoques más robustos para mejorar la calidad de las ejecuciones generadas y, por tanto, la calidad de los pools. Con este objetivo, también exploraremos modelos basados en *transformers* utilizados en áreas afines para la generación de consultas (Alaofi et al. 2023; Penha et al. 2022).

- En cuanto a nuestro trabajo sobre modelos estadísticos de retroalimentación para la adjudicación de documentos, allana el camino para futuras investigaciones sobre el uso de la retroalimentación de relevancia para la adjudicación de documentos. Los buenos resultados obtenidos empleando modelos populares de retroalimentación de relevancia indican que puede haber margen de mejora utilizando otros métodos que exploren técnicas diferentes. Por ello, tenemos previsto investigar aquí otros enfoques. En particular, queremos probar el comportamiento de la realimentación de relevancia con factorización matricial y métodos lineales (Valcarce et al. 2018a,b). También tenemos previsto investigar un método basado en Aprendizaje por Refuerzo (Montazeralghaem et al. 2020) así como variaciones de modelos de relevancia (Parapar and Barreiro 2011; Roy et al. 2019). Por último, en este escenario, existe una retroalimentación negativa explícita (documentos juzgados como no relevantes). Por lo tanto, creemos que sería interesante estudiar el uso de la retroalimentación negativa de relevancia en la estimación de los modelos utilizados para *reranking* (Wang et al. 2008).

- En el capítulo 5 propusimos una metodología más fiable para evaluar métodos de adjudicación de bajo coste para construir nuevas colecciones. A continuación, aplicamos esta metodología para realizar una evaluación exhaustiva de varios métodos de adjudicación. Creemos que a partir de las conclusiones de este trabajo hay muchas direcciones que se pueden explorar en el futuro. En esta tesis, hemos centrado nuestro análisis en dos de las colecciones más comunes utilizados en la evaluación de recuperación de información y dos de las métricas más empleadas, a saber, AP y NDCG. Creemos que sería interesante ampliarlo a más colecciones y más métricas. Hemos observado que el número de documentos relevantes detectados por un método no aumenta necesariamente el número de pares significativamente diferentes detectados. Esto sugiere que no todos los documentos relevantes de un conjunto son igual de discriminatorios. Esto abre al menos dos investigaciones interesantes que merecería la pena explorar. La primera es estudiar qué documentos relevantes serían óptimos para un *pool*, mientras que hasta ahora la atención se ha centrado en cuántos y qué tópicos utilizar en la evaluación (Buckley and Voorhees 2000; Sakai

2016b; Sanderson and Zobel 2005; Voorhees 2009; Voorhees and Buckley 2002). La segunda es estudiar qué factores son los que más influyen en el número de acuerdos activos obtenidos con pools reducidos, por ejemplo, la posición de los documentos más discriminatorios en las *runs*.

- En cuanto a nuestro trabajo sobre la creación de colecciones para tareas novedosas y, en particular, el conjunto de datos que publicamos sobre procesos de patrimonialización de entidades patrimoniales, proponemos aquí varias ideas que podrían explorarse. Tenemos previsto utilizar técnicas de procesamiento del lenguaje natural y minería de textos en la colección para identificar y extraer sitios o entidades patrimoniales. Este análisis nos permitirá elaborar mapas sobre dónde se han producido los ataques y qué entidades patrimoniales o acontecimientos históricos están implicados. También tenemos previsto enriquecer la colección con información sobre los perfiles de los usuarios de la red social cuyas publicaciones están en nuestra colección de referencia. Para ello, añadiremos a la colección de referencia publicada todas las publicaciones de todos los subreddits de cada usuario que participe en la colección actual. Esto nos permitirá disponer de un historial personal de actividad en la red social de cada uno de los usuarios que expresaron su opinión sobre las tensiones raciales y los ataques al patrimonio: qué otros temas les interesan y qué publican, el registro de sus interacciones, su experiencia en la red social. Esta información personal tiene un valor complementario para los estudios centrados en las personas: ¿Qué tipo de personas han publicado sobre las tensiones en torno a la raza y el patrimonio? ¿Qué más podemos saber sobre ellas?

REFERENCES

Nasreen Abdul-Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark D. Smucker and Courtney Wade (2004).
UMass at TREC 2004: Novelty and HARD. In: *Proceedings of TREC 2004*, pp. 1–13. URL: https://trec.nist.gov/pubs/trec13/papers/umass.novelty.hard.pdf (cited on page 35).

Marwah Alaofi, Luke Gallagher, Mark Sanderson, Falk Scholer and Paul Thomas (2023).
Can Generative LLMs Create Query Variants for Test Collections? An Exploratory Study. In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '23. Taipei, Taiwan: Association for Computing Machinery, pp. 1869–1873. DOI: 10.1145/3539618.3591960 (cited on pages 105, 120).

James Allan, Donna K. Harman, Evangelos Kanoulas, Dan Li, Christophe Van Gysel and Ellen M. Voorhees (2017).
TREC 2017 Common Core Track Overview. In: *Proceedings of The Twenty-Sixth Text REtrieval Conference*. TREC '17. Gaithersburg, Maryland, USA: NIST Special Publication 500-324. URL: https://trec.nist.gov/pubs/trec26/papers/Overview-CC.pdf (cited on page 83).

Bahadir Altun and Mucahid Kutlu (2020).
Building Test Collections Using Bandit Techniques: A Reproducibility Study. In: *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. CIKM '20. Virtual Event, Ireland: Association for Computing Machinery, pp. 1953–1956. DOI: 10.1145/3340531.3412121 (cited on pages 17, 24, 39, 41, 42, 65).

