INCREASING THE EFFECTIVENESS OF SYSTEM-BASED EVALUATION FOR INFORMATION RETRIEVAL SYSTEMS

PRABHA A/P RAJAGOPAL

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

INCREASING THE EFFECTIVENESS OF SYSTEM-BASED EVALUATION FOR INFORMATION RETRIEVAL SYSTEMS

PRABHA A/P RAJAGOPAL

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Prabha Rajagopal

Registration/Matric No: WHA130044

Name of Degree: Doctor of Philosophy

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Increasing the Effectiveness of System-based Evaluation for Information Retrieval Systems

Field of Study: Information Retrieval (Computer Science)

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date: 8th June 2018

Subscribed and solemnly declared before,

Witness's Signature Date: 8th June 2018

Name: Sri Devi Ravana

Designation: Head
Department of Information Systems
Faculty of Computer Science & Information Technology
University of Malaya

INCREASING THE EFFECTIVENESS OF SYSTEM-BASED EVALUATION FOR INFORMATION RETRIEVAL SYSTEMS

ABSTRACT

The information retrieval system evaluation is necessary to measure and quantify the effectiveness, assess user satisfaction and acceptance of the retrieval systems, and compare the performance of the retrieval systems. The relevance judgments, system rankings, and statistical significance testing are some essentials in the evaluation. This thesis makes several contributions to the information retrieval system evaluation using test collections in three different experiments. The first experiment explored issues in relation to effort needed by users to retrieve relevant contents from documents. Real users give up easily and do not put as much effort as expert judges while retrieving relevant contents. It is unknown if deeper evaluation and wider groups of systems show variation in system rankings due to effort. The experimentation aims to generate low effort relevance judgments systematically, determine the variation of system rankings evaluated at different depth and groups of systems, and explore the effectiveness in evaluating retrieval systems using low effort relevance judgment with reduced topic sizes. Low effort relevance judgments are generated using boxplot approach and standardized readability grades. The findings reveal variation on system rankings at various evaluation depths and groups of systems while reduced topic sizes evaluation shows differing outcome. The second experiment explored issues on reliability of system rankings. Evaluation of system rankings set indicates the overall reliability but not for individual systems. Evaluation by combination of metrics signifies its versatility in fulfilling different user models. The experimentation aims to propose an approach to evaluate the reliability of individual system rankings, determine suitable combination of metrics, understand generalization of system ranking reliability to other similar metrics, identify the original systems with reliable system rankings, and validate the proposed approach. The proposed intraclass

correlation coefficient approach measures the reliability of individual system rankings using relative topic ranks. The average precision and rank-biased precision metrics are recommended for measuring reliability of individual system rankings. Most experimented metrics combinations generalize well. Highly reliable systems comprise of top and mid performing systems from the original systems ranking. Also, a strong correlation coefficient between system rankings of original and proposed approach validates the proposed reliability measurement of individual retrieval system rankings. The third experiment explored issues on the usage of averaged or cut-off topic scores for statistical significance testing. Precision at *k* metric causes varying user experience while the need for total relevant documents in average precision is infeasible on the ever-changing Web. The experimentation aims to propose an approach to overcome the inaccuracy of averaged or cut-off topic scores in statistical significance test, identify a suitable sample size, and validate the effectiveness of the proposed approach. The approach uses indivisible document-level scores for statistical significance testing. The document-level scores usage produced higher numbers of statistically significant system pairs compared to the existing method. Suitable sample size selection is necessary for achieving reliable results while a high percentage of agreement between the proposed and existing reveals the effectiveness of the proposed document-level approach.

Keywords: Information retrieval evaluation, system-oriented, test collections, batch experimentation, TREC

PENINGKATAN KEBERKESANAN PENILAIAN CAPAIAN MAKLUMAT BERASASKAN SISTEM

ABSTRAK

Penilaian sistem capaian maklumat adalah perlu bagi mengukur dan mengira keberkesanan, menilai kepuasan pengguna dan penerimaan sistem capaian, serta membandingkan prestasi bagi sistem capaian tersebut. Pertimbangan yang relevan, ranking sistem, dan pengujian signifikan statistik merupakan beberapa perkara penting dalam penilaian ini. Tesis ini memberi beberapa sumbangan kepada penilaian sistem capaian maklumat dengan menggunakan koleksi ujian dalam tiga uji kaji berbeza. Uji kaji pertama menerokai isu yang berkaitan dengan usaha yang diperlukan oleh pengguna bagi membuat capaian kandungan yang relevan dari dokumen. Tidak seperti penilai yang pakar, pengguna sebenar mudah berputus asa dan tidak berusaha bersungguh-sungguh semasa membuat capaian kandungan yang relevan. Tidak pasti sekiranya penilaian mendalam dan kumpulan sistem yang luas menunjukkan variasi dalam ranking sistem disebabkan oleh usaha. Tujuan pengujikajian ini adalah untuk menghasilkan penilaian relevan yang kurang usaha secara sistematik, menentukan variasi ranking sistem yang dinilai pada kedalaman dan kumpulan sistem yang berbeza, serta menerokai keberkesanannya dalam menilai sistem capaian dengan menggunakan penilaian relevan yang kurang usaha dengan saiz topik yang kecil. Penilaian relevan yang kurang usaha dihasilkan dengan menggunakan pendekatan *boxplot* dan gred kebolehbacaan yang diselaraskan. Dapatan menunjukkan variasi dalam ranking sistem pada pelbagai kedalaman dan kumpulan sistem penilaian, manakala penilaian saiz topik yang kecil menunjukkan keberhasilan yang berbeza. Uji kaji kedua menerokai isu tentang kebolehpercayaan bagi ranking sistem. Penilaian bagi set ranking sistem menunjukkan kebolehpercayaan menyeluruh, namun bukan untuk sistem individu. Penilaian yang dibuat berdasarkan kombinasi metrik menandakan ia serba boleh dalam memenuhi model pengguna yang berbeza. Tujuan pengujikajian ini adalah untuk mencadangkan satu pendekatan bagi menilai kebolehpercayaan ranking sistem individu, menentukan kombinasi metrik yang sesuai, memahami anggapan umum bagi kebolehpercayaan ranking sistem dengan metrik serupa yang lain, mengenal pasti sistem asal dengan ranking sistem yang boleh dipercayai, serta mengesahkan pendekatan yang dicadangkan.

Cadangan pendekatan koefisien korelasi dalam kelas yang dicadangkan mengukur kebolehpercayaan ranking sistem individu menggunakan kedudukan topik yang relatif. Metrik ketepatan purata (*average precision*) dan ketepatan yang berdasarkan kedudukan (*RBP*) disyorkan untuk mengukur kebolehpercayaan ranking sistem individu. Gabungan metrik yang paling banyak diuji kaji dibuat dengan baik. Sistem yang sangat boleh dipercayai terdiri daripada sistem prestasi atas dan pertengahan dari ranking sistem asal. Juga, koefisien korelasi yang kuat antara ranking sistem pendekatan asal dan yang dicadangkan mengesahkan pengukuran kebolehpercayaan yang dicadangkan bagi ranking sistem capaian individu. Uji kaji ketiga meneroka isu tentang penggunaan skor topik purata atau *cut-off* untuk ujian signifikan statistik. Metrik P@*k* menyebabkan pengalaman pengguna yang berbeza-beza sementara keperluan untuk jumlah dokumen yang relevan dengan ketepatan purata (*average precision*) tidak dapat diterapkan pada Web yang sentiasa berubah. Eksperimen ini bertujuan untuk mencadangkan satu pendekatan bagi mengatasi ketidaktepatan skor topik purata atau *cut-off* dalam ujian signifikan statistik, mengenal pasti saiz sampel yang sesuai, dan mengesahkan keberkesanan pendekatan yang dicadangkan. Pendekatan ini menggunakan skor tahap dokumen yang boleh terbahagi untuk ujian signifikan statistik. Penggunaan skor pada peringkat dokumen menghasilkan jumlah pasangan sistem signifikan secara statistik yang lebih tinggi berbanding dengan kaedah sedia ada. Pemilihan saiz sampel yang sesuai diperlukan untuk mencapai hasil yang boleh dipercayai sementara peratusan persetujuan yang tinggi antara yang dicadangkan dan yang sedia ada menunjukkan keberkesanan pendekatan peringkat dokumen yang dicadangkan.

Kata kunci: Penilaian capaian maklumat, berorientasikan sistem, koleksi ujian, pengujikajian kelompok, TREC

ACKNOWLEDGEMENTS

First and foremost, I extend my humble salutations to God for this life and blessings.

Most importantly, I would like to express my gratitude to my supervisor, Dr. Sri Devi Ravana for her continues support, guidance, and patience. She has been one of the reasons for my study. Her unwavering knowledge has been a guidance to the completion of my thesis. I am grateful for all that I have received from her, a thank you is just not enough.

I would like to acknowledge my family for their continuous support and bearing with me through the years. My dad, Mr. Rajagopal has always been and will always be my pillar. This degree is for you 'appa'. My life partner, Raja Theva Krishnan, thank you so much for your tolerance. Despite all difficulties, you stand by me without fail. And to my sons, Raja Ayngaran Prabu and Raja Skandhan Praba, you are my precious, my motivation and my purpose of life. Thank you Ayngaran for enjoying your trips to the university with me and Skandhan for being cooperative at just few months old.

Not to forget my peers, Parnia, Masumeh, Zhang and Harshit who have been a part of my journey. It has been fun filled with your presence. My friends who have always been concerned about my study and giving encouraging words, you are the best. Also, to the academic and non-academic members of the Faculty of Computer Science and Information Technology, University of Malaya, a big thank you for your continued support and assistance.

The completion of this thesis would never have been possible without each and every one of you. Thank you from the bottom of my heart for being a part of it.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

IR	:	Information Retrieval
AP	:	Average Precision
MAP	:	Mean Average Precision
P@*k*	:	Precision at cut-off *k*
RBP	:	Rank-biased Precision
TREC	:	Text REtrieval Conference
ICC	:	Intraclass Correlation Coefficient
ANOVA	:	Analysis of Variance
VLC	:	Very Large Corpus
RPrec	:	Recall-Precision
GMAP	:	Geometric Mean Average Precision
ARI	:	Automated Readability Index
CLI	:	Coleman-Liau Index
NDCG	:	Normalized Discounted Cumulative Gain
ERR	:	Expected Reciprocal Rank

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

Information is available in abundance in different modes such as in books, journals, documents, web pages, newspapers, magazines and much more. This information can be presented in various forms such as text, audio, images or videos. People look for information to fulfill some query or doubt or gain some knowledge about something specific. However, such large amount of information can be overwhelming to start in the first place. Especially, seeking information on the Web which grows and increases by the day. A proper mechanism for indexing and retrieval could lead to success in channeling the right information to the query from the user.

The information retrieval evaluation involves measuring the performance of the retrieval systems, and is divided broadly into the system-oriented and user-oriented approaches. The user-oriented evaluation focuses on the interaction of the user with the information retrieval systems, and the user's context and situation. It takes into account the usability of the interface, user's behavior towards the search process, and user satisfaction but can be costly, time-consuming, and uncontrollable while only partially capturing the user experience.

The system-oriented evaluation usually takes place in a laboratory setting with minimal end-user involvement, and shorter experimentation time and turnaround. It is also cost-effective, quantifiable and repeatable. The system-oriented evaluation involves measuring the retrieval systems using a test collection consisting of a document corpus, topics or queries (the information need) and relevance judgment as shown in Figure 1.1.

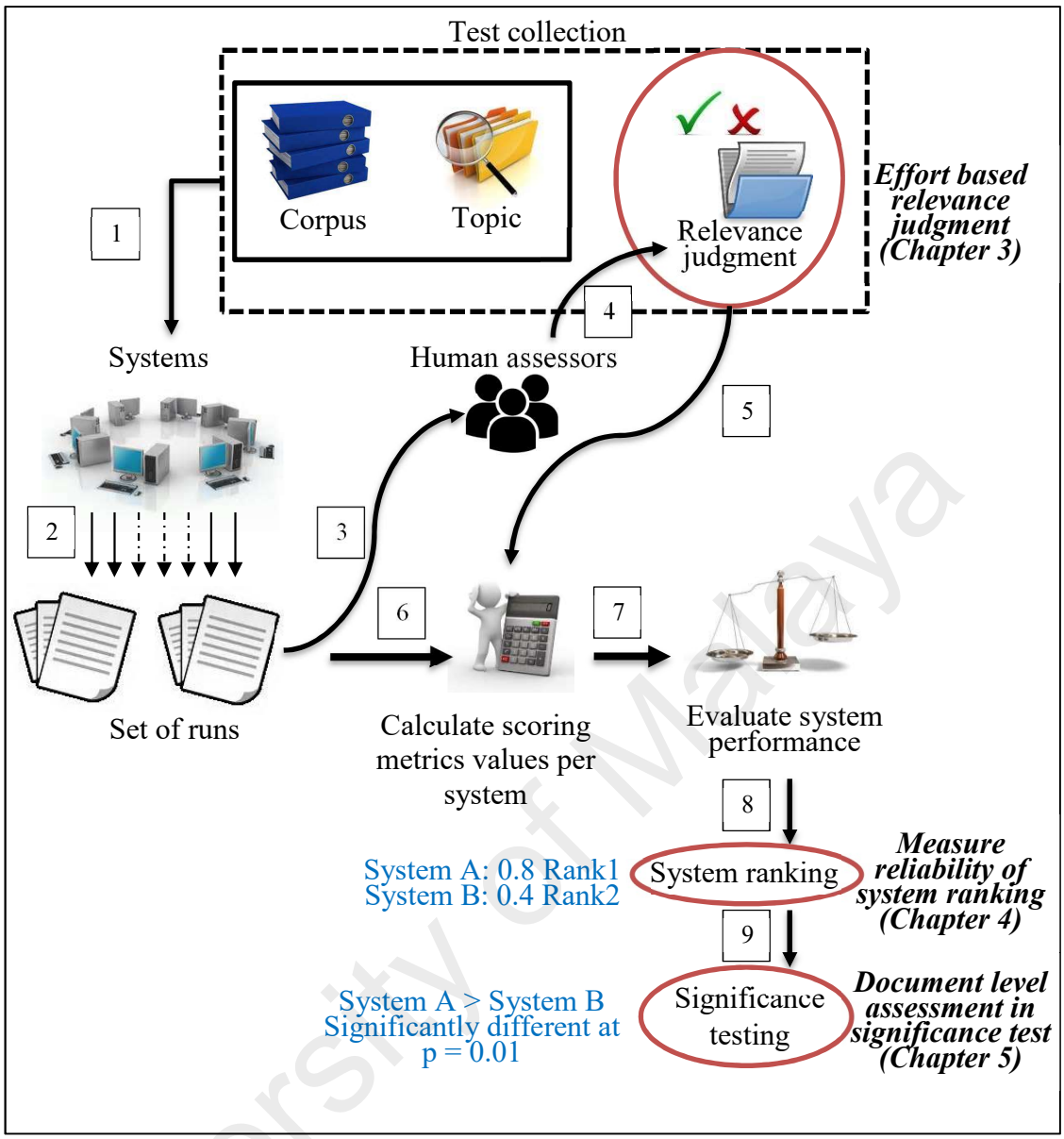

Figure 1.1: System-oriented information retrieval evaluation

The process of information retrieval involves a system known as a retrieval system or search engine with its underlying algorithm for indexing and query matching. The retrieval system could then retrieve the information and present the list to the user. The list of information or documents should be those related and relevant to the user's query. However, the relevance of these documents to the query is unknown without the generation of relevance judgment. Therefore, a relevance judgment is necessary for

information retrieval system evaluation to determine the performance or the goodness of the retrieval system in producing results that are suitable for the user's query.

Then, the performance of the retrieval systems can be scored using effectiveness metrics. At this point, the retrieval systems can be evaluated by ranking them in relation to other retrieval systems to determine the better or worse performing systems. To further confirm the differences in the effectiveness scores between two systems, a statistical significance test can be used for evaluation.

With the increasing amount of information on the Web, it has become the interest of the information retrieval research community to study the reliability, validity, and significance of the system-oriented evaluation. This thesis focuses on the system-oriented evaluation context, largely involving the test collection methodology and data from the TREC.

This thesis tackles various aspects of system-oriented evaluation that involves the relevance judgment, system rankings and significance test. The overall aim of the thesis is to increase the effectiveness of the system-oriented evaluation through disconnected experimentations involving relevance judgment, system rankings and significance test. Each experimentation consists of distinct objectives that contribute to the overall objective of this thesis to increase the effectiveness of system-oriented evaluation.

Firstly, the system-oriented evaluation involving the relevance judgment identifies the lack of effort-based relevance judgment without the involvement of human judges in evaluating the information retrieval systems. Next, the system-oriented evaluation lacks in measuring the reliability of individual information retrieval system rankings. The last aspect attempts to overcome the inaccuracy of statistical significance test results due to loss of scores from averaging and cut-offs in information retrieval system evaluation.

Following are the respective research questions and objectives for each of the gap identified in the system-oriented evaluation aspect covered in this thesis.

1.1 Research Questions

Chapter 3

RQ3.1 Does low effort relevance judgment cause variation in the system rankings when evaluated at a deeper depth of evaluation?

RQ3.2 Is there variation in system rankings for a specific group of retrieval systems when using low effort relevance judgment for evaluation?

RQ3.3 Can the retrieval systems using low effort relevance judgments be evaluated effectively with reduced topic size?

Chapter 4

RQ4.1 How to measure the reliability of individual retrieval system rankings and identify suitable combination of metrics in measuring reliability of individual system rankings?

RQ4.2 Can reliability measurement of individual retrieval system rankings be generalized to other similar metrics?

RQ4.3 Which retrieval systems from the original ranks have reliable system rankings?

RQ4.4 Does the reliability measurement of individual retrieval system rankings represent the original system rankings?

Chapter 5

RQ5.1 What is an alternate approach to measure statistically significant system pairs to overcome the drawbacks of using averaged or cut-off topic scores?

RQ5.2 What is the suitable sample size for statistical testing using the alternate approach to achieve reliable results?

RQ5.3 Can the alternate approach effectively measure statistically significant system pairs?

1.2 Research Objectives

Chapter 3 aims

OBJ3.1 To propose a systematic way of generating relevance judgments that incorporate effort and determine the variation in system rankings due to low effort relevance judgment in evaluating retrieval systems at different depth of evaluation.

OBJ3.2 To measure the variation in system rankings due to low effort relevance judgment on groups of systems in information retrieval system evaluation.

OBJ3.3 To explore the effectiveness in evaluating retrieval systems using low effort relevance judgment with reduced topic sizes.

Chapter 4 aims

OBJ4.1 To propose a method to evaluate the reliability of individual retrieval systems and determine suitable combination of metrics for measuring reliability of individual system rankings.

OBJ4.2 To extend the generalization of the system ranking reliability to other similar metrics pairs.

OBJ4.3 To distinguish the original systems with reliable system rankings.

OBJ4.4 To validate the reliability measurement of individual retrieval system rankings with the original system rankings.

Chapter 5 aims

OBJ5.1 To propose an approach suitable for evaluating retrieval systems by overcoming the drawbacks of inaccuracy using averaged or cut-off topic scores in statistical significance test.

OBJ5.2 To identify a suitable sample size in pairwise retrieval systems evaluation using the proposed approach that produces reliable results.

OBJ5.3 To validate the effectiveness of the proposed method in measuring statistically significant system pairs.

Each of the experimentation contributes to the overall objective of increasing the effectiveness of information retrieval systems evaluation. The first experiment contributes by evaluating the IR systems using effort-based relevance judgments that measures the effectiveness of the IR systems by taking real users into consideration. The second experiment contributes to the effectiveness of the IR systems evaluation by measuring the reliability of individual system rankings. The third experiment uses the document-level approach in increasing the effectiveness of IR systems evaluation. The following paragraphs detail the overall thesis structure.

1.3 Thesis Structure

Chapter 2 gives an overview of information retrieval, the different categories of information retrieval evaluation, and elements of system-oriented evaluation such as document corpus, topics, pooling, relevance judgments, and effectiveness metrics. Some of the effectiveness metrics which are used in the experimentations of this thesis are described. Chapter 2 also covers the discussion on statistical significance tests in information retrieval evaluation and analysis of variance (ANOVA). The chapter also includes details on the test collection initiatives from Text REtrieval Conference (TREC) and other similar test collections tackling different languages.

Chapters 3, 4 and 5 are the main contributions of the thesis consisting of separate experimentations tackling different aspects of the information retrieval evaluation (as noted in Figure 1.1). Chapter 3 tackles the issue of evaluating information retrieval systems with only relevance as part of the relevance judgment. The experimentation explores ways to incorporate effort in addition to relevance in the relevance judgments. The effort is known to play a major role in the satisfaction of users, and thus a systematic approach to incorporate effort in the relevance judgment is undertaken to evaluate the retrieval systems. The chapter reveals that low effort relevance judgments causes variation in the retrieval system rankings at the various depth of evaluation and may impact the way conclusions are drawn. Also, evaluation of groups of systems with low effort relevance judgments shows surprising trends between top and bottom ranked systems. Chapter 3 also attempts to evaluate the retrieval systems using low effort relevance judgments with reduced topic sizes but only few effort features produce sufficiently good results.

Chapter 4 embarks on the reliability measurement of individual retrieval system rankings using relative ranks from each topic. The reliability of the individual retrieval

systems is measured using ANOVA type of analysis known as the intraclass correlation coefficient (ICC). The ICC measures the agreement between two sets of data which results in a reliability correlation coefficient indicating the reliability of individual retrieval system. The approach highlights that certain combination of metrics is more suitable for measuring and generalizing the reliability of individual system rankings' result. There exists variation in results with the use of different combination of metrics. Through the evaluation, the retrieval systems which are highly reliable are identified. An experimental attempt with reduced topic sizes in measuring the reliability of individual retrieval systems suggests the results may skew for different topic sizes. Therefore, evaluation with reduced topic sizes should be implemented with care.

Chapter 5 tackles the inaccuracy in evaluation of retrieval systems in a pairwise manner using statistical significance test. The approach uses document-level scores to overcome the inaccuracy induced by averaged or cut-off topic scores in statistical tests. The use of indivisible document-level scores results in higher numbers of statistically significant system pairs compared to the existing method. In fact, the proposed approach shows a high percentage of agreement with the current method besides identifying more statistically significant system pairs. The document-level approach is more effective than the existing approach when compared with equal sample sizes in statistical tests. However, suitable sample sizes of document scores in the significant tests are necessary for achieving reliable results in identifying statistically significant system pairs.

Chapter 6 concludes the thesis by highlighting the thesis contribution for the three experimentations; evaluating retrieval systems using effort-based relevance judgments, measuring the reliability of individual retrieval system rankings, and document level assessment using document level scores. It also includes the future works arising from the mentioned experimentations.

CHAPTER 2: LITERATURE REVIEW

This chapter consists of the general literature review that is necessary for this study. The chapter discusses the information retrieval, and two different information retrieval evaluation, namely, the system-oriented and user-oriented approaches. Next, the chapter describes common effectiveness metrics such as precision, recall, average precision, and rank-biased precision. Subsequently, the chapter details the statistical significance test and ANOVA. The chapter ends with information on test collections from Text REtrieval Conference (TREC) and other test collection initiatives.

2.1 Information Retrieval

The information retrieval (IR) process consists of the representation of the document contents, the information need of the user and the comparison of these both (Hiemstra & Graham, 2009). The information retrieval initially aimed at methods to handle data in reference databases using title-only or abstract-only documents (Karlgren, 2000). However, the immense growth of data is constantly shifting the aims of information retrieval. The retrieval process started with indexing during the Cranfield paradigm. The indexing itself was manual (Karlgren, 2000) to enable retrieval of documents that match the specified index. Carefully constructed instructions and a set of allowed index terms reduce variation but affect the flexibility of index search. However, the evolution of indexing has led to the use of document contents as part of information retrieval. Even the manual indexing has shifted to automatic indexing. Karlgren (2000) defines indexing as

"Indexing is the practice of establishing correspondences between a set, possibly large and typically finite, of index terms or search terms and individual documents or sections thereof."

Indexing belongs within an information retrieval process as shown in Figure 2.1. The indexing prepares a document for query matching. Once indexed, the document is then ready for the retrieval process. Figure 2.1 also shows information need, which is formulated into a query. Based on the query, the indexed documents are matched and retrieved. The retrieved documents are then presented to the user for the specific query. Once the user completes the information need, his or her need may change after understanding and consuming the retrieved document. The user could then change the information need or reformulate the query. The retrieval process is then repeated to obtain a list of retrieved documents.

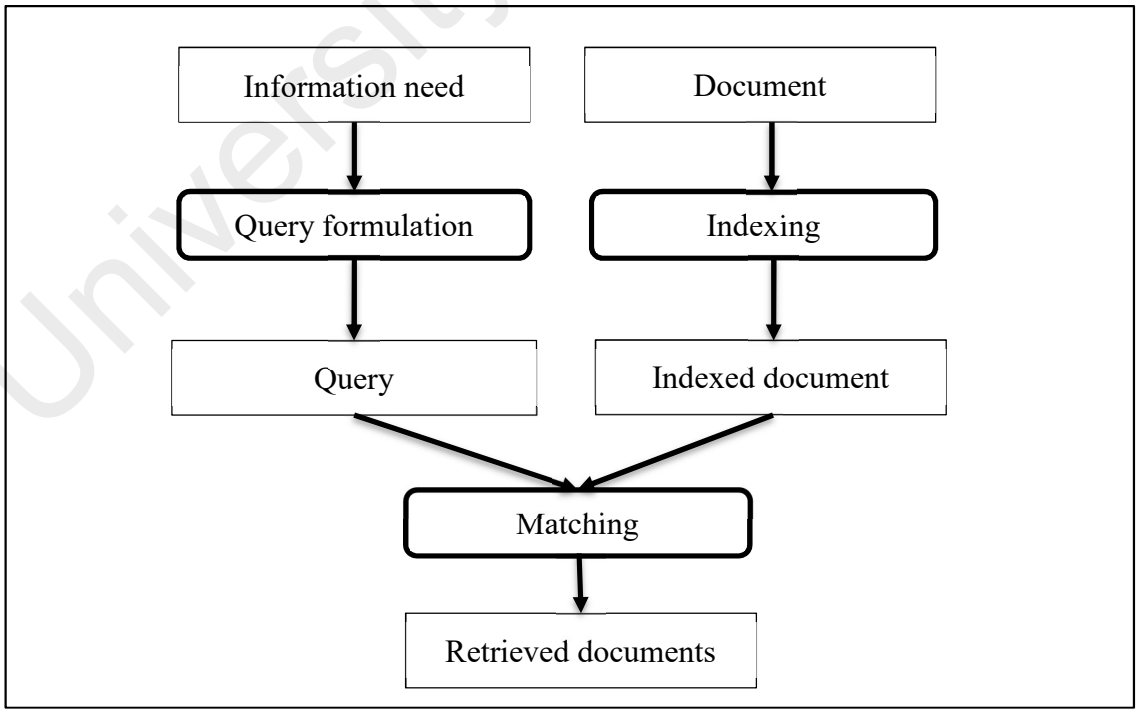

Figure 2.1: Information retrieval process (adapted from (Hiemstra & Graham, 2009))

The information retrieval process is itself an important aspect such that studies focus on indexing (De Melo & Hose, 2013; Golub et al., 2016; Hiemstra & Graham, 2009; Varathan, Sembok, Kadir, & Omar, 2014) and query formulation (Bailey, Moffat, Scholer, & Thomas, 2015; Golub et al., 2016; Hiemstra & Graham, 2009). The process of information retrieval involves a system known as a retrieval system or search engine. Each system could have its underlying algorithm for indexing and query matching, before presenting the user with its retrieved list of documents.

With the information retrieval in place, it is unclear if these retrieved lists of documents are relevant or irrelevant to the user's information need. To determine the result from the retrieval system, a mechanism is required to measure its performance based on the retrieved documents. Information retrieval evaluation is necessary to assess the performance of the systems.

2.2 Information Retrieval Evaluation

The information retrieval evaluation evolved through the Cranfield paradigm where experiments are performed on test collections. The evaluation is necessary to measure and quantify the effectiveness, and assess user satisfaction and acceptance of the information retrieval systems. It is crucial in designing, developing, maintaining effective information retrieval systems (Clough & Sanderson, 2013) and controlling the effects of different parameters (Voorhees, 2002).