Jason Arday (2021).
It's the end of the World as we know it: Racism as a global killer of Black people and their emancipatory freedoms. In: *Educational Philosophy and Theory* 53.14, pp. 1418–1420. DOI: 10.1080/00131857.2020.1782722 (cited on pages 93, 94).

Javed A. Aslam, Virgil Pavlu and Robert Savell (2003).
A Unified Model for Metasearch, Pooling, and System Evaluation. In: *Proceedings of the 12th ACM International Conference on Information and*

Knowledge Management. CIKM '03. New Orleans, LA, USA: Association for Computing Machinery, pp. 484–491. DOI: 10.1145/956863.956953 (cited on page 65).

Krisztian Balog and Robert Neumayer (2013).
A Test Collection for Entity Search in DBpedia. In: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '13. Dublin, Ireland: ACM, pp. 737–740. DOI: 10.1145/2484028.2484165 (cited on page 18).

D. Banks, Paul Over and N.-F. Zhang (1999).
Blind Men and Elephants: Six Approaches to TREC data. In: *Information Retrieval Journal* 1.1, pp. 7–34. DOI: 10.1023/A:1009984519381 (cited on page 84).

Moumita Basu, Saptarshi Ghosh and Kripabandhu Ghosh (2018).
Overview of the FIRE 2018 Track: Information Retrieval from Microblogs during Disasters (IRMiDis). In: *Proceedings of the 10th Annual Meeting of the Forum for Information Retrieval Evaluation*. FIRE '18. Gandhinagar, India: Association for Computing Machinery, pp. 1–5. DOI: 10.1145/3293339.3293340 (cited on page 18).

Rodger Benham, Luke Gallagher, Joel Mackenzie, Tadele T. Damessie, Ruey-Cheng Chen, Falk Scholer and J. Shane Culpepper (2017).
RMIT at the TREC CORE Track. In: *Proceedings of The Twenty-Sixth Text REtrieval Conference*. TREC '17. Gaithersburg, Maryland, USA: NIST Special Publication 500-324. URL: https://trec.nist.gov/pubs/trec26/papers/RMIT-CC.pdf (cited on page 20).

Tobias Blanke, Michael Bryant and Mark Hedges (2013).
Back to our data-experiments with nosql technologies in the humanities. In: *2013 IEEE International Conference on Big Data*. IEEE, pp. 17–20 (cited on page 91).

Chris Buckley, Darrin Dimmick, Ian Soboroff and Ellen M. Voorhees (Aug. 2007).
Bias and the limits of pooling for large collections. In: *Information Retrieval Journal* 10.6, pp. 491–508. DOI: 10.1007/s10791-007-9032-x (cited on page 17).

Chris Buckley and Ellen M. Voorhees (2000).

Evaluating Evaluation Measure Stability. In: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '00. Athens, Greece: Association for Computing Machinery, pp. 33–40. DOI: 10.1145/345508.345543 (cited on pages 70, 106, 120).

Chris Buckley and Ellen M. Voorhees (2005).
Retrieval System Evaluation. In: *TREC: Experiment and Evaluation in Information Retrieval*. Ed. by Ellen M. Voorhees and Donna K. Harman. The MIT Press, pp. 53–78 (cited on page 65).

Claudio Carpineto and Giovanni Romano (2012).
A Survey of Automatic Query Expansion in Information Retrieval. In: *ACM Computing Surveys* 44.1. DOI: 10.1145/2071389.2071390 (cited on page 33).

Ben Carterette (2012).
Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments. In: *ACM Transactions on Information Systems* 30.1. DOI: 10.1145/2094072.2094076 (cited on pages 45, 63, 84).

Ben Carterette (2017).
But Is It Statistically Significant? Statistical Significance in IR Research, 1995-2014. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '17. Shinjuku, Tokyo, Japan: Association for Computing Machinery, pp. 1125–1128. DOI: 10.1145/3077136.3080738 (cited on page 84).

Cyril W. Cleverdon (1962).
Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems. Aslib Cranfield Research Project (cited on pages 8, 18).

Gordon V. Cormack and Thomas R. Lynam (2006).
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. Seattle, Washington, USA: Association for Computing Machinery, pp. 533–540. DOI: 10.1145/1148170.1148262 (cited on page 84).

Gordon V. Cormack and Thomas R. Lynam (2007a).

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. Amsterdam, The Netherlands: Association for Computing Machinery, pp. 837–838. DOI: 10.1145/1277741.1277934 (cited on page 17).

Gordon V. Cormack and Thomas R. Lynam (2007b). Validity and Power of T-Test for Comparing MAP and GMAP. In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '07. Amsterdam, The Netherlands: Association for Computing Machinery, pp. 753–754. DOI: 10.1145/1277741.1277892 (cited on page 84).

Gordon V. Cormack, Christopher R. Palmer and Charles L. A. Clarke (1998). 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. Melbourne, Australia: Association for Computing Machinery, pp. 282–289. DOI: 10.1145/290941.291009 (cited on pages 10, 17, 24, 41, 52, 65, 83).

J. Cortés-Vázquez, G. Jiménez-Esquinas and C. Sánchez-Carretero (2017). Heritage and participatory governance: An analysis of political strategies and social fractures in Spain. In: *Anthropology Today* 33.1, pp. 15–18. DOI: https://doi.org/10.1111/1467-8322.12324 (cited on page 94).

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos and Jimmy Lin (2021). Overview of the TREC 2021 Deep Learning Track. In: *Proceedings of the Thirtieth Text REtrieval Conference*. Gaithersburg, Maryland, USA: NIST Special Publication 500-335. URL: https://trec.nist.gov/pubs/trec30/papers/Overview-DL.pdf (cited on pages 17, 64, 105, 119).

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Ellen M. Voorhees and Ian Soboroff (2021). TREC Deep Learning Track: Reusable Test Collections in the Large Data Regime. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '21. New York, NY, USA: Association for Computing Machinery, pp. 2369–2375. DOI: 10.1145/3404835.3463249 (cited on pages 17, 43, 83, 105, 119).

Adam Davidson-Harden (2021).

"I can't breathe": Praxis, parrhesia and the current historical moment. In: *Educational Philosophy and Theory* 53.13, pp. 1311–1315. DOI: 10.1080/00131857.2020.1779580 (cited on pages 93, 94).

Guglielmo Faggioli and Nicola Ferro (2021).
System Effect Estimation by Sharding: A Comparison Between ANOVA Approaches to Detect Significant Differences. In: *Proceedings of 43rd European Conference on IR Research*. ECIR '21. Cham: Springer International Publishing. DOI: 10.1007/978-3-030-72240-1_3 (cited on page 61).

Nicola Ferro and Mark Sanderson (2019).
Improving the Accuracy of System Performance Estimation by Using Shards. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'19. Paris, France: Association for Computing Machinery, pp. 805–814. DOI: 10.1145/3331184.3338062 (cited on pages 13, 84).

Nicola Ferro and Mark Sanderson (2022).
How Do You Test a Test? A Multifaceted Examination of Significance Tests. In: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. WSDM '22. Virtual Event, AZ, USA: Association for Computing Machinery, pp. 280–288. DOI: 10.1145/3488560.3498406 (cited on pages 13, 61, 62, 84).

Edward A. Fox and Joseph A. Shaw (1993).
Combination of Multiple Searches. In: *Proceedings of The Second Text REtrieval Conference*. TREC '2. Gaithersburg, Maryland, USA: NIST Special Publication 500-215, pp. 243–252 (cited on page 52).

Norbert Fuhr (2018).
Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. In: *SIGIR Forum* 51.3, pp. 32–41. DOI: 10.1145/3190580.3190586 (cited on page 59).

Bilal Ghanem, Jihen Karoui, Farah Benamara, Véronique Moriceau and Paolo Rosso (2019).
IDAT at FIRE2019: Overview of the Track on Irony Detection in Arabic Tweets. In: *Proceedings of the 11th Forum for Information Retrieval Evaluation*. FIRE '19. Kolkata, India: Association for Computing Machinery, pp. 10–13. DOI: 10.1145/3368567.3368585 (cited on page 18).

Mandy Tompkins Gibson and Gabriel A. Reich (2017).
Confederate Monuments: Heritage, Racism, Anachronism, and Who Gets to Decide? In: *Social Education* 81.6, pp. 356–361 (cited on page 94).

Cesar Gonzalez-Perez, Patricia Martín-Rodilla, Cesar Parcero-Oubiña, Pastor Fábrega-Álvarez and Alejandro Güimil-Fariña (2012).
Extending an Abstract Reference Model for Transdisciplinary Work in Cultural Heritage. In: *Metadata and Semantics Research*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 190–201. DOI: 10.1007/978-3-642-35233-1_20 (cited on page 96).

Hoda Hamouda, Jessica Bushey, Victoria Lemieux, James Stewart, Corinne Rogers, James Cameron, Ken Thibodeau and Chen Feng (2019).
Extending the Scope of Computational Archival Science: A Case Study on Leveraging Archival and Engineering Approaches to Develop a Framework to Detect and Prevent "Fake Video". In: *2019 IEEE International Conference on Big Data (Big Data)*, pp. 3087–3097. DOI: 10.1109/BigData47090.2019.9006170 (cited on page 91).

Donna K. Harman (2011).
Information Retrieval Evaluation. In: *Synthesis Lectures on Information Concepts, Retrieval, and Services* 3.2, pp. 1–119. DOI: 10.2200/S00368ED1V01Y201105ICR019 (cited on pages 9, 57).

Claudia Hauff, Djoerd Hiemstra, Franciska de Jong and Leif Azzopardi (2009).
Relying on Topic Subsets for System Ranking Estimation. In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. CIKM '09. Hong Kong, China: Association for Computing Machinery, pp. 1859–1862. DOI: 10.1145/1645953.1646249 (cited on page 70).

Hussein Hazimeh and ChengXiang Zhai (2015).
Axiomatic Analysis of Smoothing Methods in Language Models for Pseudo-Relevance Feedback. In: *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*. ICTIR '15. Northampton, Massachusetts, USA: Association for Computing Machinery, pp. 141–150. DOI: 10.1145/2808194.2809471 (cited on pages 36, 44).

Yosef Hochberg and Ajit C. Tamhane (1987).
Multiple Comparison Procedures. John Wiley & Sons, USA. DOI: 10.1002/9780470316672 (cited on page 62).

Mehdi Hosseini, Ingemar J. Cox, Natasa Milic-Frayling, Milad Shokouhi and Emine Yilmaz (2012).
An Uncertainty-Aware Query Selection Model for Evaluation of IR Systems. In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '12. Portland, Oregon, USA: Association for Computing Machinery, pp. 901–910. DOI: 10.1145/2348283.2348403 (cited on page 70).

Jason C. Hsu (1996).
Multiple Comparisons. Theory and methods. Chapman and Hall/CRC, USA. DOI: 10.1201/b15074 (cited on page 62).