The information retrieval evaluation divides into two broad categories: system-oriented or test collection evaluation, and user-oriented evaluation. The system-oriented evaluation usually takes place in a laboratory setting with minimal end user involvement (Clough & Sanderson, 2013). It involves a shorter experimentation time and is more cost-

effective compared with other information retrieval evaluation techniques. The system-oriented evaluation uses a collection of documents, topics or queries or the information need, and relevance judgments. In another term, system-oriented evaluation is also known as batch retrieval evaluation (Turpin & Hersh, 2001). The experimental based IR evaluation can be defined as a standard for repeated-measure where the topics are defined as experimental units, the systems as treatments with each of the systems providing a ranked system run for each topic or query (Carterette, Kanoulas, Pavlu, & Fang, 2010).

Short experimental turnaround and time span in experiments, low cost, quantifiable, and repeatable are some of the benefits of system-oriented evaluation (Moffat, Scholer, & Thomas, 2012). Although in longer duration, the cost of the system-oriented evaluation is low, its initial cost could be significant. Another drawback of system-oriented evaluation is the correlation of metrics with "user experience". It means the metric score differences does not necessarily measure the difference in the user's capacity to perform the task (Moffat et al., 2012).

On the other hand, the user-oriented evaluation focuses on the interaction of the user with the information retrieval systems, and the user's context and situation (Borlund, 2009). It also takes into account on the usability of the interface, user's behavior towards the search process, and user satisfaction. However, the users' satisfaction requires good performance concerning the information need, system speed and the user interface (Mandl, 2008). User-oriented studies can be complex, costly, time-consuming, uncontrollable and capture only part of the user experience (Smucker & Clarke, 2012).

This work focuses on system-oriented evaluation through laboratory experimentations using Text REtrieval Conference (TREC) test collections. The TREC supports research in information retrieval for large-scale evaluation of text retrieval methodologies. A test collection usually consists of a document corpus, topics and relevance judgments. The

retrieval of information is then evaluated using effectiveness metrics. The following subsections detail each aspect of the system-oriented evaluation using data from TREC as shown in Figure 2.2. The following aspects are part of the TREC evaluation cycle.

i. Document corpus and topic
ii. Pooling
iii. Relevance judgments
iv. Effectiveness metrics
v. Evaluation of system performance

Briefly, the retrieval systems use the document corpus to retrieve documents based on the topics. The retrieved lists of documents are known as system runs. Some of these system runs are pooled to create a subset of documents. Human judges who are experts in the topic will evaluate the subset documents against the topic to create a relevance judgment. In Figure 2.2, the relevance judgment is stated as partial relevance judgment because not all system runs and documents were considered for judgment in TREC. With the partial relevance judgment, a score for the system runs can be computed using effectiveness metrics. Finally, the systems' performance can be evaluated.

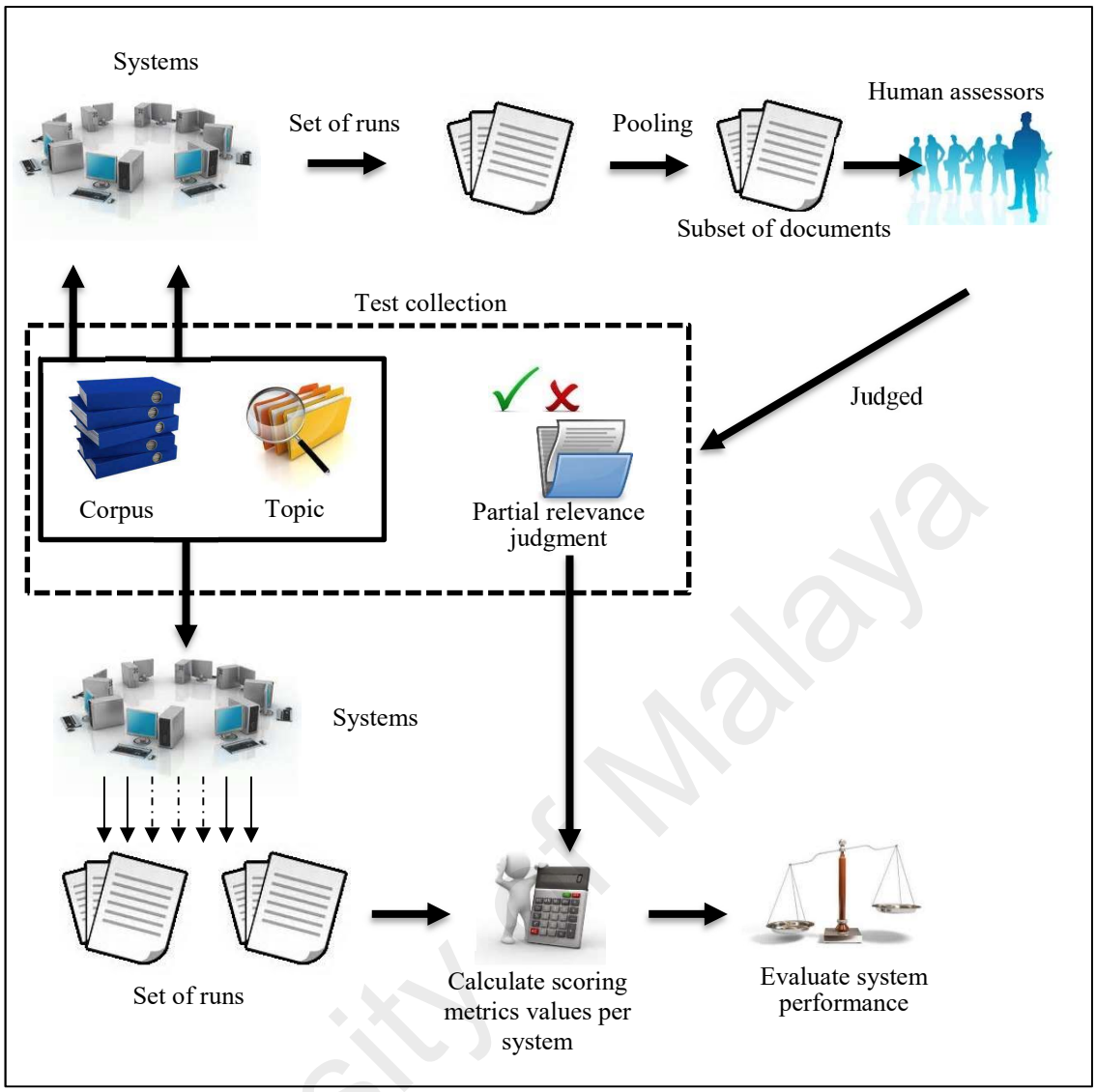

Figure 2.2: TREC evaluation cycle

2.2.1 Document Corpus and Topics

The document corpus is a collection of documents. These documents could be collected from newswire, newspaper articles or web crawls. The topic or query represents a user's information need. The information retrieval system aims to satisfy and fulfill the user's information need through its retrieval mechanisms.

Document corpus from TREC provided a large amount of text up to 27.3 TeraByte uncompressed with approximately 733 million documents in the year 2014 (Collins-

Thompson, Macdonald, Bennett, Diaz, & Voorhees, 2014). Initially, TREC document corpus consists of newswire or newspaper articles (Voorhees & Harman, 2000). However, the document corpus needs to adapt to the evolving Web by including documents crawled from the Web.

The TREC topics provide clear information need for each query. Retrieval systems would retrieve documents using the topic, either the title-only or the other parts of the topic from the document corpus to meet the information need of the user. Experts of topics usually create topics in TREC. These same experts would then judge the documents to create the relevance judgment.

In TREC-8, the topics used short titles to facilitate experiments with short queries (Voorhees & Harman, 1999). However, in TREC-9 the topics were changed to suit the Web retrieval systems (Voorhees & Harman, 2000). These topics utilized log queries submitted to Excite[1] search engine. The process of topic creation involves selection of query from the collection of sample queries. Following that, the assessors need to develop the description and narrative related to the original interpretation. The assessors would then search the web document collection to estimate the number of relevant documents for each topic. Figure 2.3 illustrates the process flow for creating topics as detailed in this paragraph.

[1] Log queries were obtained from Jack Xu of Excite's ftp site at ftp.excite.com/pub/jack on December 20th 1999.

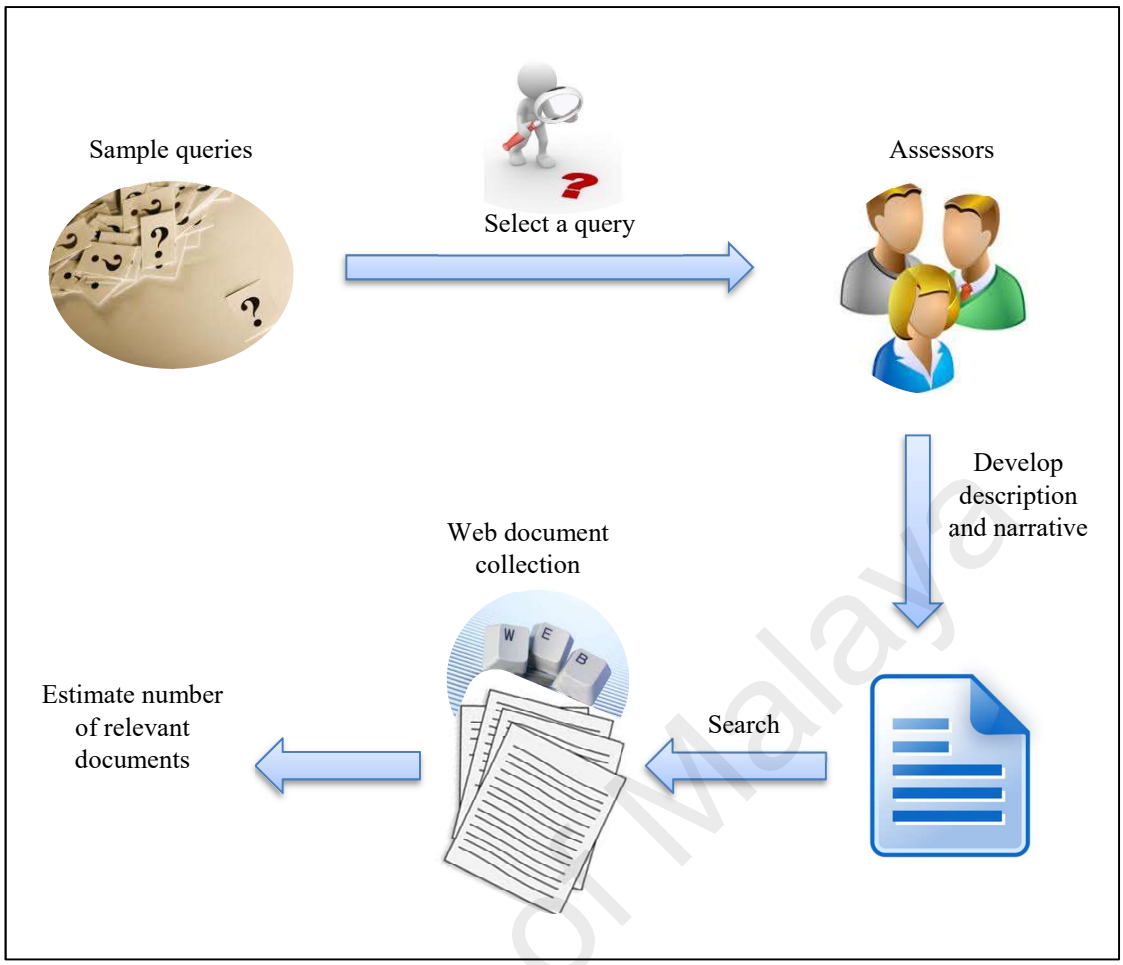

Figure 2.3: Process flow of creating topics for Web track TREC-9

The topics from TREC usually comprises of four main components:

i. Topic id or number

ii. Title

iii. Description

iv. Narrative

The topic id is a unique identifier to represent the topic. The title is usually the query submitted to a retrieval system. The description is typically not more than one sentence while the narrative provides a complete description of the document that makes it relevant

(Voorhees & Harman, 2000). Figure 2.4 shows a sample topic from TREC-9 Web track containing all the four main components of a topic.

<top>

<num> Number: 451

<title> What is a Bengals cat?

<desc> Description:

Provide information on the Bengal cat breed.

<narr> Narrative:

Item should include any information on the Bengal cat breed, including description, origin, characteristics, breeding program, names of breeders and catteries carrying bengals.

References which discuss bengal clubs only are not relevant. Discussions of bengal tigers are not relevant.

</top>

Figure 2.4: Example of topic from TREC-9 Web track

With the availability of document corpus and topics, the retrieval systems will be able to retrieve relevant documents. In TREC, researchers produce their queries using automated or manual methods with the provided document corpus and topics. A retrieval system executing a task on a test collection results in a run. Automatic runs do not involve manual intervention from the time of topics submission to the retrieval system, and the produced system runs. However, manual runs allow any amount of human intervention in producing the retrieval result. The researchers could utilize the title and other components of the topic in the retrieval process. The researchers are also encouraged to use only the title in retrieving relevant documents (Voorhees & Harman, 2005), and

submit up to 1,000 documents for each topic to NIST (National Institute of Standards and Technololgy) . The ranked retrieved documents should be from most relevant to least relevant in the system run.

While the document corpus and topics are two essential components of a test collection, a third important element is the relevance judgment. Large-scale test collections usually contain large numbers of documents and topics that need judgment. Pooling was introduced to overcome the workload of judging huge numbers of documents.

2.2.2 Pooling

Pooling became common due to large-scale test collection that makes it impractical to assess all documents. Back in the 1960s, 1970s and early 1980s, there was no pooling because the texts involved did not exceed more than 3MB (Sanderson & Joho, 2004). Pooling was first proposed to the British Library to build a huge test collection in 1975 (Jones & Rijsbergen, 1975). As test collections started to include a large number of documents, pooling helped to reduce the number of documents that need assessment by the human assessors.

Pooling selects a subset of system runs and documents from the submitted runs for judgments. Figure 2.5 shows the pooling technique used by TREC. With the pooling technique, the selected system runs are known as contributing systems. From these contributing systems, top k documents per topic are chosen, where k is the pool depth. Duplicates of the pooled documents are removed before being presented to human assessors for judgments.

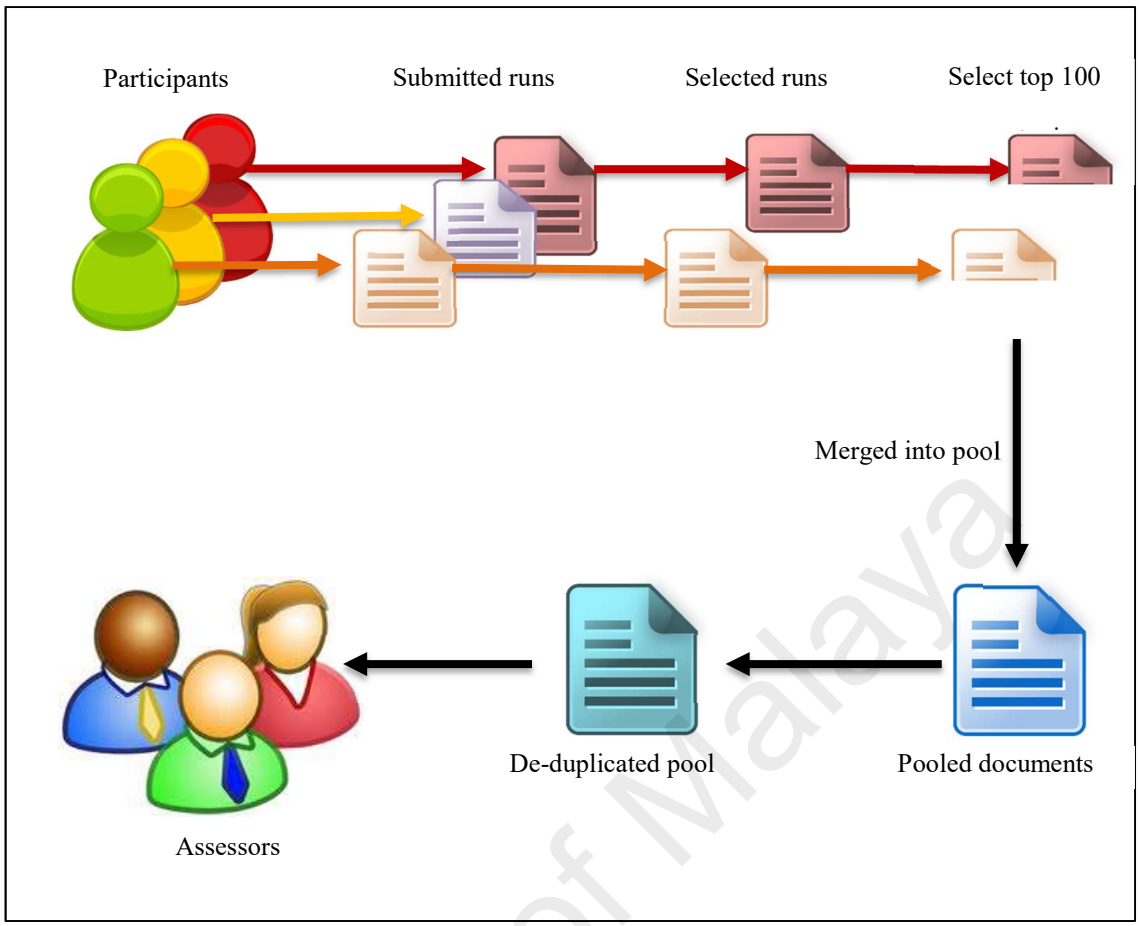

Figure 2.5: Pooling technique used in TREC

Pooling could reduce the number of documents that need judgment. Assuming there are 50 topics and each system run has 1,000 documents per topic while the number of system runs is 10. The approximate total documents (before removing duplicates) are 500,000. When pooled with top 100 documents for each topic and each system run, the complete documents to be judged are only 50,000 (before removing duplicates). There could be a further reduction when only a subset of the system runs selected with top 100 documents per topic. These reduced numbers require less effort than unpooled documents from the human assessors. The amount of time taken for judgment also can be reduced with pooling.

Assuming a single document requires 30 seconds for judgment, and there are 1 million documents. To complete judgment for one topic approximately requires 8,333 hours. For at least 50 topics, it will nearly take five decades to complete. To reduce the length of time taken in forming the relevance judgments, more topic expert judges will have to work on the document judging. Due to this, it is impractical to have all documents judged; hence, TREC uses the pooling technique to form the relevance judgment.

Pooling is cost-effective, but it is not very clear if it would produce reusable test collections because there are fewer judgments than unpooled relevance judgments. It is likely future systems will retrieve many documents that are unjudged by previous relevance judgment (Carterette, Kanoulas, & Yilmaz, 2010). Due to this, the performance of the future systems could be measured inaccurately. However, study shows pooling does not produce a significant impact on the relevance judgment when evaluating system performances (Zobel, 1998). Even though pooling leaves out some of the system runs, the results are reliable with the use of sufficient pool depth of 100 (Zobel, 1998). He also stated pooling identifies at least 50% to 70% of the relevant documents. Without pooling, there would be too many documents for judgment, consume time and impractical.

2.2.3 Relevance Judgment

The documents from pooling are presented to human assessors or judges. The human assessors would then read each of the document for the specified topic and identify its relevance. The relevance judgment consists of relevancy information of the documents for the specified topics. Due to pooling, the relevance judgment itself is known as partial relevance judgment.

There are two ways of classifying relevancy; first is binary relevance judgments' using the indication of zero for irrelevant and one for relevant documents and the second is graded or ternary relevance. The ternary or graded relevance could vary from study to study. Ternary grades were used in TREC's Web track (Craswell & Hawking, 2003), Enterprise track and Blog track (Sanderson, 2010). Although graded relevance is available, binary relevancy is more common in ad hoc style test collections (Sanderson, 2010). Figure 2.6 shows a snippet of binary relevance judgment from the TREC-8 ad-hoc track containing the topic number, document identifier, and the relevancy.

Topic	Q0	Document Id	Relevancy
401	0	FBIS3-18916	0
401	0	FBIS3-18926	0
401	0	FBIS3-18943	1
401	0	FBIS3-18946	0
401	0	FBIS3-18972	0
401	0	FBIS3-18997	0
401	0	FBIS3-19003	0
401	0	FBIS3-19032	1
401	0	FBIS3-19037	0
401	0	FBIS3-19038	1

Figure 2.6: Snippet of binary relevance judgment from TREC-8 ad-hoc track

The relevance judgment created through pooling technique consists documents that are labeled as relevant or irrelevant following the judgment by human assessors. However, unpooled and unjudged documents require classification too. These documents are usually assumed irrelevant because they were not retrieved by any of the systems contributing to the pooling. Hence, the probability of it being relevant is assumed to be very low. Nevertheless, the authors (Stefan, Clarke, Yeung, & Soboroff, 2007) have

considered this approach of assuming the unjudged documents to be irrelevant as appropriate.

Once the relevance judgments are created, it can then be used to evaluate the performance of all retrieval systems in the test collection. The relevance judgment completes a test collection when matches are discovered between topics and documents making it possible for system evaluation. The relevance judgment is beneficial when the same test collections are used in future by avoiding repeated judgments. A good alternative to remunerate the cost of large test collections is reusability of the test collections (Carterette, Kanoulas, & Yilmaz, 2010).

2.2.4 Effectiveness Metrics

The test collection makes it possible to score the system runs using effectiveness metrics. There are various effectiveness metrics available for retrieval system evaluation. The measurement of information retrieval effectiveness is the ability of the information retrieval system to identify relevant and non-relevant document for a specific query by the user. Throughout the experimentations of this study, three metrics; precision, average precision and Rank-Biased Precision (RBP) will be utilized. Hence, the following subsections focus mainly on these metrics.

2.2.4.1 Precision and Recall

Precision and recall are two basic IR metrics where precision measures the fraction of documents that are relevant to the query among all of the returned documents whereas recall is the ratio of relevant items retrieved to all relevant items in the file (Saracevic, 1995). They are single-valued metrics based on the retrieved list of documents by the

system. For systems that rank retrieved documents, it is vital to consider the order of the presented documents. Equation 2.1 indicates precision and Equation 2.2 indicates recall.

Equation 2.1

$$Precision = \frac{number\ of\ retrieved\ relevant}{total\ retrieved}$$

Equation 2.2

$$Recall = \frac{number\ of\ retrieved\ relevant}{total\ relevant}$$

Due to pooling, it is possible to obtain the number of retrieved documents for a query compared to knowing all relevant documents used in the recall. Therefore, the usage of precision in measuring effectiveness is achievable compared to recall. A system having returned relevant documents earlier in a retrieval process will have a better performance compared to a system retrieving relevant documents later or at lower rankings.

The effectiveness of a retrieval system measuring precision at a certain cut-off rank *k* is P@*k*. It does not require the total numbers of relevant documents, but it is the least stable evaluation measure compared to other commonly used measures (Manning, Raghavan, & Schütze, 2009; Webber, Moffat, & Zobel, 2010). Besides, P@*k* is not a good average metric due to strong influence of relevant documents in a query. The P@*k* is defined in Equation 2.3 where, *k* represents the top *k* documents while r_i represents the relevancy of a document at rank *i*.

Equation 2.3

$$P@k = \frac{1}{k}\sum_{i=1}^{k} r_i$$

2.2.4.2 Average Precision

The average precision (AP) is computed by averaging the precision of each document per topic. The AP is top-weighted because a relevant document in position 1 contributes more to the effectiveness score than one at position 2 and so on down the ranking. The AP is preferred due to its good probabilistic interpretation (Yilmaz & Aslam, 2006), justified by acceptable user model (Robertson, 2008) and appears to be highly informative (Aslam, Yilmaz, & Pavlu, 2005).

When computing the AP, if there are more relevant documents than the considered pooling depth, the documents' ranking should be extended until the remaining relevant documents are found. In the case where those documents are not found, such as in TREC style pooling, it is assumed that those unfound relevant documents have a precision of 0 (Moffat & Zobel, 2008). Similar to precision, the AP is measurable at a specific cut-off rank, *k*. The AP@*k* is defined in Equation 2.4 where *k* represents the rank of the retrieved document, *P(k)* is the precision at cut-off *k*, *rel(k)* is the relevancy indicator of the document at rank *k* and *R* is the number of total relevant documents.

Equation 2.4

$$AP@k = \frac{\sum_{k=1}^{n}\big(P(k)x\ rel(k)\big)}{R}$$

Table 2.1 shows an example of the retrieved documents, its relevancy, precision calculation and the AP calculation at each rank. In the example (based on relevance judgment of TREC-8), the selected query has 17 relevant documents. Calculating the AP at depth 10 (AP@10), results in 7.04 / 17 = 0.41.

Table 2.1: Example of average precision calculation for a query

Rank	Document Id	Relevancy	Total relevant document at specific rank	Precision	AP
1	LA050889-0063	1	1/1	1	1
2	FT941-15416	1	2/2	1	1
3	FT942-10464	1	3/3	1	1
4	FT921-6003	1	4/4	1	1
5	FBIS4-21346	0	4/5	0.8	0
6	FBIS3-4051	0	4/6	0.67	0
7	FT941-15415	1	5/7	0.71	0.71
8	FT942-17238	1	6/8	0.75	0.75
9	LA022689-0070	1	7/9	0.78	0.78
10	FT933-7330	1	8/10	0.8	0.8

The AP metric is sometimes denoted by a single value to indicate the quality of a retrieval system using mean average precision (MAP) (Mandl, 2008; Voorhees, 2007). The MAP of a system with multiple topics is obtained by calculating the mean of the average precision from each of the topics from the system run (Voorhees, 2007). Mandl (2008) stated that MAP is calculated based on the average precision value at each level of recall. Equation 2.5 indicates the MAP formula where *T* represents total topics in a system run, and *AP(t)* is the average precision for each of the topics in the system.

Equation 2.5

$$Mean\ Average\ Precision = \frac{\sum_{t=1}^{T} AP(t)}{T}$$

TREC uses MAP for its evaluation process, and it has become an important result by comparing the properties of MAP (Sanderson, 2010). As stated by Ravana and Moffat (2009), the MAP has become dominant and including MAP scores in IR experiments is unavoidable because most research work in IR evaluation reports MAP scores (Ravana

& Moffat, 2009). It is also one of the most commonly used measures to describe retrieval results in TREC (Voorhees, 2007). MAP tends to favor systems that retrieve relevant documents earlier and is precision-biased.

Table 2.2 shows an example of average precision for five topics of a system and the corresponding MAP calculation using Equation 2.5, which results in MAP score of 0.5. Though some topics have poorer or better AP scores compared to others, a system is represented by its ability to perform well over multiple topics. Therefore, the MAP score offers a single-valued representation of the system performance over multiple topics.

Table 2.2: Example of average precision per topic for a system

Topic	Average Precision
401	0.65
402	0.31
403	0.88
404	0.23
405	0.42

$$MAP = \frac{0.65 + 0.31 + 0.88 + 0.23 + 0.42}{5} = 0.5$$

2.2.4.3 Rank-Biased Precision

The rank-biased precision (RBP) is a rank sensitive metric using parameter p as a measure of user persistence, the probability that a user, having reached any given point in the ranked document list returned by a system, will proceed to the next rank (Moffat & Zobel, 2008). The parameter p in Equation 2.6 represents this persistence of the user. When p = 0.0, the user is assumed as either satisfied or not satisfied with the top-ranked document and will not look further down the list of the retrieved documents. On the other

hand, as *p* approaches 1.0, it is assumed that the user would look through many documents before ending their search task. The model assumes the transition probability is independent of the document relevance at *r* (Sakai & Kando, 2008). The RBP scores are always within the range of 0 and 1 (Moffat & Zobel, 2008). RBP is defined by Equation 2.6 where r_i represents the known relevance of the retrieved document at rank *i* and *p* is the persistence.

Equation 2.6

$$RBP@k = \sum_{i=1}^{k} r_i w_i$$

where,

$$w_i = (1-p)\, p^{i-1}$$

Table 2.3 shows an example of the RBP calculation for two different persistence values for each document. The example is of a topic in TREC-8. The RBP@10(p=0.8) is 0.027 and RBP@10(p=0.95) is 0.03 for the example in Table 2.3. Larger values of *p* are known to lead to deeper evaluation (Webber et al., 2010).

The RBP is also a top-weighted metrics like the AP. The scores down the ranked documents are not affected by the relevant documents as proved by (Sakai & Kando, 2008) using ideal ranked documents. Only when the number of ranked documents is 1,000, an ideal RBP score of 1 is achievable for the three different persistence. The scenario explains the unnormalized metric and the weights of RBP declining consistently with rank. The RBP is less stable than AP but more stable than P@10 (Webber et al., 2010).

Table 2.3: An example of RBP calculation for persistence 0.8 and 0.95

Rank	Document Id	Relevancy	Weight (p=0.8)	Weight (p=0.95)
1	LA050889-0063	1	0.2	0.05
2	FT941-15416	1	0.16	0.0475
3	FT942-10464	1	0.128	0.045125
4	FT921-6003	1	0.1024	0.042869
5	FBIS4-21346	0	0	0
6	FBIS3-4051	0	0	0
7	FT941-15415	1	0.052429	0.036755
8	FT942-17238	1	0.041943	0.034917
9	LA022689-0070	1	0.033554	0.033171
10	FT933-7330	1	0.026844	0.031512

2.2.4.4 Effectiveness Metrics in TREC Web Track and Robust Track

TREC uses certain effectiveness metrics to evaluate the systems. These metrics vary from year to year according to the task for Web track. In TREC-7, metrics such as P@20 and modified average precision were used (Hawking, Craswell, & Thistlewaite, 1999) while TREC-8 used mostly precision, recall and summary measures derived from precision and recall (Voorhees & Harman, 1999). In TREC-9, measures such as AP, P@10 and DCG[100] were used in evaluating the systems (Voorhees & Harman, 2000). Then, in TREC-2001 and TREC-2002 Web track, another measure known as mean reciprocal rank (MRR) was used in addition to average precision (Hawking & Craswell, 2002) and P@10 (Craswell & Hawking, 2003). The metrics remained same for TREC-2003 Web track with the use of MAP, P@10, R-Precision, MRR and additional of Success@10. The Success@10 is the proportion of queries where the correct answer appears within the top 10 (Craswell, Hawking, Wilkinson, & Wu, 2004). TREC-2003 also continued to use some of the measures from previous years but also used Recall@1000 for topic distillation queries (Craswell & Hawking, 2004).

The Web track continued after a few years of break, and in TREC-2009, the ad hoc task's primary evaluation measure was MAP and precision at various cut-off ranks (Clarke, Craswell, & Soboroff, 2009). However, from TREC-2010 to TREC-2013 the primary measure was expected reciprocal rank (ERR). Variants of nDCG were used besides the usual MAP and P@*k* (Clarke, Craswell, Soboroff, & Cormack, 2010; Clarke, Craswell, Soboroff, & Voorhees, 2011; Clarke, Craswell, & Voorhees, 2012; Collins-Thompson, Bennett, Diaz, Clarke, & Voorhees, 2013). While the common metrics remained for few years, two additional metrics were introduced in TREC-2013 for the faceted topics. These metrics are intent-aware expected reciprocal rank (ERR-IA) and α-nDCG (Collins-Thompson et al., 2013). In TREC-2014, the primary measure was ERR-IA for the faceted topics and simply ERR for single facet topic at cut-off rank 20. Other measures reported for TREC-2014 are nDCG@20, α-nDCG@10, and novelty- and rank-biased precision (NRBP) (Collins-Thompson et al., 2014).

In TREC Web track and ad hoc task, MAP measure was used in earlier years while ERR had been the primary measure for the years 2010 onwards. Nonetheless, other metrics have also been used together with MAP and ERR for the specific purpose of measurements. As for the Robust track, a few effectiveness measures were utilized for the year 2004 and 2005. In the year 2004, MAP, an average of P@10, the percentage of topics with no relevance in top 10 retrieved (%no), and the area underneath the MAP(X) versus X curve (area) were used (Voorhees, 2004). In the year 2005, MAP and average of P@10 remained with an additional measure of geometric mean average precision (GMAP) (Voorhees, 2005).

2.3 Statistics in Information Retrieval Evaluation

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data ("Statistics | Definition of Statistics by Merriam-Webster," 2017). Statistics is important to interpret the data and analyze them. Therefore, Sections 2.3.1 and 2.3.2 detail inferential statistics in terms of statistical significance tests and ANOVA respectively.

2.3.1 Statistical Significance Tests

In information retrieval evaluation, various statistical tests could be used to infer the evaluation of the retrieval systems. The statistical significance tests aim to encourage retrieval methods that are truly better instead of those that performed better by chance (Kulinskaya, Morgenthaler, & Staudte, 2014; Smucker, Allan, & Carterette, 2007, 2009). A powerful statistical significance test detects significant improvements although the advances are small while an accurate test only reports significance when it exists (Smucker et al., 2007).

Some of the statistical significance tests are Student's paired t-test, Wilcoxon signed rank test, the Sign test, bootstrap, and Fisher's randomization (permutation) test. Each statistical test has its specific criteria and null hypothesis while the bootstrap and randomization could use any criteria. Both the bootstrap and randomization tests are non-parametric significance tests (Smucker et al., 2007). A previous study suggested that the use of Wilcoxon signed rank test and Sign test for measuring the significance between system means should be discontinued due to their inadequate ability to detect significance and tendency to steer toward false detection of significance (Smucker et al., 2007). Also, randomization, bootstrap, and t-test agree with each other, but the agreement decreases with reduction in topic size. Nonetheless, t-test appears to be suitable even for small topic

sizes (Smucker et al., 2009). Both studies (Smucker et al., 2007, 2009) assumed the permutation test was optimal.

Instead, a large-scale study showed that bootstrap, t-test and Wilcoxon test performs better than permutation test (Urbano, Marrero, & Martín, 2013a). Each of these tests is identified to be optimal for certain criteria whereby bootstrap is optimal in regards to power, t-test for safety, and Wilcoxon test for exactness. The permutation test appears to be not optimal for power, safety or exactness. Besides determining the suitability of these statistical test in practical large-scale study, the study also highlighted that actual error rates are lower than nominal 0.05 (Urbano et al., 2013a).

The most widespread method of determining significance difference is through the *p*-value, by which a specific null hypothesis can be rejected. A typical null hypothesis indicates that two systems are not different, whereas rejection of the null hypothesis in a one-sided significance test could signify that one system is better or worse than the other. The *p*-value is the probability of obtaining almost equivalent or more evidence against the null hypothesis with the assumption that the null hypothesis is true. A *p*-value larger than 0.1 is not small enough to be significant, a *p*-value as small as 0.05 can seldom be disregarded, and a *p*-value less than 0.01 indicates it is highly unlikely to occur by chance (Fisher, 1995). With more data in significance testing, the possibilities of obtaining a significant result are higher. A larger sample size would lower the *p*-value (Cormack & Lynam, 2007).

In Fisher's randomization test, a null hypothesis states that a pair of system is identical on the mean average precision, MAP. It can use any test statistics. A randomization test performed by Smucker et al. (2007), 100,000 random permutations were performed for a pair of system. The difference in the MAP was measured for each pair. The initial MAP differences between the pair of system were computed to determine the statistical

significance of the system pairs. The number of pairs with equal or lower difference than the actual MAP difference of the system pair is counted. Then a two-sided *p*-value is obtained by dividing the total pairs lower or equal to the actual MAP with the number of permutations performed. Due to time constraint in computing all permutations, smaller samples may be suitable but including larger samples will produce better accuracy in estimating *p*-value (Smucker et al., 2007). The null hypothesis is rejected if the *p*-value is lower than the significance level decided.

The Wilcoxon signed rank test has a similar null hypothesis with the randomization test, and the test statistics is distribution free. The Wilcoxon Signed Rank test was preferred as an alternative to randomization test due to less computational power (Smucker et al., 2007). The Wilcoxon test takes the paired score differences and ranks their absolute value in ascending order as their test statistics. The test then replaces the true differences with ranks that approximate the magnitude differences. Although Wilcoxon test does not require much computational power, the loss of information could cause inaccurate conclusions (Smucker et al., 2007).

Test statistics of a Sign test is the number of pairs for which one system is better than the other. The test statistics has a binomial distribution. The total number of pairs is considered the number of trials, and the trials reduce for every tied pair. The Sign test is suitable for reporting number of successes in statistical testing but is sensitive to the minimum difference or the *p*-value chosen (Smucker et al., 2007).

These statistical significance tests may be suitable for different experiments depending on the purpose of the study. The randomization appears to be a good approximate in measuring the significance of systems but requires high computational power which can be overcome with sampling. It is distribution free and classified as non-parametric. The Wilcoxon test requires less computational power compared to randomization test because

it replaces the true difference with ranks. It is also distribution free and suitable for non-parametric study. The Sign test appears to be the least preferred among the Wilcoxon and randomization tests because it retains only the direction of the difference. It is suitable for binomial distribution and is also classified as non-parametric.

Another statistical significance test is the Student's t-test. A Student's t-test measures the difference in mean between two systems. In this hypothesis tests, the relationship between 2 systems' effectiveness is measured. Both system pair's effectiveness is thought to be equal, where the null hypothesis is defined as H_0: A=B. The *p*-value reflects the likelihood of a sample resulting in the observed effect if the null hypothesis is true (Borenstein, 2009). Two-sided *p*-values do not provide directions of deviations from H_0. Under the null hypothesis, the density of a *p*-value from a continuously distributed test statistic is uniform on the interval, 0 to 1.

However, the t-test is a parametric statistical test and closest to that of randomization test (Box, Hunter, & Hunter, 1978). The t-test is tested to be more reliable compared to Wilcoxon and Sign test (Sanderson & Zobel, 2005). It is also a special case of ANOVA if perceived as a linear model (Carterette, 2012). In a pairwise evaluation, the t-test is consistent with ANOVA 99% of the time (Zobel, 1998) and could be used interchangeably depending on the assumptions of data distribution. Nonetheless, the errors due to normality violations are small and cancel out (Carterette, 2012), while also being robust to assumption violations (Sakai, 2014).

2.3.2 ANOVA

The ANOVA or analysis of variance test is used to differentiate the mean scores between two or more samples. It is also possible to analyze two or more independent

variables concurrently to determine the interaction between the independent variables (Chua, 2013). In comparison to the t-test, both the ANOVA and t-test vary in measuring the difference in mean scores. The latter only compares between two mean scores while the former could compare two or more. Carterette (2012) deduced that the t-test is a special case of ANOVA. An ANOVA test that involves one independent variable is known as one-way ANOVA, and two independent variables are two-way ANOVA.

In information retrieval field, the significance measures the difference in mean performance of a pair of system. The result is used to represent a difference in the mean performance of the systems on the population (Zobel, 1998). The variability in MAP scores can be measured by variance for all set of topics and systems. There are three components to the variability, which are system variance, topic variance and system-topic interaction variance (Carterette, Pavlu, Kanoulas, Aslam, & Allan, 2008). In the ideal case, the difference in MAP should be the difference in the system performance rather than the variance from the components.

The ANOVA could be used in replacement to paired tests; the Student's t-test and Wilcoxon's signed rank test (Zobel, 1998). In the experiment comparing 1,830 pairs of systems using a t-test, ANOVA and Wilcoxon's test, the results indicate that t-test and ANOVA are consistent 99% of the time. The Wilcoxon's test is suitable for distributions that are not normal, has better discriminative power than t-test and ANOVA yet the significant results were inconsistent with the t-test and ANOVA (Zobel, 1998). However, the ANOVA and t-test are robust to assumption violations (Sakai, 2014) as the errors due to normality violations are small and cancel out (Carterette, 2012).

As such it appears the ANOVA could be used alternatively to t-test regardless of normality of data distribution. It also has good discriminative power and could be used while comparing two or more mean differences.

2.4 TREC

Text REtrieval Conference (TREC) is the first large-scale information retrieval evaluation initiative sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense. TREC started in 1992 to support research in information retrieval for large-scale evaluation of text retrieval methodologies. The TREC series has four main goals:

i. to encourage research in information retrieval based on large test collections;
ii. to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
iii. to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and
iv. to increase the availability of appropriate evaluation techniques for use by industry and academia, including the development of new evaluation techniques more applicable to current systems.

TREC approach aids research groups by providing access to new test collections annually and a channel to compare their results during the conferences. Due to such approach, TREC became a focus of interest to many researchers in the IR field (Sanderson, 2010). The first test collection, TREC-1 contained documents and topics used in the DARPA TIPSTER project which was distributed as CD-ROMs with about 1 Gigabytes of data each. The test collections consist of Wall Street Journals, AP Newswire, Federal Register and short abstracts from the Department of Energy (Harman, 1992).

The topics were written by actual users to impersonate a real user's need of a retrieval system. Topics were created by doing trial retrievals in a set of documents. If the retrieval for the topic hits between 25 and 100, that particular topic was used (Harman, 1992). Each of the topics was formatted in the same way to contain a beginning and an end marker, topic number, title, a single sentence description, and narrative providing complete description of document relevance for assessors. The relevance judgment is necessary to complete the test collection. During TREC-1, three possible methods were considered to create the relevance judgment. The first method is the most comprehensive with all documents judged for all topics, the second method is random sampling, and the third method is the pooling method that was successfully used in TREC after that (Harman, 1992).

In TREC-4, the concept of tracks was introduced, and each track was designed to focus its evaluation techniques on some specialization and related tasks (Harman, 1995). The introduction of tracks serves few purposes, the first being an incubator for new research areas, second to portray the robustness of the core retrieval technology while making it suitable for a variety of tasks, and third to provide tasks that are attractive to broader groups of researchers (Voorhees, 2002). In 1995, five tracks were run: the multilingual track, the filtering track, the "confusion" track, the database-merging track, and the interactive track. This concept of tracks was successful and became a continuation with different track topics.

In 1997, TREC-6 introduced Very Large Corpus (VLC) track to tackle the challenges imposed by the retrieval tasks required in Web. The laboratory settings of previous test collections were not suitable to Web searching that uses dynamic linking structure compared to static documents. The TREC-7 also had a VLC collection known as VLC2 containing 18.5 million Web pages collected through the Internet Archive. In 1999, Web

track was introduced in TREC-8 to conduct experiments in the context of the Web (Voorhees & Harman, 1999). The focus of TREC-8 was the evaluation of the retrieval methods to suit the growth of documents indexing by commercial search engines.

The ad-hoc track was the main task in previous test collections but was discontinued in TREC-9 to allow more track resources to build on evaluation infrastructure (Voorhees & Harman, 2000). Seven new tracks included are Cross-Language Retrieval, Filtering, Interactive Retrieval, Query Analysis, Question Answering, Spoken Document Retrieval and Web Retrieval. The Web track was developed to simulate the search retrieval on the Web. David Hawking and his colleagues from CSIRO and Australian National University obtained the snapshot of the Web from the Internet Archive from 1997 to create several subset collections (Voorhees & Harman, 2000). The snapshots divide into two subsets; WT10g and WT100g collections. These collections contained heavy-content Web pages and closed set of hyperlinks (Voorhees & Harman, 2001), similar to the Web link structure.

As TREC continues to evolve, the number of participants began to grow with the introductions of multiple tracks. The diverged tasks within each track have also drawn groups of participants to mutual interest as they start to submit their runs to fewer tracks (Voorhees, 2002). Despite the enormous response from researchers, TREC faced lots of criticism due to its tight deadline for submission of data for each cycle and discouraging analysis of individual system results (Rasmussen, 2003). Distributing results outside TREC was one of the problems as it sometimes demotivates publication of the TREC results in the journals. Although TREC received criticism, it has been successful through its conferences and has affected information retrieval research positively. The NIST and TREC continue to provide the necessary infrastructure, technology transfer opportunities

and test collections for research in large-scale information retrieval field to the researchers.

Out of the many tracks available in TREC, this study will utilize two different tracks for experimentation. The Web track and Robust track as detailed in the next two subsections.

2.4.1 Web Track

The TREC Web track test collections analyze retrieval tasks that are unique to the Web. It includes tasks over a huge collection of up to 1 billion Web pages. The TREC-8 Web track was designed as an ad hoc retrieval tasks over large Web task of 100GB and a smaller Web task of 2GB. The ad hoc retrieval task uses topics to search a static set of documents to evaluate the performance of systems (Voorhees & Harman, 1999). In this task, participants produce a set of queries from the topics provided to them by NIST using automatic or manual runs. The automatic run could be divided into title-only runs and using other parts of the topics other than the title (Hawking, 2001). The Web track started in 1999 and continued until 2003. It then took a break for a few years before starting up again in TREC-2009.

The Web track evolved over the years with specific purposes and focusing on particular tasks. The TREC-2001 Web Track focused on topic relevance task and homepage finding task (Hawking & Craswell, 2002). The TREC-2002 additionally aimed to conduct topic distillation experiments, and name page experiments (Hawking, Voorhees, Craswell, & Bailey, 1999). The TREC-2003 Web track consists of the non-interactive stream and interactive stream (Craswell & Hawking, 2004). It continued experiments on topic distillation tasks in addition to navigational task, and these two tasks

belong in the non-interactive stream while the interactive stream focused on human interaction with the topic distillation task.

The broad queries in TREC-2003 are expected to return a list of relevant home pages sites. In the previous TREC-2002, the queries could have multiple correct answers (Craswell & Hawking, 2003). The goal of TREC-2004 Web track was to find ranking approaches that work well for queries without the access to the type of query label (Craswell & Hawking, 2004). Upon the return of Web track in TREC-2009, the experiments are conducted around a new diversity task in addition to traditional ad hoc task (Clarke et al., 2009). The aim of the new diversity task in TREC-2009 was to avoid redundancy in the ranked list while also returning complete coverage for the query. The Web track continues to include a variety of tasks like ad hoc task, diversity task, spam task, and risk-sensitive retrieval task.

As the task and aim progress from year to year, the document corpus also continues to change. Initially, the TREC-8 Web track used the VLC2 collection for Large Web Task and a subset of VLC2 as WT2g collection for a small task. The VLC2 contained 18.5 million web pages while the WT2g contained 250,000 documents (Voorhees & Harman, 1999). The TREC-2002 started using the .GOV test collection consisting 1.25 million page crawls. The .GOV test collection usage was continued until 2003. The mixed query task also used the .GOV test collection in addition to a new test collection W3C specifically created for the Enterprise track (Craswell & Hawking, 2004). In TREC-2009, the Web track started using the ClueWeb09 test collection containing a billion pages (Clarke et al., 2009). However, a smaller subset of the ClueWeb09 was also provided to participants to conduct experimentation if they were unable to utilize the large document collection. The TREC-2013 started using the new enriched one billion web pages

collected from commercial search engines, Twitter, and other sources. The ClueWeb12 is the successor of ClueWeb09 (Collins-Thompson et al., 2013).

2.4.2 Robust Track

The Robust track was run for few years with the goal to improve the consistency of retrieval technology by focusing on poorly performing topics ("TREC 2004 Robust Track Guidelines," 2005). Robust track 2004 consists of 250 topics prepared by NIST and the groupings of topics are shown in Figure 2.7. A total of 150 topics is old topics that have been used in ad-hoc track in TREC-6 till TREC-8. While another 50 topics were those created for TREC-2003 Robust track and new 50 topics were created specifically for the TREC-2004 Robust track. The old topics use binary relevance judgment, and the 2004 Robust track topics use graded relevance judgment. During the evaluation, both relevant and highly relevant documents were assumed as relevant for standardization with binary relevance judgment.

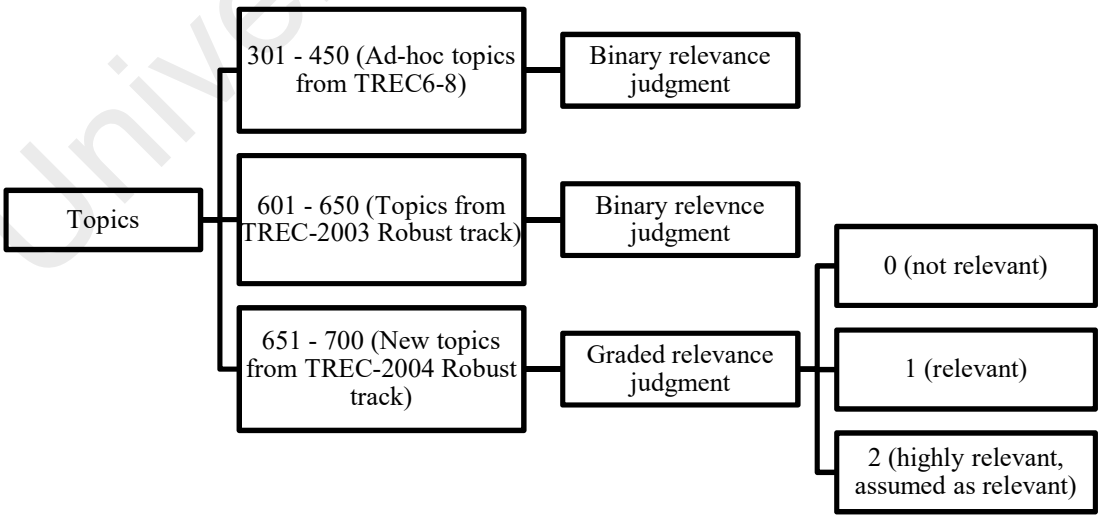

Figure 2.7: Division of topics in TREC-2004 Robust track

Using the topics provided by NIST, each participant may construct their queries to retrieve documents from the test collection. There are two categories of queries that can be created from the topic statements by the participants similar to the Web track, automatic or manual. Automatic queries do not involve any human intervention at any stage of retrieval whereas manual queries are anything else that is not automatic. It is necessary for participants submitting runs from automatic queries to use the standard queries. Automatic queries may use either the title or the description of the topic statements while the manual queries may use any combinations from the topic statements ("TREC 2004 Robust Track Guidelines," 2005).

As for the submission of the runs, each participant is required to submit two parts for each run. The first part consists of a list of retrieved ranked documents for each topic, and they must have at least one document per topic. Each topic can have a maximum of 1,000 documents. The second part consists of the system's prediction of topic difficulties, ranked between 1 till 250 inclusive ("TREC 2004 Robust Track Guidelines," 2005). The prediction of topic difficulties indicates 1 as easiest and 250 as the hardest topic for the system. Such information requested by NIST is to measure the system's ability to recognize difficult topics. The topic difficulties prediction is unique to Robust track and is not part of the Web track.

The Robust track 2005 continues to focus on evaluating retrieval systems by focusing on poorly performing topics (Voorhees, 2005). This track uses the common ad hoc retrieval task, but the evaluation methodology concentrates on a system's least effective topics. The track utilized 50 topics from previous test collection that was classified as difficult topics. Similar to Robust track 2004, it was required to predict the topic difficulties. The aim of such prediction was to steer the systems in topic-specific processing (Voorhees, 2005).

2.5 Other Test Collection Initiatives

The TREC is the first large-scale test collection, but there are other test collection initiatives in information retrieval field. The CLEF[2] Initiative (Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum) is a self-organized body that promotes research, innovation, and development of information access systems. The CLEF mainly focuses on multilingual and multimodal information system testing, tuning and evaluation. It also provides an infrastructure for investigating the use of unstructured, semi-structured, highly structured, and semantically enriched data in information access. It also creates reusable test collections for benchmarking, exploration of new evaluation methodologies and innovative using experimental data and finally providing a foundation for discussion of results, comparing approaches, exchanging ideas and knowledge transfer ("The CLEF Initiative," 2016).

Since 2010, CLEF jointly organized with a set of evaluation labs, an independent event, established by peer-reviewed conference ("The CLEF Initiative," 2016). The results of the various studies were usually presented and discussed at annual workshops in conjunction with the European Conference for Digital Libraries (ECDL), now called Theory and Practice on Digital Libraries (TPDL). The focus is performing a search across European languages but has diversified into other languages including Persian and Indian subcontinent (Braschler & Peters, 2004).

The CLEF focuses on European languages while the NII Test Collection for IR Systems (NTCIR[3]) focuses on Asian languages. The NTCIR is an evaluation exercise held every 18 months in Japan. The NTCIR Workshop is a series of evaluation workshops

[2] http://www.clef-initiative.eu//

[3] http://research.nii.ac.jp/ntcir/index-en.html

intended to enhance research in Information Access (IA) technologies. These include information retrieval, question answering, text summarization, and extraction. It was co-sponsored by Japan Society for Promotion of Science (JSPS) as part of JSPS "Research for Future Program" and National Center for Science Information Systems (NACSIS) since 1997. NTCIR has focused on the cross-language search for Asian languages such as Japanese, Chinese, and Korean (Kando, 2004).

The objectives of NTCIR is to provide a common infrastructure. The forum allows cross-system comparisons and informal environment for exchanging research ideas, cultivate research in IA through re-usable large-scale test collections, and to investigate evaluation methods of IA techniques and methods for constructing large-scale dataset reusable for experiments ("NTCIR project overview," 2016). The first NTCIR Workshop held in 1999 used a test collection called NACSIS Test Collection 1 or NTCIR-1. This test collection consists of 330,000 documents with more than half the documents presented in English-Japanese pairs (Kando et al., 1999).

Some of the other evaluation initiatives include Forum for Information Retrieval Evaluation (FIRE[4]) and The INitiative for the Evaluation of XML Retrieval (INEX[5]). The FIRE is supported by the Information Retrieval Society of India (IRSI[6]) since it started in 2008. It is very similar to CLEF and NTCIR, where the FIRE aims to encourage research on South Asian language Information Access technologies by providing reusable large-scale test collections experiments, a shared infrastructure for comparing the performance of different IR systems and discover new IR or IA tasks as the information

[4] http://fire.irsi.res.in/fire/2017/home

[5] https://inex.mmci.uni-saarland.de/data/publications.jsp

[6] http://www.irsi.res.in comparing the performance of different IR systems /

needs evolve and emerge. It covers various South Asian languages such as Bangla, Hindi, Marathi, Punjabi, Tamil, and Telugu. The test collection uses news corpora from the years 2004 until 2007 for each of the languages (Mitra, 2008).

The INEX started in 2002 with the aim to provide an infrastructure and means in the form of a large XML test collection. It also provided appropriate scoring methods for the evaluation of content-oriented retrieval of XML documents (Gövert & Kazai, 2002). The INEX created a set of XML test collection consisting of IEEE Computer Society publications between 1995 and 2002, 60 topics and graded relevance assessments. Then the retrieval effectiveness of the participating organizations' XML retrieval approaches results were evaluated and compared (Gövert & Kazai, 2002).

The CLEF, NTCIR, FIRE, and INEX are all similar to TREC which aims to encourage research in information retrieval field based on large-scale test collections. Although some focus on European languages, Asian languages and South Asian languages, they have all aimed to provide the necessary infrastructure and platform for information retrieval research.

2.6 Summary

This chapter has discussed the general literature review of information retrieval, and the system-oriented and user-oriented information retrieval evaluations. Characteristics and explanations about the common effectiveness metrics, precision, recall, average precision, and rank-biased precision were included. Two different statistical methods, the descriptive and inferential statistics were also listed. The statistical significance measures, ANOVA and correlation coefficient, were briefly discussed while comparing them to the other common statistical significance measures. The chapter ends with

information on Web track and Robust track test collections from Text REtrieval Conference (TREC) and briefly mentioning the other available test collection initiatives.

CHAPTER 3: EVALUATING INFORMATION RETRIEVAL SYSTEMS USING EFFORT BASED RELEVANCE JUDGMENTS

This chapter comprises of an experimentation to evaluate the information retrieval systems using effort based relevance judgments. The experimentation focuses on generating effort based relevance judgments and evaluating its impact on the system rankings. Section 3.1 details the background of the study, define the problem statement, lists the research questions and the objectives. Section 3.2 details the importance of effort in addition to relevance in evaluating retrieval systems, the effect of topic size and relevance on the retrieval system performance, and the various document effort features. Next, Section 3.3 explains the selection of test collections, effort feature classification and grading using the proposed approach, the creation of low effort relevance judgments and the evaluation of retrieval systems at various depths using the generated low effort relevance judgments. Section 3.4 describes the results of the experimentation and the discussions on the evaluation of retrieval systems using low effort relevance judgments, groups of systems and reduced topic size. Lastly, the summary of the chapter follows.

3.1 Background

There are two categories of information retrieval evaluation, the system-oriented evaluation, and user-oriented evaluation. In the system-oriented evaluation, the relevance judgment is one of the important aspects of a test collection. The relevance judgment usually contains information about the relevancy of documents concerning queries. In TREC environment, the topic experts judge the documents in regards to each of the queries. In user-oriented evaluations, the interaction of the actual users with the retrieval

systems is measured. Nonetheless, one of the disagreements between both the information retrieval evaluation categories is the consideration of relevance by the expert judges and the utility of the documents to actual users (Yilmaz, Verma, Craswell, Radlinski, & Bailey, 2014).

System-oriented or test collection based evaluation has always prioritized relevance as a way of measuring the effectiveness of retrieval systems. However, recent studies have shown that effort in consuming relevant documents is equally important for user satisfaction (Verma, Yilmaz, & Craswell, 2016; Yilmaz et al., 2014). The effort in this context is referring to the amount work needed by the user to find and identify the relevant content in the document. The effort can be classified as low effort or high effort. Low effort indicates less amount of work needed by the user to identify the relevant context within a document while high effort requires more work by the user to identify the relevant content within a document.

Studies have measured the amount of effort needed to judge the relevance of documents using expert judges or crowdsourced judges (Villa & Halvey, 2013; Yilmaz et al., 2014). These studies state that real users do not put in as much effort as the expert judges to judge the relevancy of the documents. The real users also tend to give up easily while looking for relevant information compared to expert judges who are trained to identify relevance from a document (Yilmaz et al., 2014).