David Hull (1993).
Using Statistical Testing in the Evaluation of Retrieval Experiments. In: *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '93. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, pp. 329–338. DOI: 10.1145/160688.160758 (cited on page 84).

IEEE Big Data CAS Workshop (2022). https://ai-collaboratory.net/cas/cas-workshops/2022-7th-cas-workshop. 17/08/2023 (Last Access) (cited on page 90).

IEEE Big Data CAS Workshop (2023). https://ai-collaboratory.net/cas/cas-workshops/2023-8th-cas-workshop. 17/08/2023 (Last Access) (cited on page 90).

Jelani Ince, Fabio Rojas and Clayton A. Davis (2017).
The social media response to Black Lives Matter: how Twitter users interact with Black Lives Matter through hashtag use. In: *Ethnic and Racial Studies* 40.11, pp. 1814–1830. DOI: 10.1080/01419870.2017.1334931 (cited on page 94).

Kalervo Järvelin and Jaana Kekäläinen (2002).
Cumulated Gain-Based Evaluation of IR Techniques. In: *ACM Transactions on Information Systems* 20.4, pp. 422–446. DOI: 10.1145/582415.582418 (cited on page 65).

Maurice G. Kendall (1938).
A new measure of rank correlation. In: *Biometrika* 30.1/2, pp. 81–93. DOI: 10.2307/2332226 (cited on pages 11, 23, 42, 59, 83).

Maurice G. Kendall (1948).
Rank Correlation Methods. Charles Griffin and Company Limited (cited on pages 11, 42, 59, 83).

Allen Kent, Madeline M. Berry, Fred U. Luehrs and J. W. Perry (1955).
Machine literature searching VIII. Operational criteria for designing information retrieval systems. In: *American Documentation* 6.2, pp. 93–101. DOI: https://doi.org/10.1002/asi.5090060209 (cited on page 18).

Matthew Kirschenbaum, Richard Ovenden, Gabriela Redwine and Rachel Donahue (2010).
Digital forensics and born-digital content in cultural heritage collections. Council on Library and Information Resources. URL: https://www.clir.org/wp-content/uploads/sites/6/pub149.pdf (cited on pages 90, 91).

John Lafferty and ChengXiang Zhai (2001).
Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '01. New Orleans, Louisiana, USA: Association for Computing Machinery, pp. 111–119. DOI: 10.1145/383952.383970 (cited on page 34).

Victor Lavrenko and W. Bruce Croft (2001).
Relevance Based Language Models. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '01. New Orleans, Louisiana, USA: Association for Computing Machinery, pp. 120–127. DOI: 10.1145/383952.383972 (cited on page 35).

Myeong Lee, Yuheng Zhang, Shiyun Chen, Edel Spencer, Jhon Dela Cruz, Hyeonggi Hong and Richard Marciano (2017).
Heuristics for assessing Computational Archival Science (CAS) research: The case of the human face of big data project. In: *2017 IEEE International Conference on Big Data (Big Data)*, pp. 2262–2270. DOI: 10.1109/BigData.2017.8258179 (cited on pages 90, 91).

Dan Li and Evangelos Kanoulas (2017).
Active Sampling for Large-Scale Information Retrieval Evaluation. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. Ed. by Ee-Peng Lim et al. CIKM '17. Singapore, Singapore:

Association for Computing Machinery, pp. 49–58. DOI: 10.1145/3132847.3133015 (cited on page 83).

Jimmy Lin, Rodrigo Nogueira and Andrew Yates (2021). Pretrained Transformers for Text Ranking: BERT and Beyond. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers. DOI: 10.2200/S01123ED1V01Y202108HLT053 (cited on pages 8, 105, 119).

Aldo Lipani, David E. Losada, Guido Zuccon and Mihai Lupu (2019). Fixed-Cost Pooling Strategies. In: *IEEE Transactions on Knowledge & Data Engineering*. DOI: 10.1109/TKDE.2019.2947049 (cited on pages 17, 24).

Aldo Lipani, Guido Zuccon, Mihai Lupu, Bevan Koopman and Allan Hanbury (2016). The Impact of Fixed-Cost Pooling Strategies on Test Collection Bias. In: *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*. ICTIR '16. Newark, Delaware, USA: Association for Computing Machinery, pp. 105–108. DOI: 10.1145/2970398.2970429 (cited on page 17).

David E. Losada, Fabio Crestani and Javier Parapar (2017). eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Foundations. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. CLEF '17. Cham: Springer International Publishing, pp. 346–360. DOI: 10.1007/978-3-319-65813-1_30 (cited on page 18).

David E. Losada, Fabio Crestani and Javier Parapar (2019). Overview of eRisk 2019 Early Risk Prediction on the Internet. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. CLEF '19. Cham: Springer International Publishing, pp. 340–357. DOI: 10.1007/978-3-030-28577-7_27 (cited on page 18).

David E. Losada, Fabio Crestani and Javier Parapar (2020). eRisk 2020: Self-harm and Depression Challenges. In: *Proceedings of 42nd European Conference on IR Research*. ECIR '20. Cham: Springer International Publishing, pp. 557–563. DOI: 10.1007/978-3-030-45442-5_72 (cited on page 18).

David E. Losada, Javier Parapar and Álvaro Barreiro (2016). Feeling Lucky?: Multi-armed Bandits for Ordering Judgements in Pooling-based Evaluation. In: *Proceedings of the 31st Annual ACM Symposium*

on Applied Computing. Ed. by Sascha Ossowski. SAC '16. Pisa, Italy: Association for Computing Machinery, pp. 1027–1034. DOI: 10.1145/2851613.2851692 (cited on pages 10, 11, 17, 24, 65).