Aspects such as the size of the document being judged increase the judgment effort and the degree of relevancy also tend to impact the amount of effort needed for judgment (Villa & Halvey, 2013). The effort is important for user satisfaction, and thus it is suggested that effort required to consume a document is included as part of ranking the retrieved documents (Yilmaz et al., 2014). In order to measure the utility of a retrieval system to an actual user, judges should be asked to provide the amount of effort needed

to find the relevant information (Verma et al., 2016; Yilmaz et al., 2014). However, acquiring additional information from judges expand their workload.

Despite the usefulness of obtaining effort from human judges, studies have shown the amount of instructions to judges cause variation between judgments in test collection (Webber, Toth, & Desamito, 2012), assessors inconsistencies (Chandar, Webber, & Carterette, 2013; Scholer, Turpin, & Sanderson, 2011; Voorhees, 2000), and assessor errors impacting system rankings (Carterette & Soboroff, 2010). Other studies attempted to understand the variation in human judgments by understanding the evolving nature of human judgments (Anderson, 2006), readability and cohesiveness which are independent of the topic (Chandar et al., 2013). As such, there exist possibilities of variations in judges providing information about the amount of effort needed to judge a particular document. Since measuring effort can be rather subjective, different judges may classify effort according to their background.

One of the focus in Verma et al.'s (2016) study is to show the difference in the performance of the retrieval systems due to the effort. The top 10 performing systems using P@10 effectiveness measure was used to compare with the systems evaluated using low effort P@10. Based on Kendall's tau values ranging between 0.53 to 0.71 for three different test collections, it was claimed that there was a significant difference in performance of the top 10 systems with low effort relevance judgments. However, the remaining systems within the test collections were not detailed for the correlation coefficient. Also, only metric P@10 was taken as a measure to determine the significant difference between the original and low effort relevance judgments' system rankings.

3.1.1 Problem Statement

The effort in addition to relevance is a major factor for satisfaction and utility of the document to the actual user. Real users give up easily and they also do not put in as much effort as expert judges while identifying relevancy in a document (Villa & Halvey, 2013; Yilmaz et al., 2014). Therefore, a retrieval system incorporating retrieval of low effort documents is preferable by the user due to lesser work in identifying relevancy of the documents compared to a retrieval system retrieving high effort documents. Consequently, it is vital to evaluate the retrieval systems based on the amount of effort needed in identifying relevancy of documents to ensure user satisfaction.

The importance of effort is measured in various ways in previous studies (Verma et al., 2016; Villa & Halvey, 2013) but limited depth of evaluation and retrieval systems (top 10 only) were used to show the differences in system rankings due to low effort relevance judgments. The differences in system rankings using original and low effort relevance judgments beyond evaluation depth 10, and the entire or different groups of retrieval systems within a test collection is unknown. Possibilities exist that the differences in system rankings may vary widely as a result of larger numbers of relevant documents found with deeper depth of evaluation. Similarly, top performing systems has always been a favorable group in retrieval system evaluation due to their ability to retrieve higher numbers of relevant documents. However, it is unclear if similar conclusions can be drawn if the group of systems is evaluated based on retrieval of low effort relevant documents.

Attempts to reduce the amount of work needed for judging relevance had been an attention of the research community. Any advancement that reduces the workload on relevance judgments without jeopardizing the quality of evaluation is an added advantage (Guiver, Mizzaro, & Robertson, 2009). Nevertheless, inquiring the amount of effort

needed for relevance judgment from the judges (Verma et al., 2016; Yilmaz et al., 2014) may cause variation in judgments (Carterette & Soboroff, 2010; Chandar et al., 2013; Scholer et al., 2011; Webber et al., 2012) and the effort needed. Possible advancements in overcoming these drawbacks are minimizing or eliminating the involvement of human judges in obtaining effort information, and the evaluation of retrieval systems with reduced topic size.

3.1.2 Research Questions

RQ1. Does low effort relevance judgment cause variation in the system rankings when evaluated at a deeper depth of evaluation?

RQ2. Is there variation in system rankings for a specific group of retrieval systems when using low effort relevance judgment for evaluation?

RQ3. Can the retrieval systems using low effort relevance judgments be evaluated effectively with reduced topic size?

3.1.3 Objectives

This study aims

OBJ1. To propose a systematic way of generating relevance judgments that incorporate effort and determine the variation in system rankings due to low effort relevance judgment in evaluating retrieval systems at different depth of evaluation.

OBJ2. To measure the variation in system rankings due to low effort relevance judgment on groups of systems in information retrieval system evaluation.

OBJ3. To explore the effectiveness in evaluating retrieval systems using low effort relevance judgment with reduced topic sizes.

3.2 Literature Review

This section details the importance of measuring retrieval systems by effort in addition to the document relevance, and the effect of topic size and relevance to retrieval system performance. Lastly, the section details the various features for measuring document effort.

3.2.1 The Importance of Effort in Addition to Relevance

Between the user-oriented and system-oriented information retrieval evaluation, preference has always been for the latter because of repeatability, short experimental turnaround, and low cost. In the system-oriented evaluation, relevance had always been a priority in measuring the performance of the systems using effectiveness metrics. Relevance has been thought to be the utility a user gains when viewing a ranked document. Therefore, a system that ranks relevant documents earlier in the ranking list is considered a more effective system (Moffat et al., 2012). Hence, an effectiveness metric should capture the gist of ranked relevant documents to score the retrieval systems.

However, there exist cases of noncorrelation between the effectiveness metrics and the real user experience (Moffat et al., 2012). Also, the outcome of system-oriented evaluation does not agree with real user satisfaction (Hersh et al., 2000; Turpin & Hersh, 2001). It is due to the factors such as effort, system and user effectiveness, and user characteristics (Al-Maskari & Sanderson, 2010) influencing user satisfaction. The effort could be divided into findability, readability and understandability factors (Verma et al., 2016). Recently, studies measured effort needed for judging relevant documents (Villa & Halvey, 2013), and identified effort as an important aspect of user satisfaction in addition to relevance (Yilmaz et al., 2014).

Experimentation on the effort needed for judging relevant documents showed that increased document size and the degree of relevance of a document indicating 'relevant' requires more effort in judging (Villa & Halvey, 2013). The experimentation was conducted using TREC HARD 2005 track and AQUANT test collections. During and after judging, the NASA TLX was used to gather the judges' action and perception of the task effort. The online study assumed that a user would look through the ranked list in ascending order. Additionally, behavioral data through the time taken to make judgments and number of topic view clicks were measured. Besides the amount of effort needed for judging, the study also showed accuracy is not affected by document length but by the degree of relevance (Villa & Halvey, 2013).

In 2014, experimentation was conducted to measure the effort needed by users to identify and consume the information from a document with regards to time (Yilmaz et al., 2014). When compared, they found a mismatch between relevance judgments from real users and expert judges. The experimentation utilized dwell time and click counts to obtain real users judgments. They argue that utility of a document to actual users is dependable in the effort needed to consume the relevant information (Yilmaz et al., 2014).

Another study focused on measuring readability, findability and understandability effort in obtaining relevance judgment and stated that effort should be a part of relevance judgment if user satisfaction is prioritized (Verma et al., 2016). They even indicated user satisfaction is a function of relevance and effort. High-effort documents are harder to consume. Thus it is less likely for the users to read it (Yilmaz et al., 2014). Experimentation also resulted in a high correlation of satisfaction and findability with user preferences compared to readability and understandability (Verma et al., 2016). It was previously known that readability effects assessor disagreement, whereby documents that are easy to read causes disagreements (Chandar et al., 2013) Also, Collins-Thompson

(2011) advocated that reading difficulties of documents are a major factor in large-scale analyses of personalized information retrieval systems.

The logistic regression models are used to predict the relevance assessment using initial user input on relevance and effort (Chandar et al., 2013; Verma et al., 2016; Yilmaz et al., 2014). With the incorporation of effort into relevance judgments, Kendall's tau correlation coefficient shows the performance of the retrieval systems are different compared to when the effort is not used in the relevance judgments (Verma et al., 2016). The result highlights the importance of effort together with relevance in satisfying user's information needs.

3.2.2 The Effect of Topic Size and Relevance to System Performance

Attempts to reduce the amount of work needed for judging relevance had been an attention of the research community. Any advancement that reduces relevance judgments workload without jeopardizing the quality of evaluation is an added advantage (Guiver et al., 2009). One possible way to that is via the reduction of topics in evaluating information retrieval systems.

Topics highly influence the comparison between retrieval systems. It is likely for a different set of topics of the same size to produce different results (Voorhees & Buckley, 2002) and some topics or topic sets predict the actual effectiveness of systems better than others (Berto, Mizzaro, & Robertson, 2013; Guiver et al., 2009). An analysis of system ranking estimation proved that performance depends heavily on a topic subset (Hauff, Hiemstra, Jong, & Azzopardi, 2009). The experiment was conducted on different estimation approaches such as data fusion, random sampling, document similarity, and document score. They also identified the random sampling approach (Soboroff, Nicholas,

& Cahan, 2001) as the most stable and best-performing estimation method. With the use of eight test collections (TREC 6 – 10, and Terabyte track 4 – 6), experimental results suggest the right topic selection could improve the performance estimation by 32% on average.

An attempt was made to understand approaches to develop better topics instead of better retrieval systems (Culpepper, Mizzaro, Sanderson, & Scholer, 2014). The two proposed approaches to quantify topic goodnesses are topic ease and topic set predictivity. Topic ease means the system runs can produce high AP values for topics (Mizzaro & Robertson, 2007). The experiment was conducted on TREC-1 to TREC-8 test collections. The topic ease decreased for the first five years and started to increase afterward. That translates to easier topics during TREC-8 and harder for TREC-9. As for the topic predictivity, Kendall's tau and linear correlation provide the result about the MAP closeness of the topic subset to the full topic. The results suggest topic subsets are less able to predict the system ranking of the full topic set. However, the authors highlight the positivity of poor topic predictivity. It means the topics are not redundant from year to year (Culpepper et al., 2014). Although there is variation in system evaluation due to topics, metrics AP, RPrec (recall-precision), P@10, and GMAP (geometric mean average precision) produced comparable results with topic sets (Guiver et al., 2009).

A good experimental design requires a sufficient number of topics in a comparative retrieval system evaluation (Voorhees & Buckley, 2002). In TREC, the preferred topic size is 50, and a minimum topic is 25 to counter the cost needed for relevance judgments (Voorhees, 2000). However, minimum judgment effort only requires 3 hours to have 95% of confidence in the evaluation outcome (Carterette, Allan, & Sitaraman, 2006). Nonetheless, a larger number of topics, stable metrics and larger differences among the system scores would increase the reliability of the results (Voorhees & Buckley, 2002).

They also stated 95% confidence could be on the system rank orderings with 50 topics, provided the absolute difference in MAP be 5% to 6%. Error rates also decreased greatly for P@10 compared to MAP with the increase in topic set size (Sanderson & Zobel, 2005).

Topics influence the comparison of system performance, but the document relevance also influences the evaluation of retrieval systems. An experiment was conducted to examine the effect of highly relevant documents to retrieval system evaluation (Voorhees, 2001). The study shows the relative effectiveness of different Web track runs differ when evaluated using only highly relevant documents versus relevant documents. However, the effect of changes due to judgments in TREC test collections shows that relative performance of the system runs was comparable (Voorhees, 2000).

Recently, a study (Jones, Turpin, Mizzaro, Scholer, & Sanderson, 2014) showed that the number of relevant documents is not an effect of evaluation inconsistencies. Evaluation inconsistencies exist with different sub-collections formed based on document source, but the number of relevant documents was uniform in each sub-collection. Nevertheless, the study did not state the effects causing the inconsistencies in the evaluation.

Previously, Voorhees (2001) mentioned the change in effectiveness could be due to small numbers of relevant documents causing the retrieval measures to be unstable. Queries with fewer than five relevant documents could even cause the average precision measure to become unstable causing variation to the system performance and rankings (Voorhees, 2000). Yet with a constant number of relevant documents, the evaluation results are inconsistent (Jones et al., 2014). A possible cause of inconsistency could be the changes in document ranking. Since small changes in document ranking can produce large variations in the evaluation scores (Voorhees, 2000).

3.2.3 Features for Measuring Document Effort

Many factors influence user satisfaction (Al-Maskari & Sanderson, 2010) and these factors can be broken down into measurable features. Figure 3.1 shows some of the document features from previous studies (Chandar et al., 2013; Verma et al., 2016; Yilmaz et al., 2014) but are not exhaustive. Simple document features, reading level features and metarank feature have been used to predict assessor disagreement (Chandar et al., 2013). Meanwhile, readability level of the document, document length, and location of query terms in the document was used to measure effort needed in judging a document (Yilmaz et al., 2014).

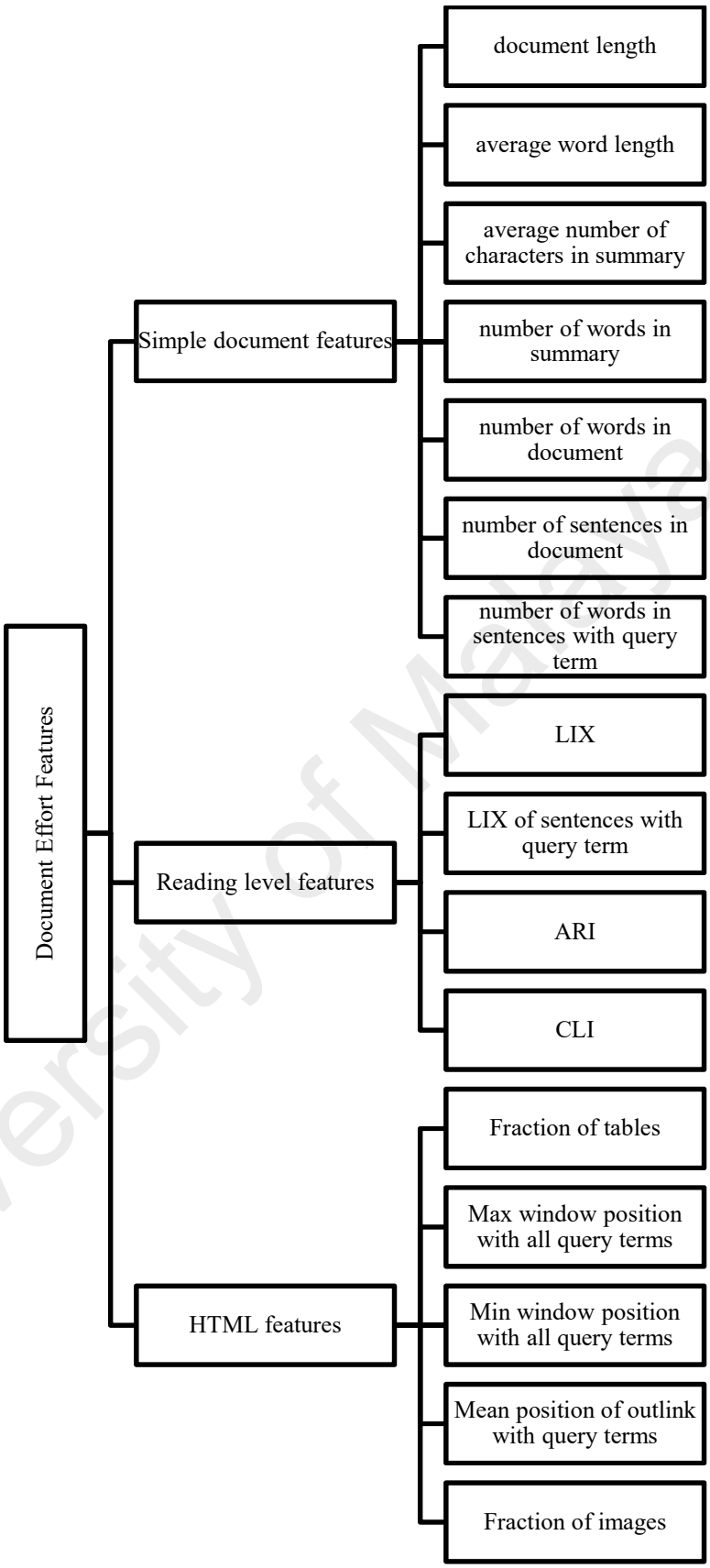

Figure 3.1: Document effort features

From the list in Figure 3.1, some features will be used as part of this study and the details or formulae for these features are provided below. The features will be used to generate the scores for each document and create the effort based relevance judgments in the context of this experimentation.

1. **Number of characters in a document (nchar)** computes the number of characters in the document. Characters include digits, alphabets, and punctuations.
2. **Number of words in a document (nword)** is simply the count of words in the document. The number of words is counted by each space (Smith & Senter, 1967).
3. **Number of sentences in a document (nsent)** is the number of sentences in the document, where sentences end with a period, question mark or exclamation mark followed by a space.
4. **LIX** is a readability measure to calculate the difficulty of reading a document. The number of periods is defined by period, colon or capital first letter. Words with 6 or more letters are considered as long words. It can be computed using the following formula.

Equation 3.1

$$\frac{number\ of\ words}{number\ of\ periods} + \frac{(number\ of\ long\ words\ \times 100)}{number\ of\ words}$$

5. **LIX of sentences with query term (LIXsent)** is LIX computation on sentences which have query terms. It uses the same formula from LIX.
6. **ARI** (Automated Readability Index) measures the readability of a document and can be defined by the following formula to obtain the grade level (*GL*) of a document.

Equation 3.2

$$GL = 0.5\left(\frac{number\ of\ words}{number\ of\ sentences}\right) + 4.71\left(\frac{number\ of\ characters\ or\ strokes}{number\ of\ words}\right) - 21.43$$

The number of sentences can be tabulated when ending with a period, question mark or exclamation mark and followed by space (Smith & Senter, 1967). The number of words is counted each time the space bar is depressed. Characters or strokes is advanced one count for each printing-key on the keyboard which includes all the letters, numbers and punctuation marks (Smith & Senter, 1967). A low-grade reading level of a document requires less effort, but it provokes assessor disagreement. Possibly reading level measure picks other aspects of document content (Chandar et al., 2013).

7. **CLI** (Coleman-Liau index) is also a readability feature and is defined by the following formula.

Equation 3.3

$$0.059L\ - 0.296S\ - 15.8$$

where, L is the average number of letters per 100 words and S is the average number of sentences per 100 words. The CLI is similar to ARI.

8. **Number of tables (ntab)** is computed by counting the number of tables within a document.
9. **Number of images (nimg)** is computed by counting the number of images within a document.

With these effort features, document scores can be calculated and an effort based relevance judgment can be created to evaluate the retrieval systems. The grading of documents based on these effort features is in the Methodology section below.

3.3 Methodology

This section describes the proposed method of generating a low effort relevance judgment systematically, the selection of test collections, classification and grading of effort features, and evaluating the retrieval systems using the newly generated low effort relevance judgments. Low effort relevance judgments should assess and identify retrieval systems which can satisfy actual users. As known earlier, real users tend to give up easily while trying to capture the relevance of the document. Systematic generation of relevance judgment indicates the adaptability and repeatability of calculating the feature scores of documents, and creating relevance judgments for different test collections with minimal experimentation variation. In an attempt to incorporate low effort in relevance judgments, this study identifies and incorporates suitable effort features from past studies. Previously, various effort features showed an impact on user satisfaction (Verma et al., 2016; Yilmaz et al., 2014). Therefore some of those features are selected and customized for the current study.

3.3.1 Test Collections

In the real Web environment, users would submit queries to search engines or information retrieval systems and expect for information that fulfills their need. Based on the submitted queries only, without any other information, the retrieval systems extract relevant information (Hawking & Craswell, 2002). Therefore, to replicate the real Web

query and retrieval systems scenario, this experimentation utilizes only retrieval systems that utilized title-only query terms from the test collections.

Two different choices of test collections for this experimentation is the TREC-9 Web track and TREC-2001 Web track. In TREC-9 test collection, 40 retrieval system runs used title-only in their retrieval process and contains 50 topics. The TREC-2001 Web track has two separate tasks, the Web ad hoc task and the home page finding task. However, only the Web ad hoc task system runs were considered for this experimentation. There was 77 title-only system runs in the Web ad hoc task and also contains 50 topics. The topics from the TREC-2001 Web ad hoc task have a similar form with TREC-9 topics, but the topic title for the prior is a real Web query taken from search engine logs (Hawking & Craswell, 2002). Both the test collections used the WT10g dataset which comprises of Web data.

The selected test collections are similar and suitable to replicate a real Web user query. Each has same numbers of topics and uses the same datasets for the retrieval of relevant documents. Although TREC-2001 Web track has slightly more system runs, the difference in number should not impact the results of the experimentation since the system runs will be evaluated and ranked relative to the other system runs within the test collection.

3.3.2 Classifying Effort Features and The Boxplot Approach for Document Grading

Various features could measure effort as discussed earlier in Section 3.2.3. Although the list is not exhaustive, some of the effort features showed an impact on user satisfaction (Verma et al., 2016) and relevance disagreements (Chandar et al., 2013). The effort

measurements are only for relevant documents since effort is a measure that can satisfy users only when the document is relevant (Verma et al., 2016). Therefore, this study will adopt the same by calculating effort for relevant documents only.

A systematic grading versatile for multiple test collections is needed. Previous studies utilized logistic regression models to predict the relationship between effort features and assessor disagreement (Chandar et al., 2013), to predict findability labels obtained from effort judging (Verma et al., 2016), and to predict the dwell time from the effort features (Yilmaz et al., 2014). This study attempts to grade and classify the effort features differently in a possible simple manner to create the effort based relevance judgments.

The effort scores for each feature can be calculated based on the description and formula of each as specified in Section 3.2.3. However, the grading of documents for each effort feature will be customized for this study using the boxplot approach. The boxplot approach is a simple yet repeatable approach in classifying documents' effort.

The boxplot usually identifies outliers and data plots within 25% and 75% interquartile range. Figure 3.2 shows the boxplot for effort feature 'number of words' for 362 relevant documents in the TREC-9 Web track ad hoc task. The upper horizontal line outside the box is the upper inner fence and the circles above this horizontal line are the suspected outliers and outliers. The suspected outliers and outliers are not removed from the evaluation but classified accordingly. The scores within 0 and the upper inner fence is divided into two to create binary relevance. The feature scores could only start from 0 and will not hold negative values, as it does not make sense to have negative scores. For example, the number of words could never be a negative value.

With the use of boxplot, documents can now be classified and graded for each effort features accordingly. Boxplot is a simple approach that could fairly divide the effort

scores while ensuring outlier scores do not skew the grading of the entire set of documents. However, the boxplot is only implemented for simple document features and HTML features. These features do not have a standard guideline as to which level of measurement constitutes to low or high grades. In contrast, the readability features have standardized grades that could be easily suited to this study.

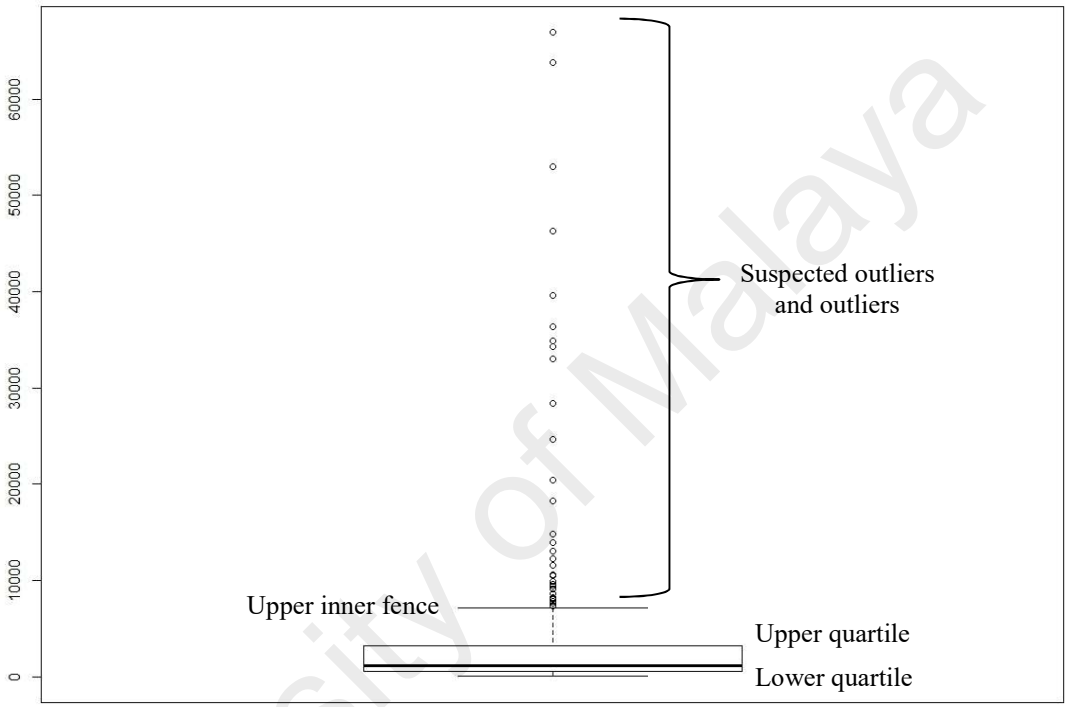

Figure 3.2: Boxplot representation with upper inner fence, suspected outliers, and outliers

If using the boxplot approach, the classifications of effort are dependable in the set of documents from that particular test collections. A drawback arising from the usage of boxplot for document grading is that the test collections may not be comparable one-to-one. It is because a document graded as a low effort in one test collection may not necessarily be graded the same in another test collection. The variation occurs based on the relevant documents within the test collection and the interquartile range computation. This shortcoming is only valid for the simple document features and HTML features

because there is no standard of classification to indicate low or high effort documents. In contrast, the readability effort features are easily comparable between test collections due to the fact there are standardized grades. Nonetheless, the correlation coefficient of system rankings can be measured reliably within the test collection and should not cause discrepancies.

Below are the effort features and the possible effort grades for a document implemented using boxplot or standardized grading.

1. Simple document features
    a. Number of sentences in a document (nsent) – Calculates the number of sentences in the document, where sentences end with a period, question mark or exclamation mark followed by a space. A low number of sentences indicate low effort.
    b. Number of words in a document (nwords) – Calculates the number of words in a document. The number of words is counted by each space (Smith & Senter, 1967). A low number of words indicate low effort.
    c. Number of characters in a document (nchar) – Calculates the number of characters in a document. It is calculated by counting the number of alphabets, numbers, and punctuations. A low number of characters indicate low effort.
2. HTML features
    a. Number of images (nimg) – Calculates the number of images within a document. Zero image is taken as high effort while the remaining number of images classification will follow the boxplot approach. Images are useful in spotting information in a document, and more images are better (Verma

et al., 2016). At the same time, too many images may not be suitable to capture the relevant content of the document.

b. Number of tables (ntab) – Calculates the number of tables within a document. Zero table is taken as high effort while the remaining number of tables classification will follow the boxplot approach. Similar to images, tables are helpful in identifying information but too many tables make it difficult to find the information needed (Verma et al., 2016).

3. Readability features

a. ARI – Low ARI level of a document requires low effort. The following classification will be made for each of the grading and ARI levels in this experimentation.

Table 3.1: Effort classification for ARI feature (first two columns taken from (Smith & Senter, 1967))

Grades	ARI level	Effort classification
5-6 years old – Kindergarten	0	Low effort
6-7 years old – First Grade	1	Low effort
7-8 years old – Second Grade	2	Low effort
8-9 years old – Third Grade	3	Low effort
9-10 years old – Fourth Grade	4	Low effort
10-11 years old – Fifth Grade	5	Low effort
11-12 years old – Sixth Grade	6	Low effort
12-13 years old – Seventh Grade	7	Low effort
13-14 years old – Eighth Grade	8	Low effort
14-15 years old – Ninth Grade	9	Low effort
15-16 years old – Tenth Grade	10	Low effort
16-17 years old – Eleventh grade	11	Low effort
17-18 years old – Twelfth grade	12	Low effort
18-22 years old – College	13	**High effort**

b. CLI – A low CLI level indicates low effort. The following effort classifications are made for each CLI level.

Table 3.2: Effort classification for CLI feature (first two columns taken from (Gústafsdóttir, 2017))

Grades	CLI level	Effort classification
Sixth Grade	≤ 6	Low effort
Seventh Grade	7	Low effort
Eighth Grade	8	Low effort
High school freshman	9	Low effort
High school sophomore	10	Low effort
High school junior	11	Low effort
High school senior	12	Low effort
College freshman	13	**High effort**
College sophomore	14	**High effort**

c. LIX – A low LIX score indicates low effort. The following effort classifications are made for each LIX scores.

Table 3.3: Effort classification for LIX feature (first two columns taken from (Gústafsdóttir, 2017))

Reading ease	LIX scores	Effort classification
Very easy to read	< 20	Low effort
Easy for practiced readers	20 - 30	Low effort
Average reading level	30 - 40	Low effort
Difficult	40 - 50	Low effort
Very difficult	> 50	**High effort**

d. LIX for sentences with query terms (LIXsent) – Calculates the LIX on sentences which have query terms. Query terms are terms that are from the title-only query (topic) which appears in sentences within a document. The classification of LIXsent effort feature is similar to LIX.

The classification of effort features is used to grade documents with the boxplot approach implementation for simple document and HTML effort features, and standardized grading approach for readability features. The readability effort features

have standardized grades and will be classified in a possible equivalent manner. For both ARI and CLI, high effort constitutes to college level. The LIX, on the other hand, classifies 'very difficult' as high effort with the assumption this level is equivalent to 'college' level from CLI and ARI features. Therefore, all three readability effort features tend to classify the documents' effort in a somewhat equivalent manner.

The boxplot approach is shown in Figure 3.3. In the boxplot approach, calculate the document scores for each feature. Then sort the document scores for each feature in ascending order. Next, calculate the interquartile range of the feature scores by subtracting the lower quartile (0.25*(N+1)) from the upper quartile (0.75*(N+1)). Following this, calculate the upper inner fence score using the formula below.

$$upper\ inner\ fence = upper\ quartile + (1.5 \times interquartile\ range)$$

Then, divide the upper inner fence feature score into two for translation as binary relevance. The lower half indicates low effort (marked as relevant) and the upper half indicates high effort (marked as non-relevant). Any document that has a feature score that falls within the lower half will be marked as low effort relevant while the remaining is characterized as high effort non-relevant.

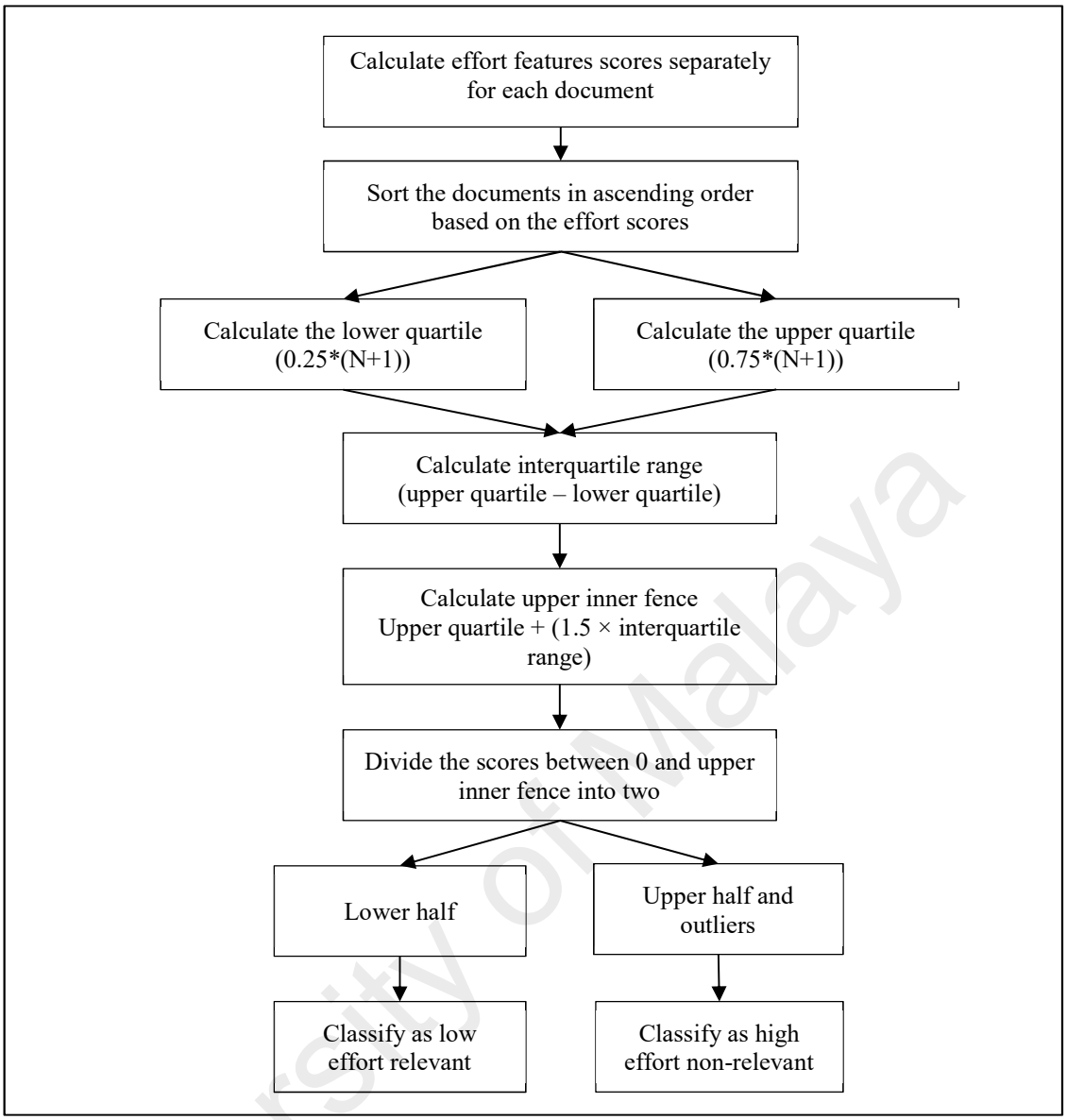

Figure 3.3: The boxplot approach in classifying document grades based on effort features

The outliers and suspected outliers are those document scores above the upper inner fence. The outlier and suspected outlier documents will be marked as high effort, thus non-relevant since their scores lie beyond the upper inner fence. With the grading of the documents based on their effort feature scores, a low effort relevance judgment can be created.

3.3.3 Generating the Low Effort Relevance Judgment and Measuring the Performance of Retrieval Systems

The low effort relevance judgment can be created from the proposed approach discussed in the previous section. Figure 3.4 shows the steps undertaken in this experimentation and the generation of low effort relevance judgment using the boxplot approach or standardized grading approach. The dotted box in the figure indicates the proposed approach while the remaining constitutes to the existing approach of creating relevance judgments and evaluating retrieval systems.

Firstly, identify and select the title-only retrieval system runs. For each of the 50 topics, choose the top *k* documents, where *k* is the evaluation depth holding values of 10, 100 and 1000. Identify the relevance for each of these documents from the original relevance judgments. Calculate the effectiveness of the retrieval system runs using specific metrics at the various evaluation depth. These steps constitute the existing method to calculate the effectiveness scores for the system runs.

As for the proposed method, calculate the document scores for each effort feature. As stated earlier, the effort is only measurable for relevant documents. Grade the documents using the boxplot approach for simple document features and HTML features or the standardized grading for readability features. The boxplot approach is directly applied to nchar, nword and nsent features belonging to the simple document features.

As for the ntab and nimg features belonging to the HTML features, there is a slight variation to the grading of effort documents. As mentioned earlier in section 3.3.2, a document containing zero table or image is considered as high effort and thus non-relevant. Since, an image or table increases the chances of identifying relevant information (Verma et al., 2016). Therefore, the nonexistence of a table or image is assumed to be a high effort document for user consumption. The remaining of the

documents are classified accordingly based on the boxplot approach. A document having feature score within the lower half will be marked as low effort relevant (except a document containing zero image or table) while the remaining is graded as high effort, non-relevant.

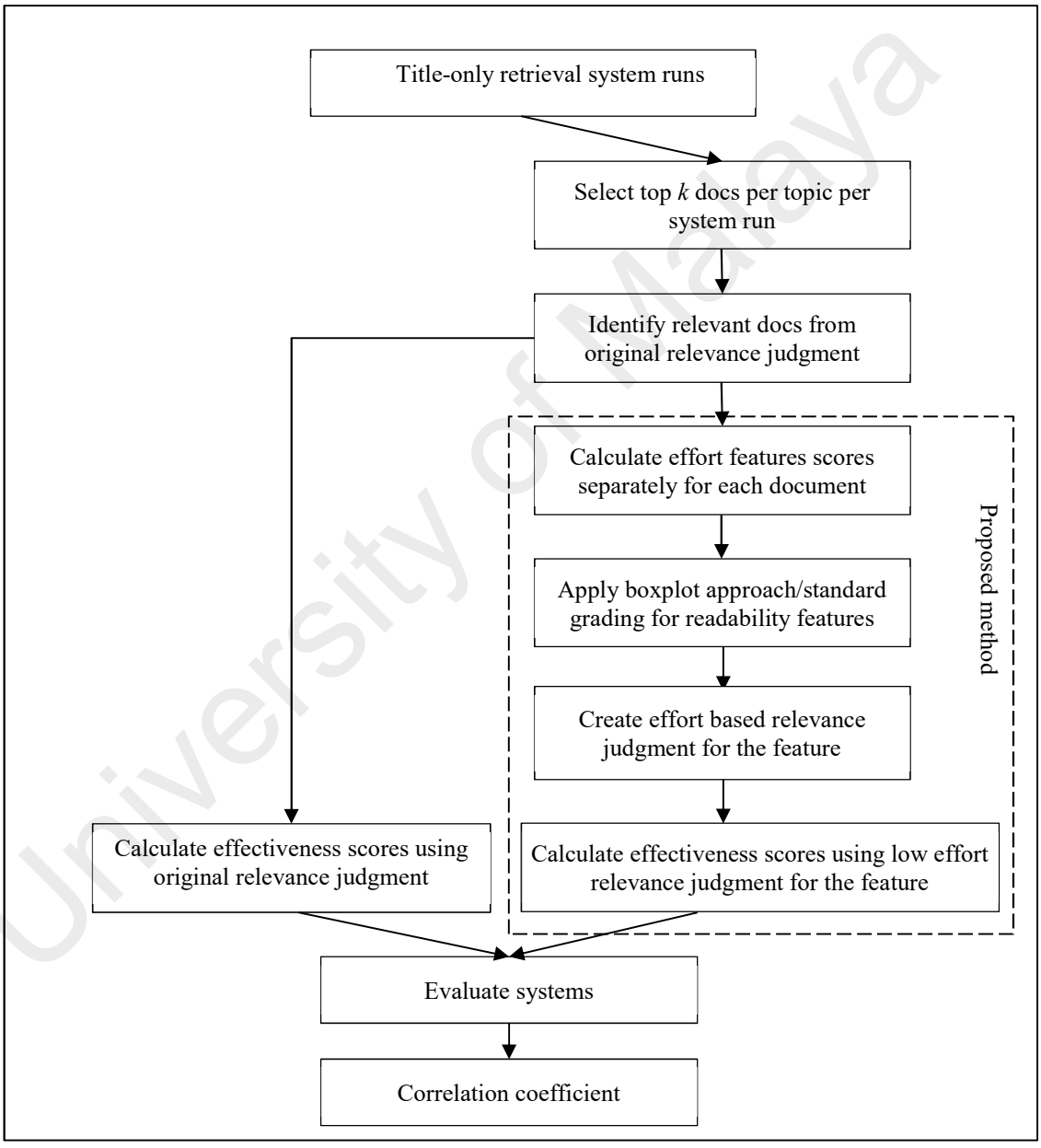

Figure 3.4: Generating effort based relevance judgment with the proposed boxplot approach or standardized grading approach and evaluation of retrieval systems

As for the readability features, ARI, CLI, LIX and LIXsent, the classification of effort is according to the standardized grades mentioned in section 3.3.2. These features do not employ boxplot approach but use a standardized grading, making it easily repeatable and comparable between test collections.

Now with the effort graded documents, generate a low effort relevance judgment, *eqrel* (refer Figure 3.4). Repeat document score calculation for each effort features separately, grade the documents accordingly with the proposed approach, and generate the low effort relevance judgment for all effort features and evaluation depth.

Each feature and evaluation depth combination has its own eqrel. For example, there are nine effort features and three evaluation depth considered. Therefore, there are a total of 27 eqrels for evaluating the retrieval systems in a test collection. The effectiveness of each system run is computed using the individually created eqrel (refer Figure 3.4). The effectiveness scores obtained through the use of eqrel can only be lower or equal to that of original relevance judgment, *qrel*. The effectiveness scores could never exceed those from qrel because effort classification is for relevant documents found in the qrel only. Therefore, the eqrel will never have more relevant documents than those found in the qrel. Only in best case scenario, an eqrel could have marked all the relevant documents from qrel as low effort.

Finally, the system runs can be evaluated using Kendall's tau correlation coefficient between the system ranks from qrel and the system ranks from the eqrel. The Kendall's tau would indicate the similarities or differences in system rankings using the existing and the newly created low effort relevance judgments. A strong correlation coefficient would indicate incorporating low effort relevant documents in measuring system effectiveness does not influence the retrieval system ranks. While a weak correlation

coefficient could mean the system effectiveness scores are impacted due to the low effort relevant documents as claimed by Verma et al. (2016).

3.3.4 Evaluating Retrieval Systems Using Low Effort Relevance Judgments

The evaluation of retrieval systems using low effort relevance judgments is conducted for different depth of evaluation. The evaluation depths attempted are 10, 100 and 1000. The purpose of different evaluation depth is to determine the variation in system rankings due to eqrel with deeper evaluation. For each of this evaluation depth, the system effectiveness is calculated using qrel, and each feature's eqrel, separately. The correlation coefficient of system ranks is performed between the same evaluation depth from both qrel and eqrel. Two different metrics were attempted, P@k and AP@k, where k = 10, 100, and 1000.

The evaluation of retrieval systems using eqrel is extended into specific groups of systems. There are possibilities that top 10 or bottom 10 systems from the original qrel show variation in system rankings due to evaluation with the low effort relevance judgment. Therefore, this study attempts to identify the impact of eqrel in regards to their system rankings for these groups of systems. The experimentation continues to use the two metrics specified earlier and the various evaluation depth.

The final assessment using low effort relevance judgment is by reduced topic size. The experimentation is conducted to determine if lesser topic sizes could effectively evaluate the retrieval systems with eqrel. A reduction in topic size should reduce the work in generating relevance judgment but should not jeopardize the quality of evaluation.

A total of 50 topics were available for both test collections. The reduced topic evaluation is shown in Figure 3.5. Firstly, obtain the system ranks from the qrel for

metrics P@*k* or AP@*k*. The original system ranks consider all 50 topics for each system. These constitute as the baseline system ranks for comparison. Next, randomly select a reduced topic size, for example, 30 topics. Compute the effectiveness scores for the systems using these 30 topics with the qrel. Then, perform Kendall's tau correlation coefficient between the system ranks from baseline and reduced topic size from the qrel.

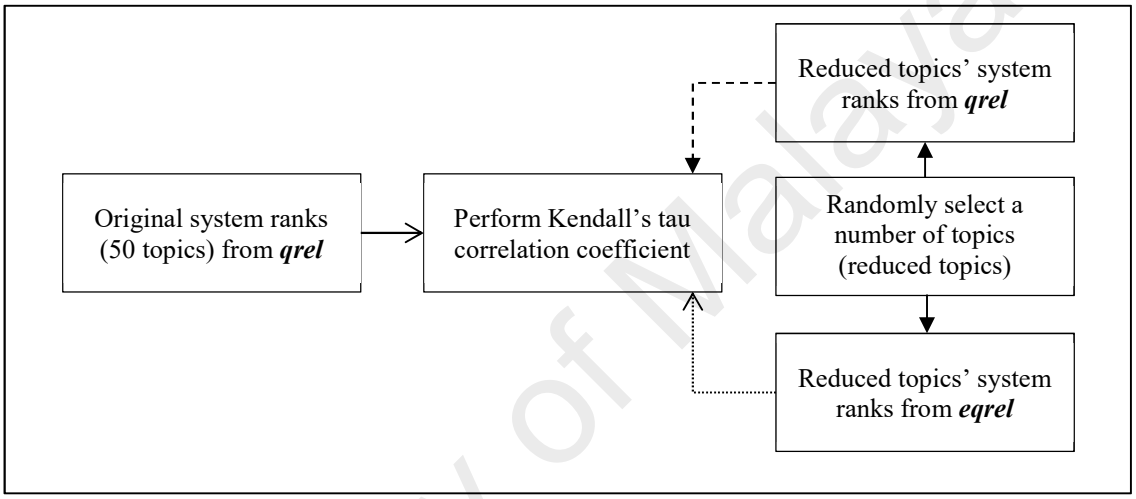

Figure 3.5: Measuring Kendall's tau correlation coefficient for reduced topic size using qrel and eqrel.

Then, with the same topics used in the qrel, compute the system effectiveness scores using eqrel. Rank the systems and perform Kendall's tau correlation coefficient between the system ranks from baseline and the reduced topic size from eqrel. For each eqrel, a different set of random topics is selected. However, ensure the same topics are used in the qrel to obtain the effectiveness, system rank and correlation coefficient for the evaluation. Since a different set of topics evaluate the retrieval systems differently (Voorhees & Buckley, 2002), the same topics are selected for evaluating with qrel and eqrel to measure only the changes due to effort features. Whereas, the random selection

of topics for each feature is to allow different combinations of topics to measure the retrieval systems.

The Kendall's tau values from reduced topic size using qrel and eqrel will be evaluated to determine the effectiveness of using eqrel for reduced topic size. If Kendall's tau value from eqrel is equally good or better than Kendall's tau value from qrel, the reduced topic could be a good alternative in measuring the system effectiveness using low effort relevance judgments.

3.4 Results and Discussions

This section consists of the experimental results and discussions. The results are presented in different subsections for each of the research questions. Firstly, Section 3.4.1 presents the correlation coefficient results between system ranks of original and the low effort relevance judgments for different depth of evaluation. Next, Section 3.4.2 details the correlation coefficient results between system rankings of original and low effort relevance judgment for groups of systems. Lastly, Section 3.4.3 presents the results of retrieval systems evaluation using reduced topic sizes for low effort relevance judgment.

3.4.1 Correlation Coefficient Evaluation of Individual Effort Features for Different Depth of Evaluation

The experimental results reveal the variation in system rankings due to low effort relevance judgments when evaluated at different depth and thus corresponds to RQ1. The impact on the system rankings will be evaluated with Kendall's tau correlation coefficient to determine the variation in the system rankings. Low effort relevance judgments were

created for each of the effort features separately and one with all features combined (labeled as AllFeature). The eqrel for AllFeature uses the feature grades of the documents from the individual effort features and deduce a final grade using mode. If five out of nine features have graded a document as low effort relevant, then by using mode, the document is also graded as low effort relevant. The performance of the retrieval systems is then measured using these low effort relevance judgments. Also, note that the effectiveness scores measured from the eqrel will always be lower than or at most equal to that of the qrel because low or high effort classifications are on relevant documents only. Nonetheless, the correlation coefficient should not be affected by the lower effectiveness scores using eqrel since Kendall's tau uses system ranks for evaluation.

The experimentation measured Kendall's tau correlation coefficient between the system rankings of original and low effort relevance judgments. A strong correlation coefficient (≥0.8) would indicate that the system rankings are not affected by the effort feature while moderate to poor correlation coefficient (<0.8) would indicate the system rankings are affected by the effort feature. The intention of such observation is to recognize any changes in the correlation coefficient due to the depth of evaluation.

Figure 3.6 shows Kendall's tau correlation coefficient for the various effort features evaluated at different depths for both test collections. Two different metrics, P@*k* and AP@*k,* were used for a combination depth, *k* = 10, 100 or 1000. The x-axis represents the individual features and AllFeature, and the y-axis represents Kendall's tau correlation coefficient values. The dotted horizontal lines in the graph indicate the tau value of 0.8. The detailed numbers used for plotting the Figure 3.6 is available in APPENDIX N (TREC-9) and APPENDIX O (TREC-2001).

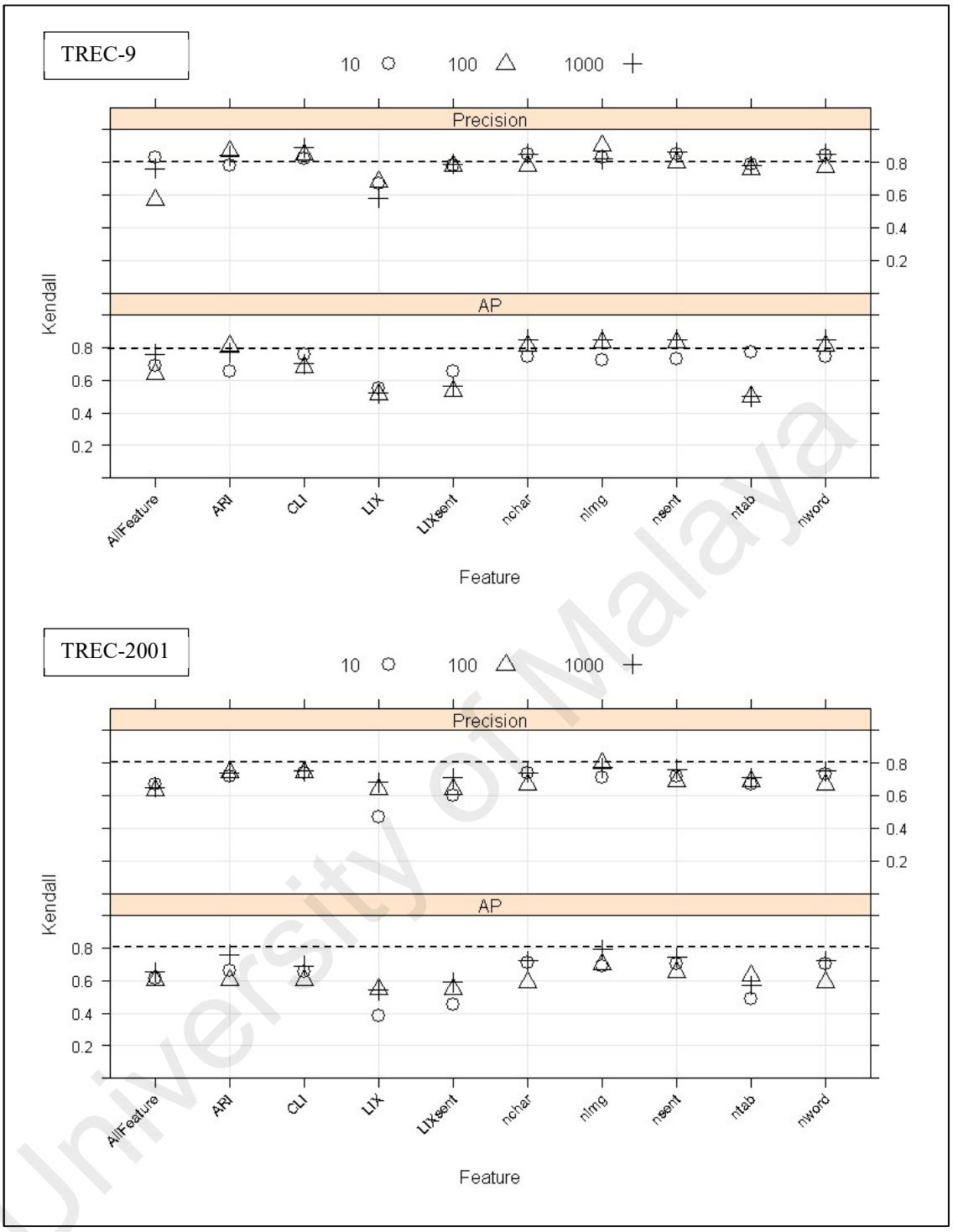

Figure 3.6: Kendall's tau correlation coefficient for various metrics between original system rankings and proposed method's system ranking using low effort relevance judgments, for different test collections.

Let's analyze the Figure 3.6 part by part. Firstly, the plots for TREC-9 P@*k* shows that depth of evaluation using eqrel affect the system rankings for some features. The system rankings for features such as LIX, LIXsent and ntab are all influenced by effort for all depth of evaluation. Whereas, the system rankings for features CLI, nimg and nsent are not impacted by the evaluation depth as all tau values fall at or above 0.8. As for the remaining features, ARI, nchar, nword, and AllFeature, the effort impacts the system rankings for some depth of evaluation.

As for the assessment with TREC-9 AP@*k* metric, AllFeature, CLI, LIX, LIXsent and ntab features have Kendall's tau values below 0.8 for all depth of evaluation. These features show the low effort relevance judgments have now caused changes in the systems' performances regardless of the evaluation depth. The remaining features show varying effect on the system rankings as some correlation coefficient values appear above while others are below 0.8, depending on the depth of evaluation.

The effort features that affect the system rankings have been identified through the correlation coefficient values. However, did the correlation coefficient become stronger or weaker with the depth of evaluation? For TREC-9 P@*k*, the increase in depth from 10 to 100 or 1000 has caused the correlation coefficient to remain equally good or stronger for ARI, CLI, LIXsent, and nimg features. The increase in Kendall's tau values with the increase in depth of evaluation shows effort does not necessarily impact system rankings with deeper depth of evaluation. Meanwhile, AllFeature, LIX, nchar, nsent, nword, and ntab have all caused Kendall's tau values to decrease with the increase in evaluation depth. A possible cause of the correlation coefficient changes could be the number of relevant documents in the different depth of evaluation.

Analyzing the same for TREC-9 AP@*k* shows that features ARI, nchar, nword, nsent and nimg have a stronger correlation coefficient with deeper depth while the correlation

coefficient gets weaker for CLI, LIX, LIXsent and ntab. The only exception is AllFeature that show the depth of 100 has weaker tau compared to depth 10, and stronger tau for depth 1000 compared to the other two depths. The possible cause of such changes could be the low effort relevant documents or the number of relevant documents due to deeper evaluation.

Next, let's analyze the results for TREC-2001 to examine the change in correlation coefficient with the depth of evaluation. It can be observed that all the features have caused an impact on the system rankings due to low effort relevance judgments, for both P@*k* and AP@*k* metrics. All the effort features have Kendall's tau values below 0.8. Despite the variation in system rankings due to effort, it will be interesting to analyze if the correlation coefficient becomes stronger or weaker with the depth of evaluation.

The features LIX, LIXsent, nimg and ntab have all caused the correlation coefficient to increase with depth of evaluation. The trend is same for both P@*k* and AP@*k*. Such trend indicates low effort relevance judgments may have a lesser influence on the system rankings with deeper evaluation. Also, evaluating with metric AP@*k* has an increasing correlation coefficient with deeper evaluation for AllFeature. However, evaluation of AllFeature using metric P@*k* displays decreasing correlation coefficient values with increased evaluation depth. The same observations appear for nchar effort feature when assessed using P@*k* and AP@*k* with the addition to nword feature for AP@*k*.

When assessed using P@*k*, the ARI and CLI features both produce consistent correlation coefficient despite the change in evaluation depth. The results indicate the system rankings are consistent or rather the effectiveness of the systems changed consistently among the different depth of evaluation such that the correlation coefficient hardly changed. As for the other features such as nsent and nword evaluated using P@*k*, the results show varying correlation coefficient with the increase in depth while the ARI,

CLI and nsent features evaluated using AP@*k* also shows the same trend. The Kendall's tau value is weaker for depth 100 compared to depth 10 but Kendall's tau values for depth 1000 is always stronger than that of depth 10.

Evaluating at depth 100 using low effort relevance judgment appears to have a larger influence on the system rankings. Most likely the change in relevant documents between ranks 11 to 100 using eqrel have caused this alteration in the system rankings, and thus lowering Kendall's tau correlation coefficient.

Although the results from both the test collections do not show exact replication, it confirms influences of low effort relevance judgments to system rankings for different depth of evaluation. More times, the depth of 10 has weaker correlation coefficient than the remaining two depths experimented. The results reveal that system evaluation using low effort relevance judgments may have a lesser impact on the system rankings for deeper evaluation depth.

In the past study (Verma et al., 2016), Kendall's tau correlation coefficient for system rankings generated from P@10 original and low effort relevance judgments show tau values between 0.53 and 0.71 for TREC Web Track Ad hoc task 2012-2014 test collections. Also, only the correlation coefficient for top 10 systems was considered. When considering all systems for evaluation, the results from this study do indicate some similarities and differences with the past study for various features. It is also important to highlight that Verma et al. (2016) used a linear regression model to create the low effort relevance judgment. The current study, however, uses a boxplot approach and a standardized grading for readability features to determine the low effort relevant documents. Nonetheless, this study explored low effort relevance judgment with regards to depth of evaluation and highlighted a stronger influence of effort in shallow evaluation compared to deeper evaluation.