David E. Losada, Javier Parapar and Álvaro Barreiro (2017). Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. In: *Information Processing & Management* 53.5, pp. 1005–1025. DOI: 10.1016/j.ipm.2017.04.005 (cited on pages 39, 41, 42, 58, 65, 83).

Yuanhua Lv and ChengXiang Zhai (2009). A Comparative Study of Methods for Estimating Query Language Models with Pseudo Feedback. In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. CIKM '09. Hong Kong, China: Association for Computing Machinery, pp. 1895–1898. DOI: 10.1145/1645953.1646259 (cited on pages 34, 35).

Yuanhua Lv and ChengXiang Zhai (2014). Revisiting the Divergence Minimization Feedback Model. In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. CIKM '14. Shanghai, China: Association for Computing Machinery, pp. 1863–1866. DOI: 10.1145/2661829.2661900 (cited on page 36).

Richard Marciano, Victoria Lemieux, Mark Hedges, Maria Esteva, William Underwood, Michael Kurtz and Mark Conrad (2018). Archival records and training in the age of big data. In: *Re-Envisioning the MLS: Perspectives on the future of library and information science education*. Emerald Publishing Limited (cited on page 90).

Patricia Martín-Rodilla, Marcia L. Hattori and Cesar Gonzalez-Perez (2019). Assisting Forensic Identification through Unsupervised Information Extraction of Free Text Autopsy Reports: The Disappearances Cases during the Brazilian Military Dictatorship. In: *Information* 10.7. DOI: 10.3390/info10070231 (cited on page 96).

Jasmine McNealy (2011). The privacy implications of digital preservation: social media archives and the social networks theory of privacy. In: *Elon Law Review* 3 (cited on page 91).

Lynn Meskell (2002).

Negative Heritage and Past Mastering in Archaeology. In: *Anthropological Quarterly* 75.3, pp. 557–574 (cited on page 94).

Stefano Mizzaro and Stephen Robertson (2007).
Hits hits TREC: Exploring IR Evaluation Results with Network Analysis. In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '07. Amsterdam, The Netherlands: Association for Computing Machinery, pp. 479–486. DOI: 10.1145/1277741.1277824 (cited on page 70).

Alistair Moffat, Falk Scholer and Paul Thomas (2012).
Models and Metrics: IR Evaluation as a User Process. In: *Proceedings of the Seventeenth Australasian Document Computing Symposium*. ADCS '12. Dunedin, New Zealand: Association for Computing Machinery, pp. 47–54. DOI: 10.1145/2407085.2407092 (cited on page 61).

Alistair Moffat, William Webber and Justin Zobel (2007).
Strategic System Comparisons via Targeted Relevance Judgments. In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '07. Amsterdam, The Netherlands: Association for Computing Machinery, pp. 375–382. DOI: 10.1145/1277741.1277806 (cited on pages 17, 83).

Ali Montazeralghaem, Hamed Zamani and James Allan (2020).
A Reinforcement Learning Framework for Relevance Feedback. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '20. Virtual Event, China: Association for Computing Machinery, pp. 59–68. DOI: 10.1145/3397271.3401099 (cited on pages 105, 120).

Devon Mordell (2019).
Critical questions for archives as (big) data. In: *Archivaria* 87.87, pp. 140–161 (cited on page 91).

Makoto P. Kato, Yiqun Liu, Noriko Kando and Charles L. A. Clarke, eds. (2019).
NII Testbeds and Community for Information Access Research - 14th International Conference. Lecture Notes in Computer Science. Tokyo, Japan: Springer. DOI: 10.1007/978-3-030-36805-0 (cited on page 17).

Benedicta Obodoruku (2016).

Social Networking: Information Sharing, Archiving and Privacy. In: *24th BOBCATSSS Conference Proceedings & Abstracts* (cited on page 91).

David Otero (2019).
Exploiting Pooling Methods for Building Datasets for Novel Tasks. In: *Proceedings of the 9th PhD Symposium on Future Directions in Information Access*. FDIA '19. Milan, Italy: CEUR-WS.org, pp. 96–102. URL: http://ceur-ws.org/Vol-2537/paper-16.pdf (cited on page 89).

David Otero, Patricia Martin-Rodilla and Javier Parapar (2021).
Building Cultural Heritage Reference Collections from Social Media through Pooling Strategies: The Case of 2020's Tensions Over Race and Heritage. In: *Journal on Computing and Cultural Heritage* 15.1, pp. 1–13. DOI: 10.1145/3477604 (cited on pages 18, 89).

David Otero, Javier Parapar and Álvaro Barreiro (2020).
Beaver: Efficiently Building Test Collections for Novel Tasks. In: *Proceedings of the Joint Conference of the Information Retrieval Communities in Europe*. CIRCLE '20. Samatan, Gers, France: CEUR-WS.org. URL: http://ceur-ws.org/Vol-2621/CIRCLE20_23.pdf (cited on page 89).

David Otero, Javier Parapar and Álvaro Barreiro (2021).
The Wisdom of the Rankers: A Cost-Effective Method for Building Pooled Test Collections without Participant Systems. In: *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. SAC '21. New York, NY, USA: Association for Computing Machinery, pp. 672–680. DOI: 10.1145/3412841.3441947 (cited on page 18).

David Otero, Javier Parapar and Álvaro Barreiro (2023).
Relevance feedback for building pooled test collections. In: *Journal of Information Science*. DOI: 10.1177/01655515231171085 (cited on pages 17, 31).