Analyses of Kendall's tau results have demonstrated the variation in system rankings due to effort for different depth of evaluation. It cannot be shunned that the change in system rankings could have been due to the number of relevant documents in the relevance judgments. A total number of relevant documents influence the performance of the systems (Voorhees, 2001) and thus their ranking as well. Nevertheless, it is also possible that the number of relevant documents is not affecting the evaluation inconsistencies (Jones et al., 2014). An analysis is necessary to reestablish the cause of changes in correlation coefficient due to evaluation depth. And a further question, is the system ranking change due to low effort relevant documents or just the number of relevant documents in the eqrel?

The percentage of agreement of relevant documents between qrel and eqrel can indicate the proportion of relevant documents in eqrel. The percentage of agreement between the qrel and eqrel is measured separately for each feature. Figure 3.7 shows the steps involved in computing the relevant documents agreement between the original and the low effort relevance judgments.

Firstly, the top *k* documents from the qrel are retrieved for each topic per system. Then identify the unique, relevant documents. The unique relevant documents constitute the top *k* relevant documents for qrel, #rel_ori_qrel_top*k*. Using the top *k* relevant documents of original relevance judgments, identify low effort relevant documents for each effort feature respectively. For each feature, count the number of low effort relevant documents from top *k* from the respective eqrel. The number represents the top *k* relevant documents for the feature, #rel_featureS_eqrel_top*k*. The percentage of relevant documents' agreement is then calculated for each feature using formula #rel_featureS_eqrel_top*k* divided by #rel_ori_qrel_top*k*. The *k* in the top *k* takes values 10, 100 or 1000. The value of *k* should be the same for both the qrel and eqrel.

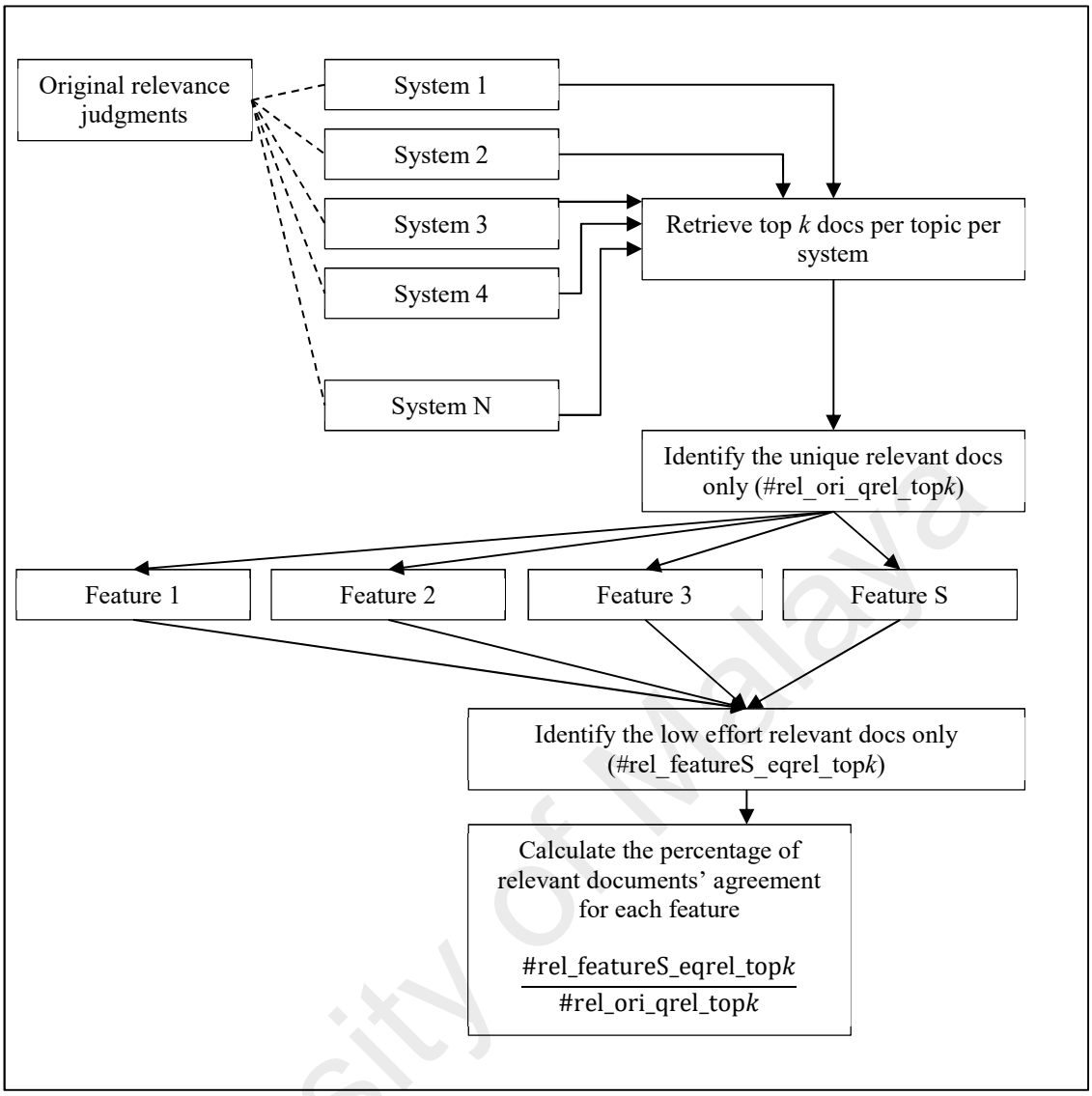

Figure 3.7: Steps for computing agreement of relevant documents between original and low effort relevance judgments

Variations in the percentage of agreement could only happen due to the relevant documents. The reason for this is that low effort relevant documents in the eqrel originated from the qrel relevant documents. Meanwhile, those relevant documents graded as high effort become non-relevant in the eqrel. Hence they would not have a relevancy match with the qrel. If the percentage of agreement is high, there are more low effort relevant documents that match the existing relevant documents from qrel. If the

percentage of agreement is low, there are lesser low effort relevant documents from eqrel that matches the existing relevant documents from qrel.

The analysis determines if the change in the correlation coefficient (due to system ranking changes using eqrel) with regards to evaluation depth is due to the influence of low effort relevant documents instead of just the number of relevant documents. If Kendall's tau value increases with the percentage of agreement, the change in tau could have occurred due to the increase in the number of relevant documents in the eqrel. In contrast, if Kendall's tau value increases while the percentage of agreement remains same, the number of relevant documents is unlikely the cause. Similarly, if the percentage of agreement increases but Kendall's tau value remains consistent, the effect is not due to the increase in the number of relevant documents.

The results of the analysis are shown in Figure 3.8 for TREC-9 and Figure 3.9 for TREC-2001 respectively. The x-axis represents Kendall's tau values, and the y-axis represents the percentage of agreement (relevant documents). The graphs are divided into two rows; the top row is for metric P@*k* and the bottom row for AP@*k*. The columns represent each effort feature experimented and are labeled accordingly along the top of each row.

From Figure 3.8, features ARI, CLI, LIX, nchar, nimg, nsent, and nword have a similar tabulation of plots for P@*k*. The position of the plots appears almost horizontal and close to each other. Horizontal positioning implies minimal changes in the percentage of agreement that causes shifts in Kendall's tau value. It means the correlation coefficient of the system rankings is changing despite the constant percentage of relevant documents. The same is seen for AP@*k* for features ARI, CLI, LIX, LIXsent, and nimg. Therefore, the change in the system rankings is not due to number of relevant documents.

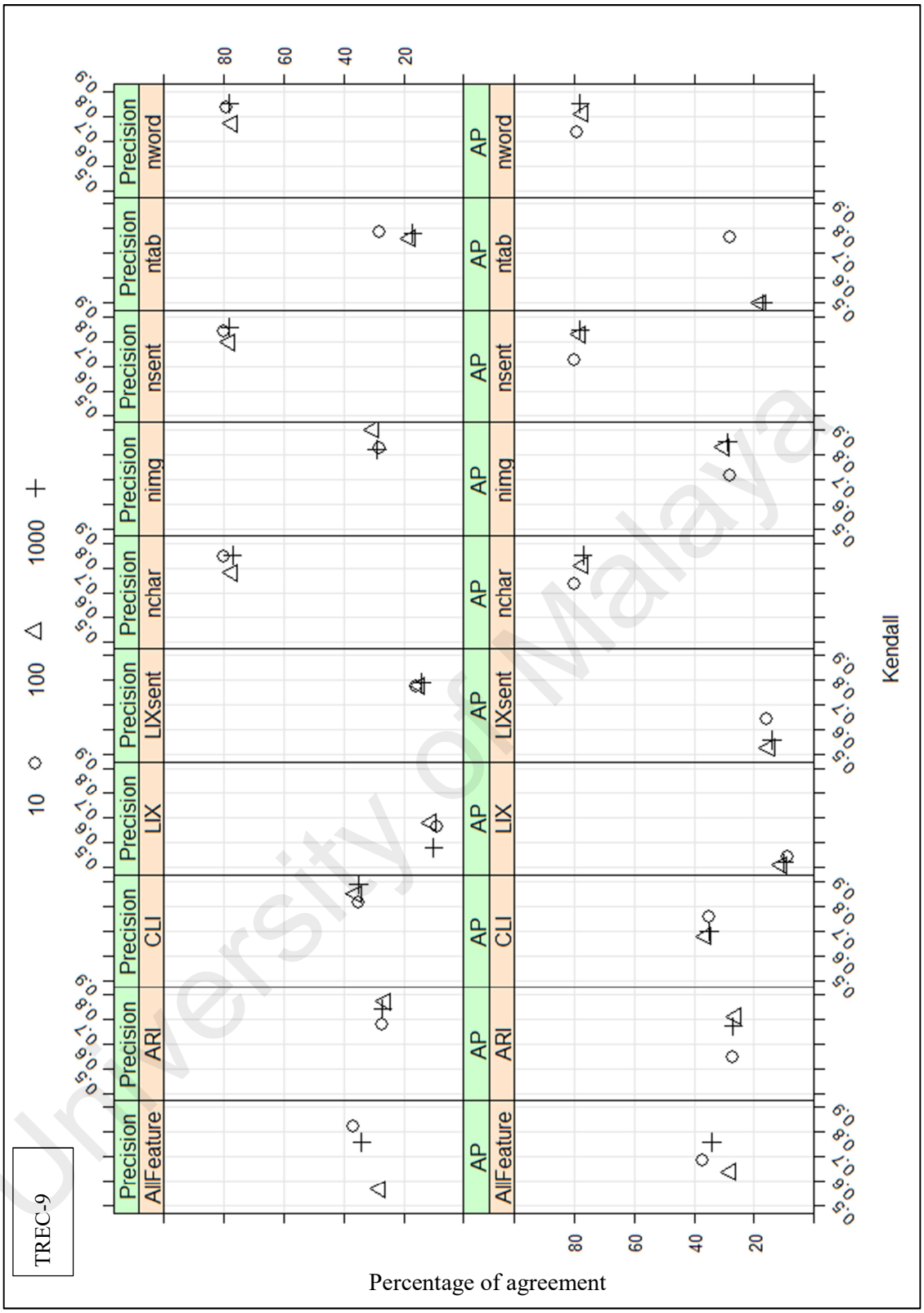

Figure 3.8: TREC-9 ─ Percentage of relevant documents agreement versus Kendall's tau correlation coefficient for different depth of evaluation.

The features nchar, nsent, and nword for AP@*k* have a decreasing percentage of agreement for deeper evaluation compared to shallow depth. However, the declining percentage of agreement results in stronger Kendall's tau value. Decreasing percentage of agreement means there are lesser relevant document matches between eqrel and qrel but still yields strong correlation coefficient. The analysis suggests the number of relevant documents is not causing the change in system rankings. However, due to a lower percentage of agreement, the effectiveness scores could be much lower for the systems as well. The change in effectiveness could be due to small numbers of relevant documents causing the retrieval measures to be unstable (Voorhees, 2001). Nonetheless, notice that these features actually have approximately 80% of agreement with qrel. Also, deeper depth with a lower percentage of agreement but stronger Kendall's tau. Hence, the system ranking change is likely due to the low effort relevant documents.

The LIXsent feature for P@*k* has an almost vertical plot. There are small changes in the percentage of agreement, but Kendall's tau value remains consistent. In fact, the percentage of agreement decreases slightly with the depth of evaluation. It could mean the similarity of ranks between the qrel and eqrel did not change with the depth. Despite variations in the number of relevant documents, the system rankings are not impacted. In this case, the change in system ranking is due to low effort relevant documents instead of just the number of relevant documents.

As for ntab feature, for both P@*k* and AP@*k*, a higher percentage of agreement yields better Kendall's tau value. In fact, the shallow depth evaluation appears to have a better percentage of agreement compared to deeper evaluations. Such scenario could indicate the number of relevant documents is the cause positive change in system rankings. AllFeature has a similar observation. Although the changes in Kendall's tau appears to have occurred due to the number of relevant documents, it is important to note, the

grading of AllFeature was contributed by all the individual effort features. Almost all the features, except ntab, demonstrated the effect of low effort relevant documents to Kendall's tau. Hence, it is acceptable to state the change in system rankings for the AllFeature is also due to the low effort relevant documents and not just the number of relevant documents.

Let's analyze the graph for TREC-2001 (Figure 3.9). It is plotted the same way as Figure 3.8 in regards to the axes, metrics, and features. The results show most features (AllFeature, LIX, LIXsent, nchar, nsent, ntab, and nword for P@*k*, and AllFeature, ARI, CLI, LIX, LIXsent, nchar, nsent, ntab and nword for AP@*k*) have their plots positioned horizontally. The horizontal positioning means small changes in the percentage of agreement while Kendall's tau correlation coefficient have significant changes. As stated earlier, such results did not occur due to the number of relevant documents but most likely due to the low effort relevant documents. Similarly, low effort relevant documents contribute to the observations in P@*k* for features ARI and CLI. These features have their plots positioned vertically. Even though the percentage of agreement decreases with the depth of evaluation, Kendall's tau value remains consistent across the depths. Therefore, the number of relevant document did not cause the changes in system ranking.

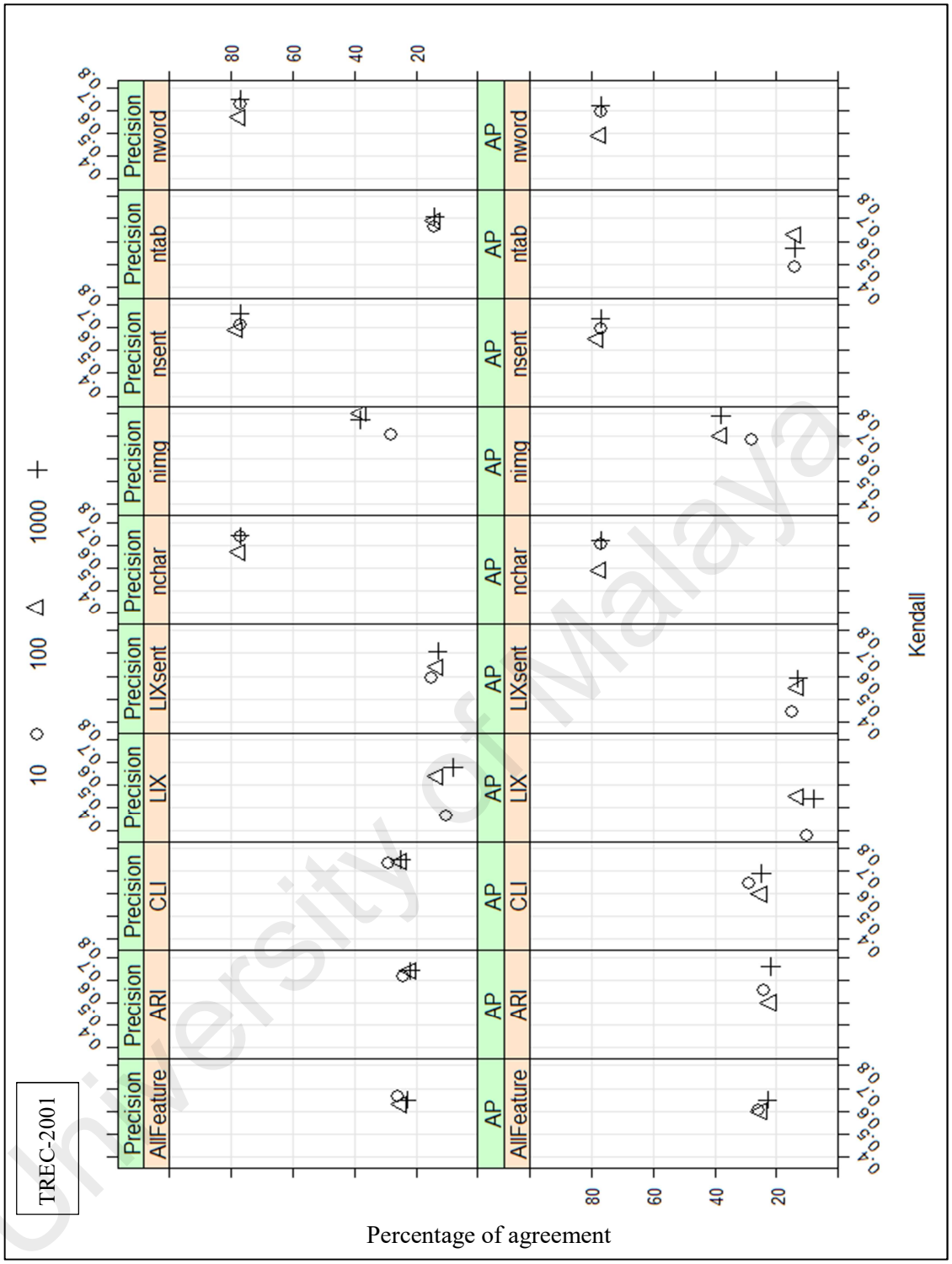

Figure 3.9: TREC-2001 ─ Percentage of relevant documents agreement versus Kendall's tau correlation coefficient for different depth of evaluation.

The only feature that could suggest the role of relevant documents' number to the variation in Kendall's tau is the nimg feature. The percentage of agreement increases with the depth, which could mean more relevant documents in eqrel. Also, these higher percentage of agreement appears to have a positive influence to Kendall's tau correlation coefficient. However, the AllFeature has its plots positioned horizontally. It is different compared to TREC-9 observation. Horizontal positioning is a clear indication that the number of relevant documents is not the cause of changes in system ranking. The analysis directly demonstrates the system rankings change is due to low effort relevant document.

The experimental results reveal that low effort relevance judgments cause changes to the system rankings at different depth of evaluation. Deeper depth of evaluation also tends to have a lesser impact on the system rankings due to low effort relevance judgments compared to shallow depth evaluation. The analysis confirmed the correlation coefficient changes were caused by low effort relevant document instead of just the number of relevant documents in the eqrel. Sometimes, a small number of relevant documents produced better or consistent Kendall's tau. The position of the low effort relevant document could have impacted the system rankings. Another possibility is the magnitude of difference in the effectiveness scores. Due to lower numbers of relevant documents in the eqrel, the effectiveness scores of the systems are lower than those from original qrel. Therefore, the difference in scores could have changed consistently, such that the system ranks did not vary between using qrel and eqrel.

3.4.2 Correlation Coefficient Evaluation for Groups of Systems with Low Effort Relevance Judgments

The variation in system rankings due to low effort relevance judgments is evaluated for groups of systems and covers RQ2. The groupings are top 10 systems, and bottom 10

systems, that were identified through the MAP scores for specific depth from the evaluation using original relevance judgments. The systems were ranked separately for each depth of evaluation.

Figure 3.10 shows Kendall's tau correlation coefficient for the top 10 and bottom 10 groups of systems for the various effort features and depth of evaluation. The plot for the entire set of systems is also included. The x-axis represents Kendall's tau correlation coefficient values, and the y-axis represents the features. Each row consists of effectiveness metrics at a specific depth of evaluation. The figure includes graphs for both test collections. The detailed numbers used for plotting Figure 3.10 is available in APPENDIX P (TREC-9) and APPENDIX Q (TREC-2001).

The plots for TREC-9 show that correlation coefficient of top 10 systems is lower than that of bottom 10 systems for all features except LIX and CLI for P@10, and CLI for P@100. The results indicate top performing systems are more susceptible to low effort relevance judgments. With the increase in evaluation depth to 100 for metric P@*k*, the correlation coefficient values for the top 10 systems tend to increase but insufficient to have a strong correlation coefficient value. When the depth is increased further to 1000 for top 10 systems using metric P@*k*, Kendall's tau value decreases largely. However, the results for top 10 systems from metric AP@*k* shows mostly a dip in Kendall's tau with depth 100 and a rise for depth 1000.

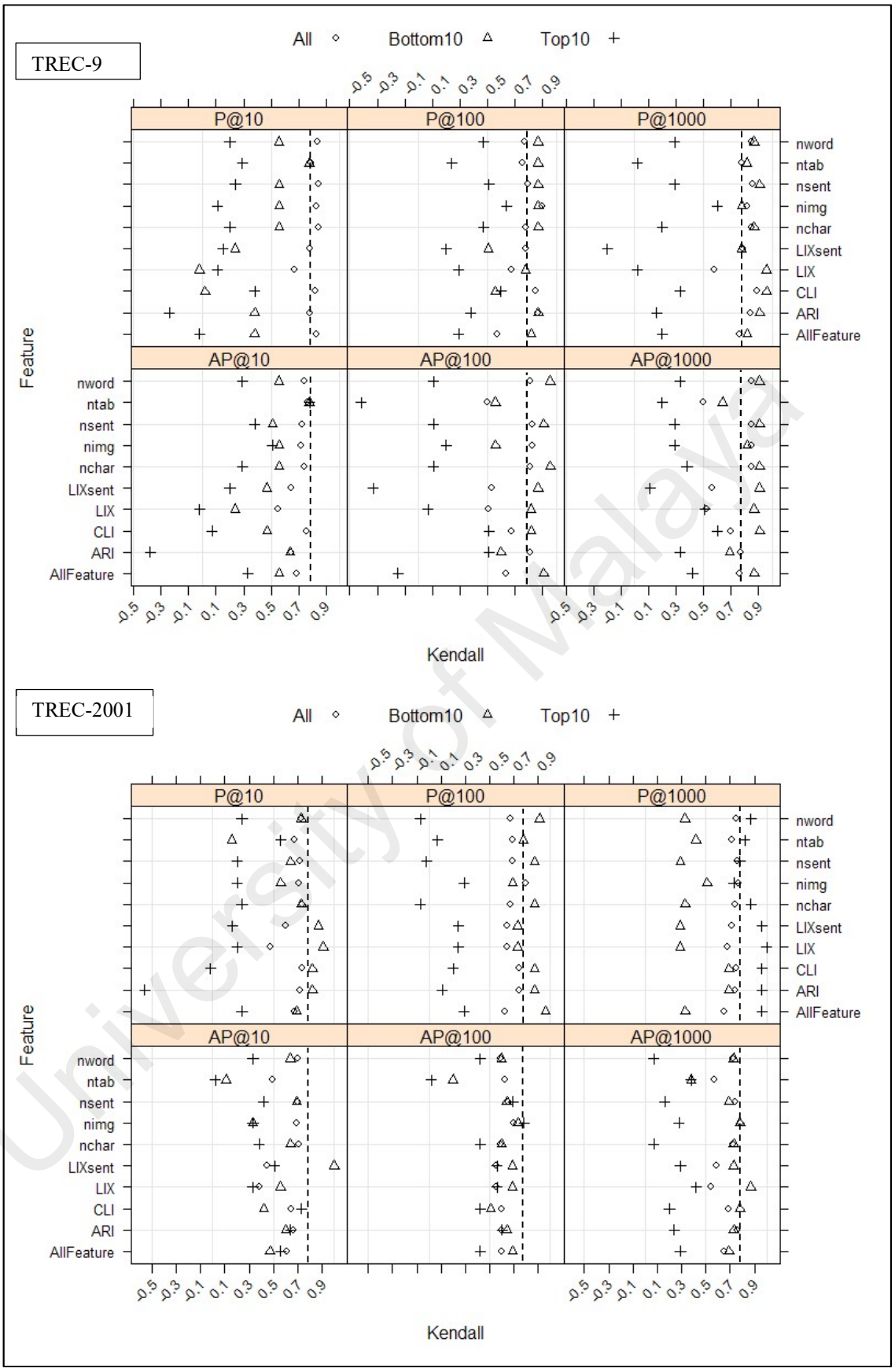

Figure 3.10: Correlation coefficient for groups of systems for the various effort features

In contrast, the bottom 10 systems are becoming less vulnerable with deeper evaluation since the correlation coefficient increases. The observations are parallel for both metrics. A possible reason for the outcome is that the bottom 10 systems had originally retrieved relevant documents that were low effort. When these systems were evaluated with the low effort relevance judgments, their performance was not largely impacted. Hence, Kendall's tau is mostly above 0.8 for P@100, P@1000, AP@100, and AP@1000. However, for P@10 and AP@10, Kendall's tau values are weak to moderate. Nonetheless, the bottom 10 systems score better tau than the top 10 systems for P@10 and AP@10 metrics.

The plots for TREC-2001 show varying results for the groups of systems and metrics combination. For most of the evaluation metrics, the top 10 systems score a lower correlation coefficient value compared to bottom 10. Even in the past study, poor Kendall's tau correlation coefficient revealed the top 10 system have a difference in their performance when measured using low effort P@10 (Verma et al., 2016). One clear exception is for evaluation metrics P@1000, whereby the top 10 systems have strong correlation coefficient, in addition, to have scored better than bottom 10 systems and overall systems. Besides, there are few other features such as CLI, ARI and AllFeature that show top 10 systems have better correlation coefficient than the bottom 10 systems for AP@10.

The analysis reveals the top performing systems have focused on retrieving relevant documents but not documents that could satisfy real users because they are likely high effort documents. Although the systems measure top performance among the rest of the systems using original relevance judgment, an actual user could face difficulties in consuming the information from their retrieved documents. As mentioned in a previous

study, high-effort documents are harder to consume. Hence, it is less likely for the users to read it (Yilmaz et al., 2014).

Nonetheless, the bottom 10 systems are not affected by the low effort relevance judgments since this group achieves strong correlation coefficients. The poor performing systems may, in fact, be retrieving low effort documents suitable for the utility of actual user. However, the change in observation between bottom 10 and top 10 systems in TREC-2001 P@1000 and few others, raises a concern for the swap.

If the correlation coefficient of the top 10 systems could be better, it means the rankings of these systems are very close to the original system ranks. For that to happen, the performance of these systems is sustained relative to others. The relevant documents translate to these performances. Therefore, the top 10 systems should have higher matches of the relevant documents with the original relevance judgments compared to those from bottom 10 systems.

The percentage of document matches between qrel and eqrel is derived to determine the role of relevant documents between the top 10 and bottom 10 systems, especially the TREC-2001 P@1000. The percentage of average document matches is computed by dividing the total matching relevant documents with a total number of documents for the specific system groups and multiplying them by 100%.

A high percentage of matches would mean there are more relevant documents in the eqrel for the group of systems. A low percentage of matches means there are less relevant documents in the eqrel. If the percentage of average document matches is high and yet the Kendall's tau value is weak, the underlying reason could be the position of the ranked relevant documents that are now considered as a low effort. However, the P@*k* is not influenced by the rank or position of the relevant documents. It only uses the total relevant

documents against the total document retrieved at specific cut-off *k*. However, the AP@*k* scores will be influenced by the rank or position of the relevant documents, which in turn could cause the effectiveness scores to change and thus their rankings as well.

Figure 3.11 shows the percentage of average documents matches (relevant) for each feature for the top 10 and bottom 10 system groups. The x-axis represents the feature while the y-axis represents the average matched relevant documents in percentage. The average matched relevant documents is shown for each feature based on the depth of evaluation. The dotted lines within the graph are placed as a guideline to visualize the placement of plots in the chart.

Let's first look at the plots for TREC-2001 for evaluation depth of 1000. The top 10 systems had better correlation coefficient than the bottom 10 systems for TREC-2001 and P@1000 metrics. Since the top 10 systems had better correlation coefficient, it is expected that these systems would have a higher percentage of average matched relevant documents compared to the bottom 10 systems.

From the figure (observe the + symbols), it appears that top 10 systems have almost equal percentage of matched relevant documents with the bottom 10 systems. To be exact, features nsent, nword, and nchar for top 10 systems have 4% matches each, while bottom 10 systems have 3% matches each. The LIX feature for top 10 systems has 0.38% matches while the bottom 10 systems have 1% matches. The remaining features have exactly the same percentage of matches. Therefore, it appears that the number of relevant documents is not the cause of the correlation coefficient change or the swap between the top 10 and bottom 10 systems. Due to a small percentage of relevant document matches, the total number of relevant documents could be small as well and thus causing variability in system ranks. Small numbers of relevant documents, fewer than 5 in each query has caused instability in the system rankings (Voorhees, 2001).