David Otero, Javier Parapar and Nicola Ferro (2023).
How Discriminative Are Your Qrels? How To Study the Statistical Significance of Document Adjudication Methods. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. CIKM '23. New York, NY, USA: Association for Computing Machinery. DOI: 10.1145/3583780.3614916 (cited on pages 17, 39, 41, 42, 49, 57).

David Otero, Daniel Valcarce, Javier Parapar and Álvaro Barreiro (2019).

Building High-Quality Datasets for Information Retrieval Evaluation at a Reduced Cost. In: *Proceedings of The 2nd XoveTIC Conference (XoveTIC 2019)*. MDPI. DOI: 10.3390/proceedings2019021033 (cited on page 18).

Javier Parapar and Álvaro Barreiro (2011).
Promoting Divergent Terms in the Estimation of Relevance Models. In: *Proceedings of the Third International Conference on Advances in Information Retrieval Theory*. ICTIR'11. Bertinoro, Italy: Springer-Verlag, pp. 77–88. DOI: 10.1007/978-3-642-23318-0_9 (cited on pages 105, 120).

Javier Parapar, David E. Losada and Álvaro Barreiro (2021).
Testing the Tests: Simulation of Rankings to Compare Statistical Significance Tests in Information Retrieval Evaluation. In: *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. SAC '21. Virtual Event, Republic of Korea: Association for Computing Machinery, pp. 655–664. DOI: 10.1145/3412841.3441945 (cited on pages 13, 84).

Javier Parapar, David E. Losada, Manuel A. Presedo-Quindimil and Álvaro Barreiro (2020).
Using score distributions to compare statistical significance tests for information retrieval evaluation. In: *Journal of the Association for Information Science and Technology* 71.1, pp. 98–113. DOI: 10.1002/asi.24203 (cited on pages 13, 84).

Nathaniel Payne (2018).
Stirring the cauldron: redefining computational archival science (CAS) for the Big Data domain. In: *2018 IEEE International Conference on Big Data (Big Data)*, pp. 2743–2752 (cited on page 91).

Gustavo Penha, Arthur Câmara and Claudia Hauff (2022).
Evaluating the Robustness of Retrieval Pipelines with Query Variation Generators. In: *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*. Stavanger, Norway: Springer-Verlag, pp. 397–412. DOI: 10.1007/978-3-030-99736-6_27 (cited on pages 105, 120).

Anxo Pérez, Javier Parapar and Álvaro Barreiro (2022).
Automatic depression score estimation with word embedding models. In: *Artificial Intelligence in Medicine* 132, p. 102380. DOI: https://doi.org/10.1016/j.artmed.2022.102380 (cited on page 18).

Jay M. Ponte and W. Bruce Croft (1998).

A Language Modeling Approach to Information Retrieval. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '98. Melbourne, Australia: Association for Computing Machinery, pp. 275–281. DOI: 10.1145/290941.291008 (cited on pages 8, 20).

Jennifer Pybus (2013).
Social networks and cultural workers: Towards an archive for the prosumer. In: *Journal of Cultural Economy* 6.2, pp. 137–152 (cited on page 90).

Md Mustafizur Rahman, Mucahid Kutlu and Matthew Lease (2019).
Constructing Test Collections Using Multi-Armed Bandits and Active Learning. In: *The World Wide Web Conference*. WWW '19. San Francisco, CA, USA: Association for Computing Machinery, pp. 3158–3164. DOI: 10.1145/3308558.3313675 (cited on pages 58, 83).

Thorsten Ries and Gábor Palko (2019).
Born-digital archives. In: *International Journal of Digital Humanities* 1.1. Ed. by Gábor Palko, pp. 1–11 (cited on page 90).

Pilar Rivero, Iñaki Navarro and Borja Aso (2020).
Educommunication Web 2.0 for Heritage: A View From Spanish Museums. In: *Handbook of Research on Citizenship and Heritage Education*. Ed. by Emilio José Delgado-Algarra and José María Cuenca-López. Hershey, PA: IGI Global, pp. 450–471 (cited on page 89).

Stephen Robertson (2004).
Understanding inverse document frequency: on theoretical arguments for IDF. In: *Journal of Documentation* 60.5, pp. 503–520. DOI: 10.1108/00220410410560582 (cited on page 36).

Stephen Robertson and Ian Soboroff (2002).
The TREC 2002 Filtering Track Report. In: *Proceedings of The Eleventh Text REtrieval Conference (TREC 2002)*. TREC '02. Gaithersburg, Maryland, USA: NIST Special Publication 500-251, pp. 27–39. DOI: 10.6028/NIST.SP.500-251 (cited on page 52).

Stephen Robertson and Hugo Zaragoza (Apr. 2009).
The Probabilistic Relevance Framework: BM25 and Beyond. In: *Foundations and Trends in Information Retrieval* 3.4, pp. 333–389. DOI: 10.1561/1500000019 (cited on pages 8, 20).

Kevin Roitero, J. Shane Culpepper, Mark Sanderson, Falk Scholer and Stefano Mizzaro (2020).
Fewer topics? A million topics? Both?! On topics subsets in test collections. In: *Information Retrieval Journal* 23.1, pp. 49–85. DOI: 10.1007/s10791-019-09357-w (cited on page 70).

Dwaipayan Roy, Sumit Bhatia and Mandar Mitra (2019).
Selecting Discriminative Terms for Relevance Model. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'19. Paris, France: Association for Computing Machinery, pp. 1253–1256. DOI: 10.1145/3331184.3331357 (cited on pages 105, 120).