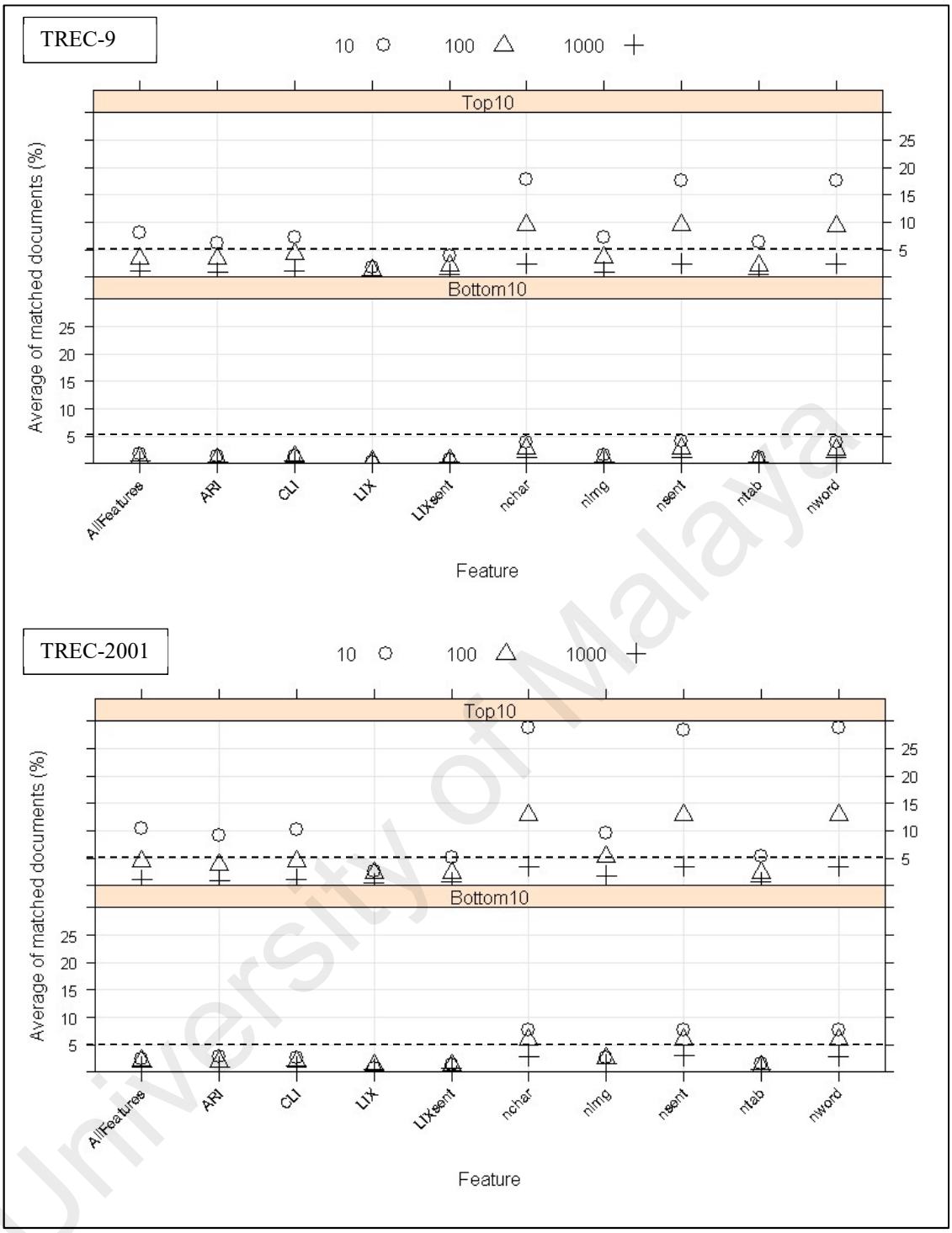

Figure 3.11: Average of matched relevant documents (%) between eqrel and qrel for various effort features.

As for the other features in TREC-2001 evaluation such as CLI, ARI and AllFeature for AP@100 also resulted in top 10 systems having better correlation coefficient than the bottom 10 systems. Could this also be due to instability in the system rankings due to low

numbers of relevant document matches? From Figure 3.11 (observe the triangle symbols), it can be seen that top 10 systems have a higher percentage of relevant document matches compared to the bottom 10 systems. Perhaps the number of relevant documents is the cause of better correlation coefficient for top 10 systems compared to bottom 10 systems.

However, notice that for both test collections, the top 10 systems have a higher percentage of relevant document matches than the bottom 10 systems. These numbers should have caused the top 10 systems to have better correlation coefficient than the bottom 10 systems, but that is not the case. In fact, most times the bottom 10 systems had stronger correlation coefficient compared to the top 10 systems. In such cases, the position of the low effort relevant documents in the ranked list could have impacted the system rankings of groups of systems. Another possibility is the consistent change in effectiveness scores in the bottom 10 systems using eqrel such that the system ranks remained unchanged.

3.4.3 Evaluation of Retrieval Systems with Reduced Topic Size Using Low Effort Relevance Judgments

The study then attempts to measure the effectiveness of the systems using low effort relevance judgments with reduced topic sizes and consequently answering to RQ3. A total of 50 topics were available for evaluation for both test collections. The system rankings from original relevance judgment measured for 50 topics will be the baseline for comparison of this experimentation. Randomly selected topics will be used to measure the effectiveness of the systems and rank them. These ranks will then be correlated with the baseline. The approach taken to evaluate the retrieval systems with reduced topic sizes was detailed in Section 3.3.4.

The results of the experiment to reduce the topic size while measuring the effectiveness of the systems using low effort relevance judgments is plotted in Figure 3.12 and Figure 3.13 for TREC-9 P@*k* and AP@*k* respectively. The x-axis represents the topic size, and the y-axis represents Kendall's tau values. Each panel within the graph represent the features. The symbol of the plot is matched to assist in understanding the chart. For example (refer to the legend), the triangle is used for evaluation depth of 10. The triangle is not filled for eqrel and filled for qrel. The colors also match. Similar pairing is done for other depth of evaluation. The dotted horizontal lines in the graph are placed as a guideline to easily visualize the placement of the plots. The detailed numbers used for plotting Figure 3.12 and Figure 3.13 are available in APPENDIX R and APPENDIX S respectively.

The Kendall's tau measures the ordering of the system ranks between the baseline and reduced topic size. For the system ranks of reduced topic size using qrel, a strong Kendall's tau would indicate the representation of the reduced topic to the full topic. However, a strong Kendall's tau between baseline and reduced topic using eqrel, would not necessarily mean a representation of the full topic set. But, it could mean there is no impact on the system rankings due to low effort relevance as stated in earlier sections of this study. From the earlier experimental results, it is already known that low effort relevance judgments have an impact on the system rankings at different depth. Therefore, instead of interpreting Kendall's tau as the impact of low effort relevance judgment, it will be used as a predictivity of the full topic system ranking. Also, an equal or better eqrel correlation coefficient than the qrel should be observed as a suitable prediction of the proposed approach with the reduced topics. Because now the reduced topics from eqrel are able to predict the full topics as good or better than the reduced topics from qrel.

Based on the observations for P@*k*, topic sizes 5 to 20 for a depth of 10 appears to have about seven features that have better Kendall's tau values than the ones using qrel. These features are AllFeature, ARI, CLI, nchar, nsent, nword, and nimg. Out of these features, some have achieved strong correlation coefficient values 0.8; the nword (size 15), nchar (size 10 and 20), and nsent (size 10 and 20). As for depth 100, nimg (size 20), and CLI (size 20) features have better Kendall's tau values using eqrel compared to qrel. Evaluating at depth 1000, ntab (size 20), LIXsent (size 5), nchar (size 5), nimg (size 15 and 20), CLI (size 10 and 20) and AllFeature (size 10) have equally good or better correlation coefficient than using qrel. For depth 100 and 1000, none of the correlation coefficient values were above 0.8.

As the topic size is increased from 25 to 40 for depth of 10, only ARI and CLI effort features have Kendall's tau values as good as those obtained from qrel. There were no instances when Kendall's tau values from eqrel are better than qrel for topic sizes 25 to 40. However, some features using eqrel appear to have strong correlation coefficient. They are ntab (size 30), nword (size 30 and 40), nchar (size 30 and 35), nsent (size 35 and 40), CLI (size 30 and 40), and AllFeature (size 35 and 40). When evaluating at depth 100, only the ARI feature shows better Kendall's tau value than qrel for topic size 30. But nchar (size 35), nimg (size 30 and 35), nsent (size 40), CLI (size 35 and 40), and ARI (size 35 and 40) features have tau values above 0.8. As for depth 1000, none of the eqrel Kendall's tau values are better than qrel, but some features have produced values above 0.8. These features are ntab (size 30), nimg (size 40), and CLI (size 35).

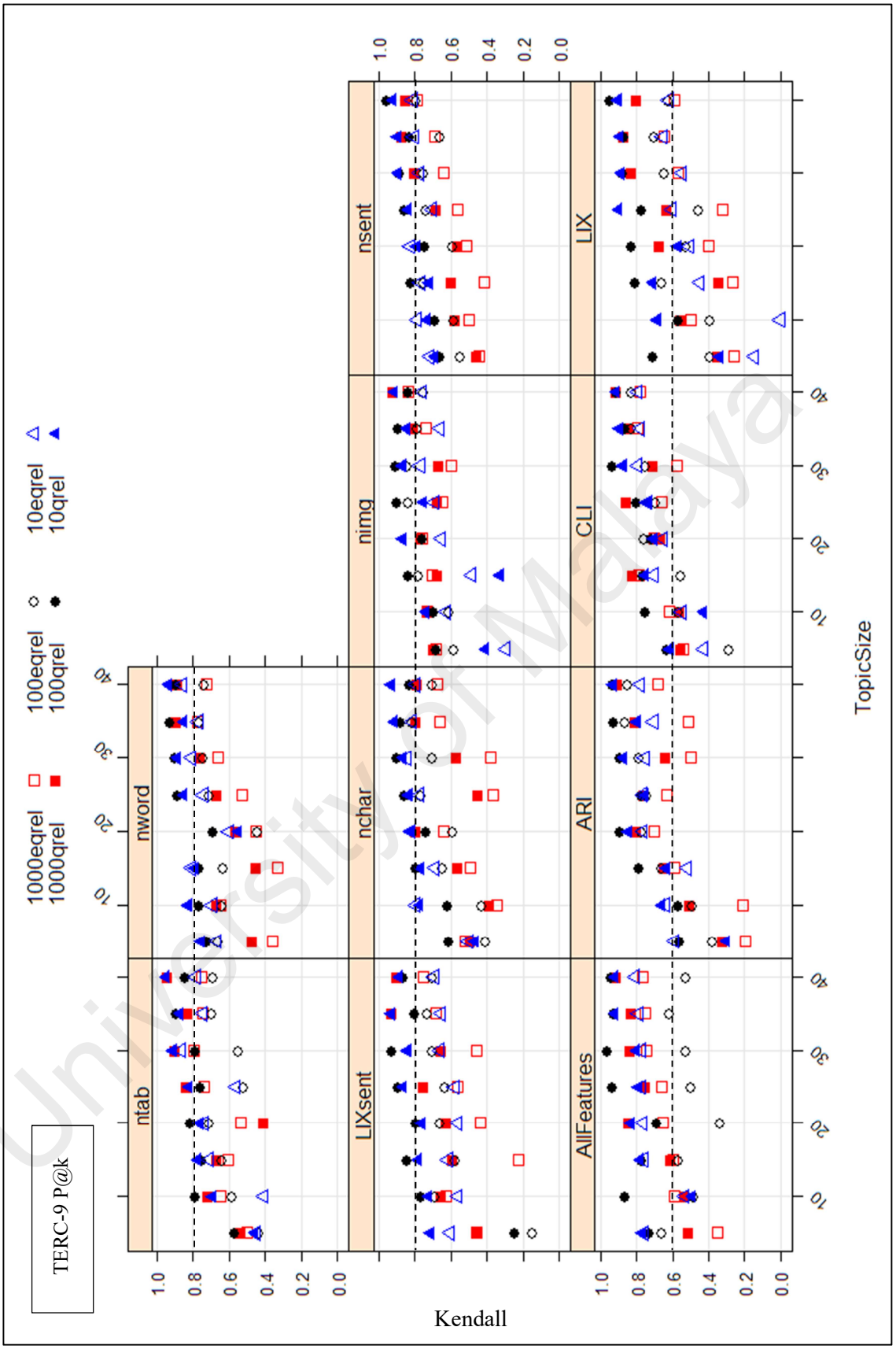

Figure 3.12: TREC-9 ─ Kendall's tau correlation coefficient for reduced topic sizes using metrics P@*k*

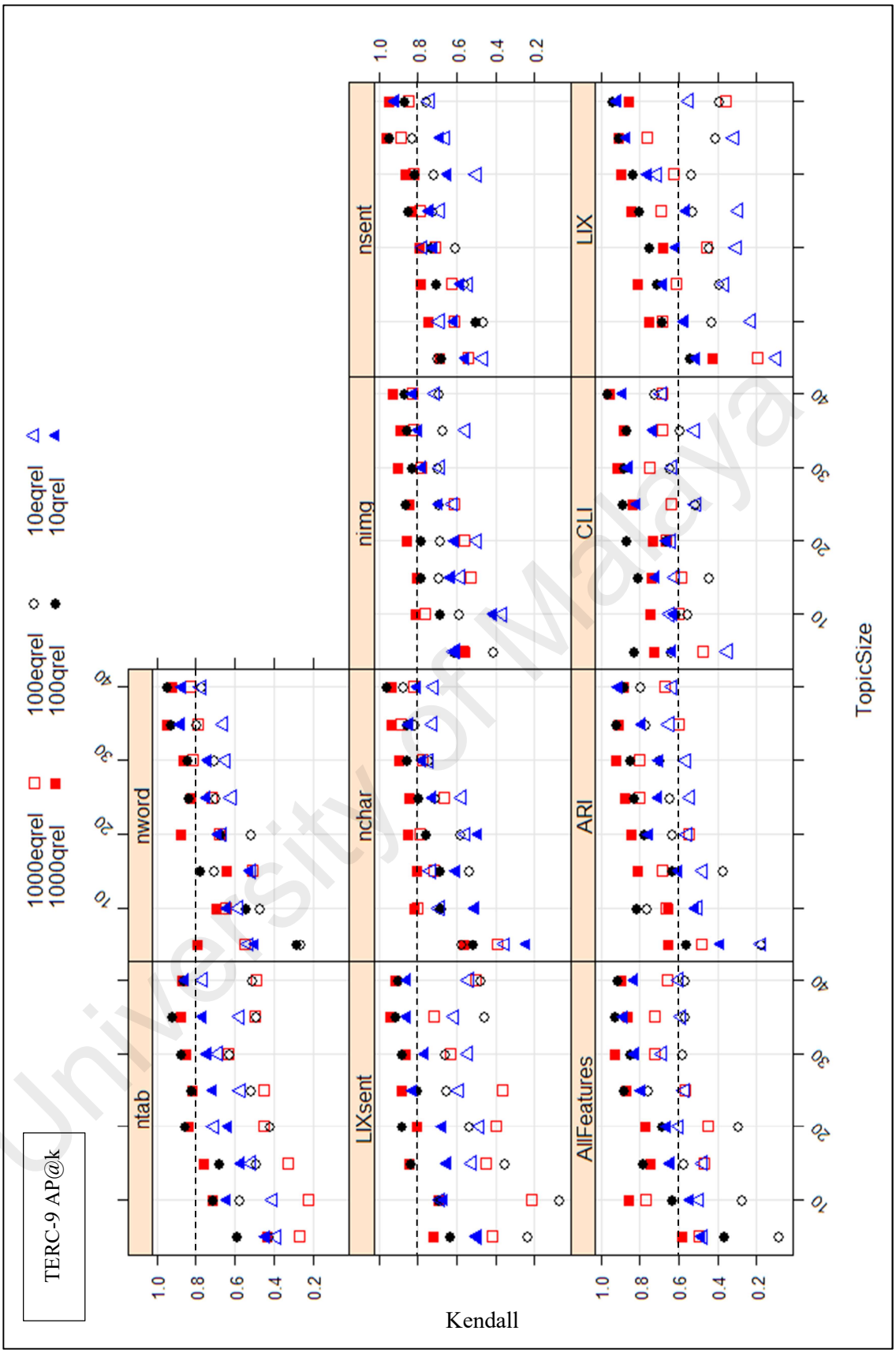

Figure 3.13: TREC-9 ─ Kendall's tau correlation coefficient for reduced topic sizes using metrics AP@*k*

From the analysis for P@*k* using shallow depth of evaluation, topic sizes 10 to 20 appear to be suitable alternatives in evaluating retrieval systems for features nword, nchar, and nsent. Although not always Kendall's tau value is above the one produced by qrel, these topic sizes are capable of producing sufficiently good system ranking with the original 50 topics. No suitable topic sizes for deeper evaluation depth (100 and 1000) and small topic sizes. For larger topic sizes (30 to 40), nword is suitable for shallow depth of evaluation, and nimg and ARI are suitable for deep depth of evaluation. Meanwhile, ntab, nchar, nsent and CLI are appropriate for both shallow and deep depth of evaluation for larger topic sizes.

As for the observations of small topic sizes (5 to 20) in TREC-9 AP@*k* with evaluation depth of 10, a few features appear to have better Kendall's tau values using eqrel compared to using qrel. These features are ntab (size 20), nword (size 5), nchar (size 15 and 20), nimg (size 5), nsent (size 10), LIXsent (size 10 and 20) and CLI (size 10). None of these features using eqrel had a correlation coefficient value above 0.8 for any topic sizes. With evaluation depth of 100 and small topic sizes, nsent (size 5) feature has better tau using eqrel than using qrel. However, nchar (size 10) feature using qrel has Kendall's tau value above 0.8. As for the deeper evaluation of 1000 and small topic sizes, features nimg and ARI have better tau values using eqrel compared to using qrel. Nonetheless, the nchar (size 10) feature produced tau values above 0.8.

As for evaluations for various depths using larger topic sizes (25 to 40) using TREC-9 AP@*k*, no features using eqrel score better correlation coefficient than using qrel. However, there are some features with larger topic sizes that have strong correlation coefficient. For evaluation depth 100, features nchar (size 35 and 40), and nsent (size 35), and for evaluation depth 1000, features nword (size 30 and 40), nsent (size 30, 35 and 40), nimg (size 35 and 40), and nchar (size 35 and 40) have strong correlation coefficient.

Also, note there is no feature having strong correlation coefficient for shallow depth of evaluation using larger topic sizes.

From the analysis for AP@*k* using shallow depth of evaluation, there are no suitable reduced topic sizes in evaluating retrieval systems. However, for small topic sizes and deep depth of evaluation, the nchar feature appears to be a good alternative to using 50 topics. However, larger topic sizes (25 to 40) using nchar, nsent, nword and nimg features could be a good alternative in evaluating retrieval systems for deeper depth of evaluation.

Overall, the nchar feature appears to be effective for retrieval system evaluation using reduced topic sizes for both shallow or deep evaluation depth. This feature seems to have strong Kendall's tau correlation coefficient for small and big topic sizes when evaluated using P@*k* and AP@*k* metrics.

The same experiment was conducted for TREC-2001 test collection. The results of the experimentation are plotted in the graphs in Figure 3.14 and Figure 3.15 for P@*k* and AP@*k* respectively. The detailed numbers used for plotting Figure 3.14 and Figure 3.15 are available in APPENDIX T and APPENDIX U respectively. Based on the observations, TREC-2001 P@*k* evaluations for small or big topic sizes show no Kendall's tau values using eqrel that is better than using qrel for evaluation depth 10. No features evaluated using eqrel achieved strong correlation coefficient with depth 10 as well. For evaluation depth 100, features nchar and nword have tau values using eqrel better than qrel, but none have achieved strong correlation coefficient for small topic size. As for evaluation depth 1000, features nword and nsent for small topic size, and nsent for big topic size have Kendall's tau values using eqrel stronger than that produced by qrel. However, these features were able to achieve moderate correlation coefficient ranging between 0.4 and 0.8 tau values.

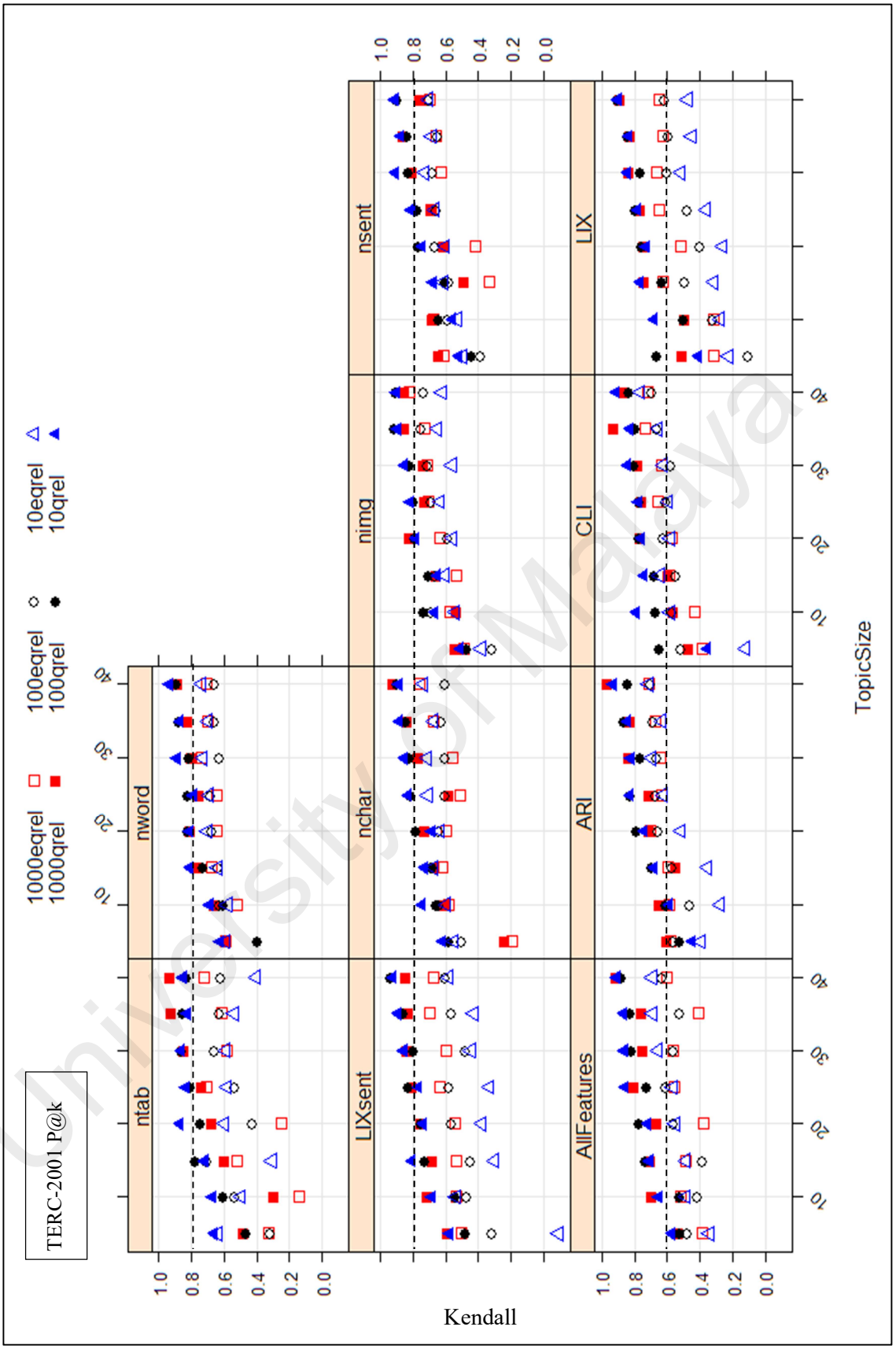

Figure 3.14: TREC-2001 ─ Kendall's tau correlation coefficient for reduced topic sizes using metrics P@*k*

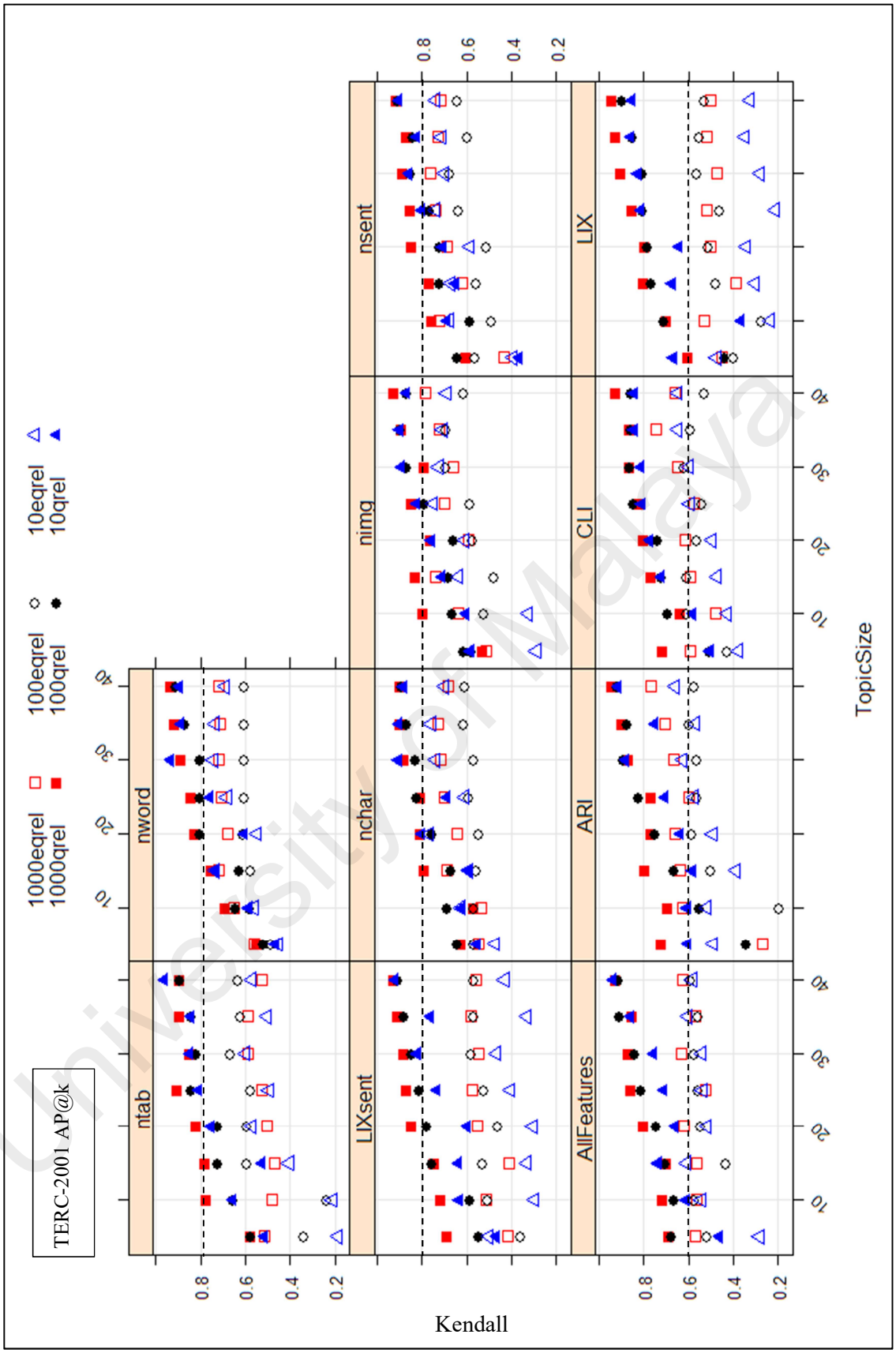

Figure 3.15: TREC-2001 ─ Kendall's tau correlation coefficient for reduced topic sizes using metrics AP@*k*

Evaluating the TREC-2001 with AP@*k* also has a similar outcome with that produced using P@*k* for this test collection. Only a few features achieved better tau values using eqrel compared to qrel. These features are nword, nchar and nsent for evaluation depth 10, and nword for evaluation depth 1000 using only small topic sizes. No features had produced a tau value above 0.8 for any feature, depth of evaluation and topic size combination.

It appears that for this test collection, evaluation of retrieval systems using reduced topic sizes is not suitable since no features were able to achieve a strong correlation coefficient. However, there are many instances whereby the retrieval system evaluation using eqrel and qrel have both produced moderate Kendall's tau values. The tau values were achievable with small topic sizes (10 to 15) for all depth of evaluation, and both metrics experimented. However, bigger topic sizes tend to have Kendall's tau closer to 0.8. Despite the variation in retrieval system evaluation due to topic sizes, the results produced by both metrics are comparable, as stated by Guiver et al. (2009). Recall that different topics of the same size could produce different results while some topics may be more suitable for evaluating the retrieval systems (Guiver et al., 2009; Voorhees & Buckley, 2002).