Tetsuya Sakai (2016a).
Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015. In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '16. Pisa, Italy: Association for Computing Machinery, pp. 5–14. DOI: 10.1145/2911451.2911492 (cited on page 84).

Tetsuya Sakai (2016b).
Topic set size design. In: *Information Retrieval Journal* 19.3, pp. 256–283. DOI: 10.1007/s10791-015-9273-z (cited on pages 70, 106, 120).

Tetsuya Sakai (2018).
Laboratory Experiments in Information Retrieval. Vol. 40. The Information Retrieval Series. Springer. DOI: 10.1007/978-981-13-1199-4 (cited on pages 45, 62, 63).

Tetsuya Sakai (2021).
On Fuhr's Guideline for IR Evaluation. In: *SIGIR Forum* 54.1. DOI: 10.1145/3451964.3451976 (cited on page 59).

Tetsuya Sakai, Noriko Kando, Chuan-Jie Lin, Teruko Mitamura, Hideki Shima, Dong-Hong Ji, Kuang-hua Chen and Eric Nyberg (2008).
Overview of the NTCIR-7 ACLIA IR4QA Task. In: *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*. National Institute of Informatics (NII), pp. 77–114. URL: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/NTCIR7/C1/IR4QA/01-NTCIR7-OV-IR4QA-SakaiT.pdf (cited on page 65).

Tetsuya Sakai, Sijie Tao and Zhaohao Zeng (2022).
Relevance Assessments for Web Search Evaluation: Should We Randomise or Prioritise the Pooled Documents? In: *ACM Transactions on Information Systems* 40.4. DOI: 10.1145/3494833 (cited on page 24).

G. Salton, A. Wong and C. S. Yang (1975).
A Vector Space Model for Automatic Indexing. In: *Communications of the ACM* 18.11, pp. 613–620. DOI: 10.1145/361219.361220 (cited on pages 8, 20).

Gerard Salton (1968).
Automatic Information Organization and Retrieval. McGraw Hill Text (cited on page 7).

Nubras Samayeen, Adrian Wong and Cameron McCarthy (2022).
Space to breathe: George Floyd, BLM Plaza and the monumentalization of divided American urban landscapes. In: *Educational Philosophy and Theory* 54.14, pp. 2341–2351. DOI: 10.1080/00131857.2020.1795980 (cited on page 94).

Cristina Sanchez (2013).
Significance and social value of Cultural Heritage: Analyzing the fractures of Heritage. In: *Science and Technology for the Conservation of Cultural Heritage*. DOI: 10.1201/b15577 (cited on page 94).

Mark Sanderson (2010).
Test Collection Based Evaluation of Information Retrieval Systems. In: *Foundations and Trends in Information Retrieval* 4.4, pp. 247–375. DOI: 10.1561/1500000009 (cited on pages 32, 57).

Mark Sanderson and Hideo Joho (2004).
Forming Test Collections with No System Pooling. In: *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '04. Sheffield, United Kingdom: Association for Computing Machinery, pp. 33–40. DOI: 10.1145/1008992.1009001 (cited on pages 30, 52, 116).

Mark Sanderson and Justin Zobel (2005).
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. Salvador,

Brazil: Association for Computing Machinery, pp. 162–169. DOI: 10.1145/1076034.1076064 (cited on pages 70, 84, 106, 121).

Jacques Savoy (1997).
Statistical inference in retrieval effectiveness evaluation. In: *Information Processing & Management* 33.4, pp. 495–512. DOI: 10.1016/S0306-4573(97)00027-7 (cited on page 84).

Guenther Schefbech, Dimitris Spiliotopoulos and Thomas Risse (2012).
The Recent Challenge in Web Archiving: Archiving the Social Web. In: *Proceedings of the International Council on Archives Congress*. Brisbane, Australia, pp. 1–5 (cited on pages 90, 91).

Stephan A. Schwartz (2020).
Police brutality and racism in America. In: *EXPLORE* 16.5, pp. 280–282. DOI: 10.1016/j.explore.2020.06.010 (cited on page 94).

Ian Soboroff and Stephen Robertson (2003).
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. Toronto, Canada: Association for Computing Machinery, pp. 243–250. DOI: 10.1145/860435.860481 (cited on page 52).

K. Spärck Jones and Cornelis J. van Rijsbergen (1975).
Report on the need for and provision of an 'ideal' information retrieval test collection. In: *Computer Laboratory* (cited on pages 9, 58).

H. Stančić (2018).
Computational archival science. In: *Moderna arhivistika. Časopis arhivske teorije in prakse (Journal of Archival Theory and Practice)* 1.2, pp. 323–329 (cited on page 90).

Richard S. Sutton and Andrew G. Barto (2018).
Reinforcement Learning: An Introduction. Second. Adaptive computation and Machine Learning. MIT Press (cited on pages 10, 83).

Franklyn Herbert Upward (2005).
The records continuum. In: *Archives: Recordkeeping in society*. Centre for Information Studies, pp. 197–222 (cited on page 91).

Julián Urbano, Harlley Lima and Alan Hanjalic (2019).

Statistical Significance Testing in Information Retrieval: An Empirical Analysis of Type I, Type II and Type III Errors. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'19. Paris, France: Association for Computing Machinery, pp. 505–514. DOI: 10.1145/3331184.3331259 (cited on pages 13, 84).

Julián Urbano, Mónica Marrero and Diego Martín (2013).
A Comparison of the Optimality of Statistical Significance Tests for Information Retrieval Evaluation. In: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '13. Dublin, Ireland: Association for Computing Machinery, pp. 925–928. DOI: 10.1145/2484028.2484163 (cited on pages 13, 61, 84).

Daniel Valcarce, Javier Parapar and Álvaro Barreiro (Jan. 2018a).
Document-based and Term-based Linear Methods for Pseudo-Relevance Feedback. In: *SIGAPP Applied Computing Review* 18.4, pp. 5–17. DOI: 10.1145/3307624.3307626 (cited on pages 105, 120).

Daniel Valcarce, Javier Parapar and Álvaro Barreiro (2018b).
LiMe: Linear Methods for Pseudo-Relevance Feedback. In: *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. SAC '18. Pau, France: Association for Computing Machinery, pp. 678–687. DOI: 10.1145/3167132.3167207 (cited on pages 105, 120).

Ellen M. Voorhees (2002).
The Philosophy of Information Retrieval Evaluation. In: *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum*. CLEF '01. Darmstadt, Germany: Springer, pp. 355–370. DOI: 10.1007/3-540-45691-0_34 (cited on pages 8, 10, 17, 32, 43, 57, 116).

Ellen M. Voorhees (2009).
Topic Set Size Redux. In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '09. Boston, MA, USA: Association for Computing Machinery, pp. 806–807. DOI: 10.1145/1571941.1572138 (cited on pages 70, 106, 121).

Ellen M. Voorhees (2018).
On Building Fair and Reusable Test Collections Using Bandit Techniques. In: *Proceedings of the 27th ACM International Conference on Information*

and Knowledge Management. CIKM '18. Torino, Italy: Association for Computing Machinery, pp. 407–416. DOI: 10.1145/3269206.3271766 (cited on pages 43, 49, 58, 83, 85, 104, 116, 119).

Ellen M. Voorhees (2019).
The Evolution of Cranfield. In: *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*. Cham: Springer International Publishing, pp. 45–69. DOI: 10.1007/978-3-030-22948-1_2 (cited on pages 8, 17).

Ellen M. Voorhees and Chris Buckley (2002).
The Effect of Topic Set Size on Retrieval Experiment Error. In: *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '02. Tampere, Finland: Association for Computing Machinery, pp. 316–323. DOI: 10.1145/564376.564432 (cited on pages 70, 106, 121).

Ellen M. Voorhees, Nick Craswell and Jimmy Lin (2022).
Too Many Relevants, Whither Cranfield Test Collections? In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '22. Madrid, Spain: Association for Computing Machinery. DOI: 10.1145/3477495.3531728 (cited on pages 49, 58, 83).

Ellen M. Voorhees and Donna K. Harman (2000).
Overview of the Eighth Text REtrieval Conference (TREC-8). In: *Proceedings of the Eighth Text REtrieval Conference*. Gaithersburg, Maryland, USA: NIST Special Publication 500-246, pp. 1–24. DOI: 10.6028/NIST.SP.500-246 (cited on page 64).

Ellen M. Voorhees and Donna K. Harman (2005).
TREC: Experiment and Evaluation in Information Retrieval. The MIT Press (cited on pages 17, 24, 41, 57, 58, 65).

Ellen M. Voorhees, Ian Soboroff and Jimmy Lin (2022).
Can Old TREC Collections Reliably Evaluate Modern Neural Retrieval Models? DOI: 10.48550/arXiv.2201.11086 (cited on pages 17, 64, 83, 105, 119).

Xuanhui Wang, Hui Fang and ChengXiang Zhai (2008).
A Study of Methods for Negative Relevance Feedback. In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and*

Development in Information Retrieval. SIGIR '08. Singapore, Singapore: Association for Computing Machinery, pp. 219–226. DOI: 10.1145/1390334.1390374 (cited on pages 106, 120).

William Webber, Alistair Moffat and Justin Zobel (2008).
Statistical Power in Retrieval Experimentation. In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. CIKM '08. Napa Valley, California, USA: Association for Computing Machinery, p. 57100580. DOI: 10.1145/1458082.1458158 (cited on page 84).

Jane Winters and Andrew Prescott (2019).
Negotiating the born-digital: a problem of search. In: *Archives and Manuscripts* 47.3, pp. 391–403 (cited on page 91).

Emine Yilmaz, Javed A. Aslam and Stephen Robertson (2008).
A New Rank Correlation Coefficient for Information Retrieval. In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '08. Singapore, Singapore: Association for Computing Machinery, pp. 587–594. DOI: 10.1145/1390334.1390435 (cited on pages 42, 83).

ChengXiang Zhai and John Lafferty (2001a).
A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '01. New Orleans, Louisiana, USA: Association for Computing Machinery, pp. 334–342. DOI: 10.1145/383952.384019 (cited on page 35).

ChengXiang Zhai and John Lafferty (2001b).
Model-Based Feedback in the Language Modeling Approach to Information Retrieval. In: *Proceedings of the Tenth International Conference on Information and Knowledge Management*. CIKM '01. Atlanta, Georgia, USA: Association for Computing Machinery, pp. 403–410. DOI: 10.1145/502585.502654 (cited on pages 35, 36).

ChengXiang Zhai and John Lafferty (2004).
A Study of Smoothing Methods for Language Models Applied to Information Retrieval. In: *ACM Transactions on Information Systems* 22.2, pp. 179–214. DOI: 10.1145/984321.984322 (cited on page 35).

Justin Zobel (1998).

How Reliable Are the Results of Large-Scale Information Retrieval Experiments? In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '98. Melbourne, Australia: Association for Computing Machinery, pp. 307–314. DOI: 10.1145/290941.291014 (cited on pages 10, 17, 43, 58, 116).