It is important to highlight at this point that for each feature, different combinations of topics for each topic set were used. It is unlikely that any two topics set were same. The difference in topics could have been a major contributor to variations in the system ranking evaluations.

3.5 Summary

System-oriented or test collection based evaluation has always prioritized relevance but the effort in consuming the relevant documents is equally important for measuring the effectiveness of retrieval systems. The OBJ1 is achieved by systematically generating and evaluating retrieval systems with low effort relevance judgments at different depth of evaluation. The systematic approach in evaluating retrieval systems using low effort relevance judgments has a stronger influence on shallow depth of evaluation compared to deeper depth. It is proved that difference in the system ranking is not due to just the number of relevant documents. At times, low numbers of relevant documents have shown strong correlation coefficient results compared to the higher number of relevant documents. It can be concluded that low effort relevance judgment is the cause of poor correlation coefficient between the system rankings from original and low effort relevance judgments. The rank position of these low effort relevant documents has an influence on the variation of the system rankings. Therefore, it is crucial to evaluate retrieval systems at shallow depth using low effort relevance judgments. Besides, it is necessary to use low effort relevance judgments in evaluating retrieval systems to determine retrieval systems that satisfy users.

The system rankings of the top performing systems are more vulnerable to the low effort relevance judgments compared to the bottom performing systems. Although the top performing systems have high performance, the documents retrieved requires high effort and less likely consumed by the end user. On a separate note, the influence of low effort relevance judgments decreases with increased depth for bottom performing systems. These systems have been retrieving low effort relevant documents despite being the group of systems with poor performance. Additionally, low numbers of relevant documents caused instability in the system rankings. Even though the top performing systems have

more low effort relevant documents compared to bottom performing systems, the top performing systems lack correlation coefficient with the original system ranks. The position of the low effort relevant documents in the ranked list has caused the change in the system rankings of top performing systems. Retrieval systems that have been capable of retrieving high numbers of relevant documents should focus on retrieving and ranking low effort documents to ensure user satisfaction. Meanwhile, those retrieval systems already retrieving low effort documents should prioritize on increasing the number of low effort documents to fulfill user's need. With that, the OBJ2 is met.

For OBJ3, the prediction of system rankings using reduced topic sizes with low effort relevance judgments is inconclusive due to the effect of topic variation. A limited number of features were predictive of the full topic set but not consistent. Further studies are necessary to obtain better understanding on the effectiveness of using reduced topic size. A best topic set may suggest a clearer understanding of the full topic predictivity using low effort relevance judgment. However, it is not within the scope of this study to explore the best topic set as an option of full topic prediction. Rather, it aimed to understand the effectiveness of evaluating the retrieval systems with reduced topic size using low effort relevance judgment. At this point, it is at least known that reduced topic size is effective in predicting the performance of the retrieval system for specific effort features. Opportunities exist for evaluation of retrieval systems with reduced topic set since certain features have better correlation coefficient when evaluated using reduced topic size from low effort relevance judgments compared to original relevance judgments.

CHAPTER 4: MEASURING THE RELIABILITY OF SYSTEMS RANKING IN INFORMATION RETRIEVAL SYSTEMS EVALUATION

This chapter covers an experimentation measuring the reliability of ranking in information retrieval systems evaluation. The focus of the experimentation is to measure individual system's reliability using its ranks. Section 4.1 details the background, problem statement, research questions and the objectives of this experimentation. Section 4.2 details the literature review on the past studies measuring reliability, the different categories of reliability measurement, the difference between Pearson and intraclass correlation coefficient, and the model and forms of the intraclass correlation coefficient. Next, Section 4.3 details the selection of test collections, the fitting of intraclass correlation coefficient to the experimentation, the proposed method to measure reliability of individual retrieval system rankings, and the evaluation of individual retrieval system rankings' reliability. Then, Section 4.4 details the results of the experimentation and the discussions pertaining the results. Lastly, the summary of the chapter follows.

4.1 Background

Information retrieval systems use a number of topics and document corpus for retrieval of relevant documents. The retrieval is an important aspect of the system such that studies focus on indexing (De Melo & Hose, 2013; Golub et al., 2016; Hiemstra & Graham, 2009; Varathan, Sembok, Abdul Kadir, & Omar, 2014) and query formulation (Bailey et al., 2015; Golub et al., 2016; Hiemstra & Graham, 2009). These retrieved ranked documents together with the relevance judgment allows for evaluation of the retrieval system.

The performance of an information retrieval system can be measured using various metrics. Each metric evaluates the systems differently according to the user model implementation (Moffat et al., 2012). Usually, adaptive user model such as average precision is used to calculate scores for topics before aggregating them into an overall effectiveness score. Based on this effectiveness score, the performance of the system can be evaluated.

The evaluation of the retrieval system is the other important aspect of information retrieval, providing information about the performance of the system relative to others. A system performing well in a test collection does not necessarily perform similarly when another test collection is used (Pavlu, Rajput, Golbus, & Aslam, 2012). In addition, the selection of topics is also important in the quality of system ranking (Hauff et al., 2009; Voorhees & Buckley, 2002). Such that some topics tend to represent the system better than other topic sets (Guiver et al., 2009; Voorhees & Buckley, 2002). Therefore, topic size selection is dependable in the test collection and task (Urbano et al., 2013a). However, it is inevitable that a system would perform differently for different topics or metrics.

Due to various effects in the evaluation of retrieval system, the reliability of the test collections (Urbano et al., 2013a), relevance assessment (Blanco et al., 2013; Ruthven, 2014; Voorhees, 2000), and effectiveness metrics (Baccini, Déjean, Lafage, & Mothe, 2012; Sakai, 2007) were studied.

There exists variation in reliability of test collections. An aspect of test collection is the relevance judgments. The reliability of relevance judgments was measured through the variation in system rankings. Relevance judgments from different expert judges (Voorhees, 2000) or crowdsourced judges (Blanco et al., 2013) did not impact the reliability of the system rankings. The study (Voorhees, 2000) focused on measuring the

reliability of the system rankings using only the MAP metrics, while Blanco et al.'s (2013) study used MAP, NDCG, and P@10. The reliability of judgment was measured using Fleiss' Kappa (Blanco et al., 2013).

The system rankings are in fact translations from the effectiveness metrics scores. Suitable metrics was suggested based on binary and graded relevance (Sakai, 2007), and metrics that represent clusters of similar metrics based on some mathematical properties (Baccini et al., 2012). The underlying user models of the clustered metrics were, however, different. The experimentation (Baccini et al., 2012) used effectiveness scores instead of the translated ranks to identify metrics clusters for retrieval system evaluation. As least one metric from each cluster was suggested as a representative to sufficiently evaluate the retrieval systems. But, the study did not mention which of those metrics are suitable for measuring the reliability of the system rankings.

A reliable system is crucial in satisfying users' need. In evaluating the retrieval systems, different metrics could be based on different user models, fulfilling different user needs (Moffat et al., 2012). Past studies (Blanco et al., 2013; Sakai, 2007; Urbano et al., 2013a; Voorhees, 2000) have measured reliability for a set of systems through experimentations based on the differences in scores or ranks obtained from a test-retest kind of setting. None of them have explored individual system's ranking reliability. Evaluation measuring similarities of a set of system rankings were explored due to variations from relevance judgments, metrics or test collections. When system rankings are evaluated in a set of systems, the results from Kendall's tau only indicates the overall strength of correlation coefficient but not for the individual systems.

As an example, consider a scenario in Table 4.1 where the effectiveness scores are different but not the ranks. The ranks are from different metrics which measure the

systems according to their user model representation. However, a system's reliability can be observed by their consistency in rankings.

Table 4.1: Example of consistency in system ranking

System	Topic	P@5	Rank	AP@5	Rank
Sys A	Topic 1	0.8	1	0.81	1
Sys B		0.6	2	0.45	2
Sys A	Topic 2	0.2	1	0.04	1
Sys B		0	2	0	2
Sys A	Topic 3	0.4	1	0.28	1
Sys B		0.2	2	0.25	2

In the example from Table 4.1, each of the systems has been reliable in their rankings despite their changing effectiveness scores. It is known that effectiveness scores would vary due to the topic. The sustainment of the individual system's performance measured by their ranks provides the confidence a user can gain upon repeated queries. That means a user can be certain that the retrieval system is providing results that are reliable, let's say 80% of the time. However, this reliability measure is not dependent on good or poor performing system but rather its ability to sustain its performance in regards to other systems for different queries.

4.1.1 Problem Statement

Past studies (Blanco et al., 2013; Sakai, 2007; Urbano et al., 2013a; Voorhees, 2000) measured reliability for a set of systems based on the differences in scores or ranks obtained from a test-retest setting. The Kendall's tau indicates the overall strength of correlation coefficient of a set of system rankings but not for the individual systems. At

the time of this study and of best knowledge, previous studies have not explored the reliability of individual system's ranking.

A large number of metrics were experimented to group metrics in terms of their mathematical properties, although the metrics have different user models (Baccini et al., 2012). One metric per cluster was recommended as a representative to sufficiently evaluate the retrieval systems. However, it was not stated which of those metrics are suitable for measuring the reliability of the system rankings.

A reliable system is necessary for satisfying users' need through its consistent performance. A single metric could measure the effectiveness of a system for a specific user model. Although the same system can be scored using different metrics, the reliability of a system ranking from a combination of metrics has not be explored. Evaluation of retrieval systems with a combination of metrics could accomplish multiple user models and satisfy users' information need.

4.1.2 Research Questions

RQ1. How to measure the reliability of individual retrieval system rankings and identify suitable combination of metrics in measuring reliability of individual system rankings?

RQ2. Can reliability measurement of individual retrieval system rankings be generalized to other similar metrics?

RQ3. Which retrieval systems from the original ranks have reliable system rankings?

RQ4. Does the reliability measurement of individual retrieval system rankings represent the original system rankings?

4.1.3 Objectives

This study aims

OBJ1. To propose a method to evaluate the reliability of individual retrieval systems and determine suitable combination of metrics for measuring reliability of individual system rankings.

OBJ2. To extend the generalization of the system ranking reliability to other similar metrics pairs.

OBJ3. To distinguish the original systems with reliable system rankings.

OBJ4. To validate the reliability measurement of individual retrieval system rankings with the original system rankings.

4.2 Literature Review

This section provides information about past studies on measuring reliability, the different reliability testing, the difference between Pearson and intraclass correlation coefficient (ICC), and the models and forms of intraclass correlation coefficient as a measure of reliability.

4.2.1 Previous Related Studies

Information retrieval evaluation has seen many aspects of measuring reliability such as reliability of relevance judgment (Blanco et al., 2013; Ruthven, 2014; Voorhees, 2000), reliability of test collections (Urbano, Marrero, & Martín, 2013b) and the reliability of system rankings (Hosseini, Cox, Milic-Frayling, Shokouhi, & Yilmaz, 2012; Sanderson & Zobel, 2005).

Back in the year 2000, NIST initiated a study to identify the changes in system rankings due to the differences in the relevance judgment by different expert judges. All the NIST assessors had the similar training for the task undertaken, and three different judges judge each topic. Strong Kendall's tau correlation coefficients proved the system rankings of the retrieval systems are reliable despite the differences in the relevance judgments (Voorhees, 2000).

The reliability of relevance judgments was measured for intra-assessor consistency (Ruthven, 2014), and inter-assessor agreement (Blanco et al., 2013). As part of their study, intra-assessor consistency measure was used to examine the frequency of an assessor making the same decision multiple times (Ruthven, 2014). The assessors needed to select or rate expansion terms from a supplied list. These expansion terms appeared in multiple forms while the same assessors repeat the assessment of expansion terms on the same topic. Such setup is to allow comparison on the consistency of the term selection and revealed that intra-assessor consistency is 75%. The reliability measurement used Spearman's rho (Ruthven, 2014).

Similarly, another study (Blanco et al., 2013) conducted experiments on relevance judgments using crowdsource and showed that the judgments are reliable upon repeated experiments. Three different experiments used two crowdsourced judgments and one expert judges' judgments in measuring inter-assessor agreement using Fleiss's kappa. The Fleiss's kappa is suitable for use with categorical data when there are more than two raters ("Inter-rater reliability," 2017). The study has shown that although the reliability of the crowdsourced judges is lower than expert judges, the system rankings from crowdsourced judges were same as that obtained from expert judges (Blanco et al., 2013).

A reliable test collection implies the conclusions can be replicated with another test collection (Urbano et al., 2013a) although a system performing well with one test

collection does not always perform well with another test collection (Pavlu et al., 2012). To overcome the difficulties of interpreting the Generalizability Theory, the study (Urbano et al., 2013a) proposed a tool based on interval estimates of the stability indicators. Their approach helps in making decisions regarding the number of queries needed in the experimental design. Also, query set size selection depends on the task and test collection as there was significant variation among the test collections. They concluded that some test collections were not reliable (Urbano et al., 2013a).

Reliable effectiveness metrics is also important since variations in retrieval system evaluation could occur due to effectiveness metrics. The effectiveness score of a system is an indicator of the system's performance relative to other systems (Sanderson & Zobel, 2005; Zobel, 1998). A previous study claims that effectiveness from large numbers of topics is more reliable compared to smaller numbers of topics (Sanderson & Zobel, 2005). The study utilized MAP scores from multiple queries and ten samples. The standard deviation of the MAP scores from the ten samples was calculated to determine the reliability of using varied topic sizes. The 'Adaptive' method was experimented to select queries and determine the reliability of system rankings using Kendall's tau and Pearson correlation coefficient (Hosseini et al., 2012). The method produced reliable system rankings.

Numerously available effectiveness metrics for retrieval system evaluation makes it tough to decide the best metric. However, a study proved that some metrics could be grouped in terms of mathematical properties (Baccini et al., 2012). As such, seven groups of metrics were determined, and any one of the metrics within the group could be used as an evaluation metric. As such, few metrics were suggested, the mean average precision (MAP) as a measure of average precision, precision at ten (P@10) to measure satisfaction

of user with the first retrieved document, and the exact_recall to measure the ability of the system to retrieve most of the relevant documents (Baccini et al., 2012).

The average precision (AP) is favorable due to its stability and robustness whereby if the differences are statistically significant, AP-based differences between systems on one set of topics can be observed on another set of topics (Moffat & Zobel, 2008). The RBP, on the other hand, measures the rate at which utility is gained by a user at a given degree of persistence, *p* (Moffat & Zobel, 2008).

4.2.2 Various Categories of Reliability Measurement

Reliability is concerned about the amount of random error within a measurement. If the measurement error is small, the more reliable is the measurement (Rubin & Babbie, 2009). Reliability does not ensure the accuracy of the measurement as it can be subject to systematic error. When taking measurements, the concern behind the measurement is to know how precise the measurement is. Two important aspects of precision are reliability and validity. Reliability implies if a particular measurement is repeatable or reproducible, whereas, validity implies closeness of the measured value to its true value (Hopkins, 2000).

Variance can measure the strength of the estimated reliability. It is a measure of the spread or distribution of a set of scores (Trochim, Donnelly, & Kanika, 2015). A small variance indicates the data are close to the mean and one another, while wide variance means otherwise. Equation 4.1 measures the strength of estimated reliability.

Equation 4.1

$$\frac{Variance\ of\ the\ true\ score}{Variance\ of\ the\ measure}$$

There are multiple ways of measuring reliability which include inter-rater reliability, test-retest reliability, parallel-forms reliability or split-half reliability, and internal consistency reliability (Trochim et al., 2015). Figure 4.1 shows the different reliability categories and a brief explanation about each category. Following subsections detail each category and the type of reliability approach to determine a suitable reliability measure for the experimentation.

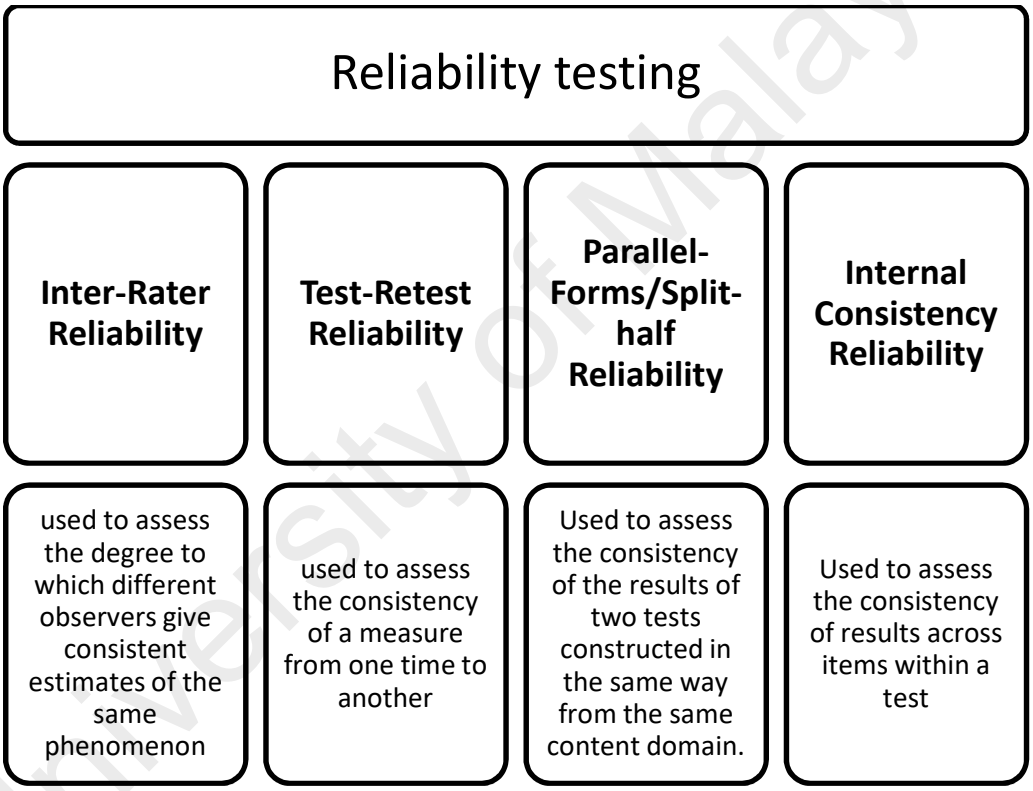

Figure 4.1: Reliability testing categories

4.2.2.1 Inter-rater reliability

The inter-rater reliability is the degree of agreement among the raters based on the ratings given by them. The Pearson product-moment correlation coefficient (Rust & Golombok, 2015), Cohen's kappa, Fleiss' kappa and intraclass correlation coefficient are

examples of inter-rater reliability. Inter-rater reliability is suitable for subjective measurement such as student creativity. One rater might rate the student as very creative (grade 1) while the other rates the student as creative (grade 2). Although the rating is done at the same time, the level of observation by both the raters vary. The inter-rater reliability overcomes such variations.

An example for inter-rater reliability can be explained using two raters A and B as in Table 4.2. These two raters would categorize 100 items into two different categories 'good' and 'bad'. From the categorization, both raters A and B categorized 30 items as 'good' and 40 items as 'bad' while the remaining items were classified differently between the two raters. This inter-rater reliability indicates the two raters are in agreement 70% of the time. Hence, the percent agreement, p_a is 0.7. However, there are possible errors in measuring inter-rater reliability in this way. First is that either rater could have guessed the category and by chance, it is same with the categorization of the other rater. Second, a rater could have randomly selected the category of the item which happens to be the same as that of the other rater. These two scenarios occurred by chance but the reliability measure should be beyond chance (Gwet, 2014).

Table 4.2: Distribution of 100 items by raters and categories

Rater A	**Rater B**		
	Good	**Bad**	**Total**
Good	30	20	50
Bad	10	40	50
Total	40	60	100

Cohen addressed the problem with the occurrence of chance. From the example in Table 4.2, Cohen's kappa uses probabilities of the raters classifying the items in the categories. Based on Cohen's kappa, the chance agreement is now 0.5.

$$p_e = \frac{50}{100} \times \frac{40}{100} + \frac{50}{100} \times \frac{60}{100} = \frac{50}{100} = 0.5$$

Therefore, the inter-rater reliability between rater A and B measured using Cohen's kappa is 0.4.

$$\widehat{K_c} = \frac{p_a - p_e}{1 - p_e} = \frac{0.7 - 0.5}{1 - 0.5} = \frac{0.48}{0.68} = 0.4$$

On the other hand, the Fleiss's Kappa is used to measure percent chance agreement among 3 or more raters. The raters are selected randomly from a group of raters by replacement. These raters would then rate the items that are also randomly selected from a set *n* items (Gwet, 2014). The Fleiss' kappa is used for nominal scale ratings ("Fleiss' kappa," 2017; Gwet, 2014) and the Cohen's kappa is usually used in categorical or qualitative items ("Cohen's kappa," 2017).

The intra-class correlation (ICC) is another form of inter-rater reliability measure. It is a basis of variance analysis and estimation of various variance components (Bartko, 1966). The Winer's approach produces the same intraclass correlation coefficient if the rater's variance-covariance matrix remains unchanged, although original data ratings are modified (Algina, 1978). However, Bartko argued that computing reliability coefficient based on the unchanged variance-covariance matrix is not justifiable (Bartko, 1976). The product-moment correlation employs reliability coefficient on the unchanged variance-covariance. Therefore it is not a recommended measure of reliability (Algina, 1978).

The ICC measures the correlation between one rating (single or mean) of a subject and another rating for the same subject (Shrout & Fleiss, 1979). For ICC(1), correlation 1.0 can be achieved when the within-subjects variance is 0 while the mean square between is

greater than 0. When the within-subject variance is 0, it means the ratings of each subject is identical. In other words, all the ratings by the judges or raters agree for each subject. Hence, a reliability correlation of 1.0, and a perfect agreement (Bartko, 1976). Further details on ICC is in Sections 4.2.3 and 4.2.4.

Depending on the study, average measures could be used as the ratings for intraclass correlation. The reliability coefficient indicates a correlation between different sets of the same number of random raters' average ratings of the subjects (Bartko, 1976). The Spearman-Brown prediction also uses average ratings to assess the reliability. When the ratings are dichotomous, values 0 or 1, the Kuder-Richarson Formula Number 20 is a suitable approach.

4.2.2.2 Test-retest reliability

Test-retest reliability method is one of the methods of reliability testing in quantitative research. In this method, the same test is repeated after a period of time to assess the consistency of a measure from one time to another (Chua, 2013). In retest correlation, reliability determines the closeness of the repeated test to that of the initial testing. It is similar to using correlation coefficient that determines the closeness of system ranks to the original TREC system ranks. The results from the initial and repeated tests are compared, and if their results are identical, the system has high reliability. The research system is reported as reliable when the correlation value is 0.65 and above. However, human measurement using this method may cause skewing (Chua, 2013).

4.2.2.3 Parallel forms and split-half reliability

Parallel forms reliability measures the reliability of different questions and question sets that aim to evaluate the same construct ("Types of reliability," 2017). For example, in one form a question 2 + 3 is stated while in the second form question 4 + 1 is stated. This scenario is measuring the same construct but in separate forms. The reliability is then measured using Pearson product-moment correlation between the scores from both the forms (Rust & Golombok, 2015). However, there is a drawback in using parallel forms reliability testing since two forms are needed. It also means double the work or effort. Due to its drawback, this reliability measure is uncommon compared to its more improved version, the split-half reliability.

The split-half reliability is common in quantitative research and divides the form into two equal halves or groups. If the halves are divided randomly, biasness is prevented (Rust & Golombok, 2015). Also, the split-half answer must be 0 or 1 or simple right or wrong (Shuttleworth, 2009). However, the Pearson product-moment correlation does not measure the reliability of the forms since it is only representing half the form. Nonetheless, the Spearman-Brown can measure the reliability of the entire test (Rust & Golombok, 2015). If the correlations are high, the items are said to have a high level of internal reliability (Chua, 2013).

4.2.2.4 Internal consistency reliability

The internal consistency reliability assesses the consistency of results across items within a test. The Cronbach's α is the most commonly used internal consistency reliability which is equivalent to the mean of all possible split-half coefficients ("Psychometrics," 2017). The Cronbach's alpha reliability method uses the identification of correlation

value between scores for each item in the test and the total score for all items in the test (test index score). Correlation is measured between the items and the test index score. High correlation shows high reliability while low reliability will be discarded.

Through Cronbach's alpha reliability coefficient, the reliability level of the research instrument can be identified. An alpha value of 0.65 to 0.95 is a satisfactory result. A low alpha value indicates the ability of the items in the research instrument to measure the concept (or the variable) is low, and too high alpha indicates all items are similar to one another and is not encouraged (Chua, 2013). Cronbach's alpha can be used for dichotomous scales and multipoint scales.

4.2.3 The Difference Between Pearson Correlation and Intraclass Correlation Coefficient

In the information retrieval evaluation field, Pearson correlation, Kendall's tau, and Spearman rho are the usual correlation coefficients (Hauke & Kossowski, 2011). In short, the Pearson measures linearity between two sets of variables, Kendall's tau measures the rank swaps between two sets of system rankings while Spearman's rho/rank correlation measures the strength and direction of the association between two variables for non-normal distribution of data.

Both the Pearson correlation and intraclass correlation coefficient can be classified as inter-rater reliability, while the ICC can also fit as intraclass. The interclass and intraclass correlation methods are closely related to correlation while sometimes it is more useful and accurate to be measured in regards to analysis of variance (Fisher, 1969). The difference between Pearson correlation and intraclass correlation coefficient (two-way random effects model) can be shown using an example as below.

Table 4.3: Sample data for topics and ranks by different raters

Topic	Rater 1	Rater 2	Rater 3
1	1	6	1
2	2	7	2
3	3	8	3
4	4	9	4
5	5	10	5

Table 4.4: Correlation coefficient for Pearson and intraclass correlation coefficient (ICC) between ranks from different raters (data from Table 4.3)

Rater1	Rater2	Rater3
Pearson	1	1
ICC	0.167	1

Based on the example given in Table 4.3, the Pearson correlation between Rater 1 and Rater 2 is 1, while the correlation between Rater 1 and Rater 3 is also 1. This occurs because of the variance-covariance matrix in unchanged from the original data set (Bartko, 1976). Due to the nature of Pearson correlation, it is not recommended to measure reliability (Bartko, 1976) using Pearson correlation. Meanwhile, the intraclass correlation using two-way random effects model between Rater 1 and Rater 2 is 0.167 while the correlation between Rater 1 and Rater 3 is 1. The intraclass correlation measures the agreement between the two raters. The more uniform the measurement between the raters, higher will be the reliability (Landers, 2015).

4.2.4 Models and Forms of Intraclass Correlation Coefficient

The intraclass correlation coefficient (ICC), is the correlation between one measurement on a target and another measurement obtained on that same target (Shrout & Fleiss, 1979). The technique of finding ICC is a basis of variance analysis and estimation of various variance components known as reliability index. The ICC can only

be interpreted as correlation coefficient if the denominator includes the total variance (Bartko, 1966). Although the ICC is based on analysis of variance, the ANOVA and ICC are different. The ANOVA focuses on measuring the statistical significance between groups but the ICC measures the correlation coefficient ("One-way ANOVA - An introduction to when you should run this test and the test hypothesis | Laerd Statistics," 2017). According to Landers (2015b), ICC shouldn't be used with categorical data but the ANOVA could be applied to two or more independent, categorical groups ("Two-way ANOVA in SPSS Statistics - Step-by-step procedure including testing of assumptions | Laerd Statistics," 2017). ICC restriction suggests that every subject to be rated must have the same number of ratings (Landers, 2015).

The selection of ICC depends on the decomposition of a rating made by the i^{th} judge on the j^{th} target in regards to various effects (Shrout & Fleiss, 1979). An ICC can take values between 1 and -1/(*k*-1) (Bartko, 1976). However, a negative ICC is usually taken to be zero reliability (Algina, 1978). There are three important decisions to select a proper ICC;

(1) whether the ANOVA should be one way or two ways,

(2) whether raters' effect is considered random or fixed, and

(3) whether the analysis unit is a single measurement or mean of few raters.

Figure 4.2 shows the first decision in selecting the ICC model. The first decision is to identify one-way or two-way ANOVA. If the study has non-consistent raters, whereby different raters will rate each of the targets, the ANOVA model is one-way random Model 1. If the study uses consistent raters, it is necessary to determine if the raters are a sample or population. Consistent raters here mean the same raters would rate each target. Use

Model 2 if the raters are samples from a larger population, and Model 3 for population raters or the only raters of interest (Landers, 2015; Shrout & Fleiss, 1979).

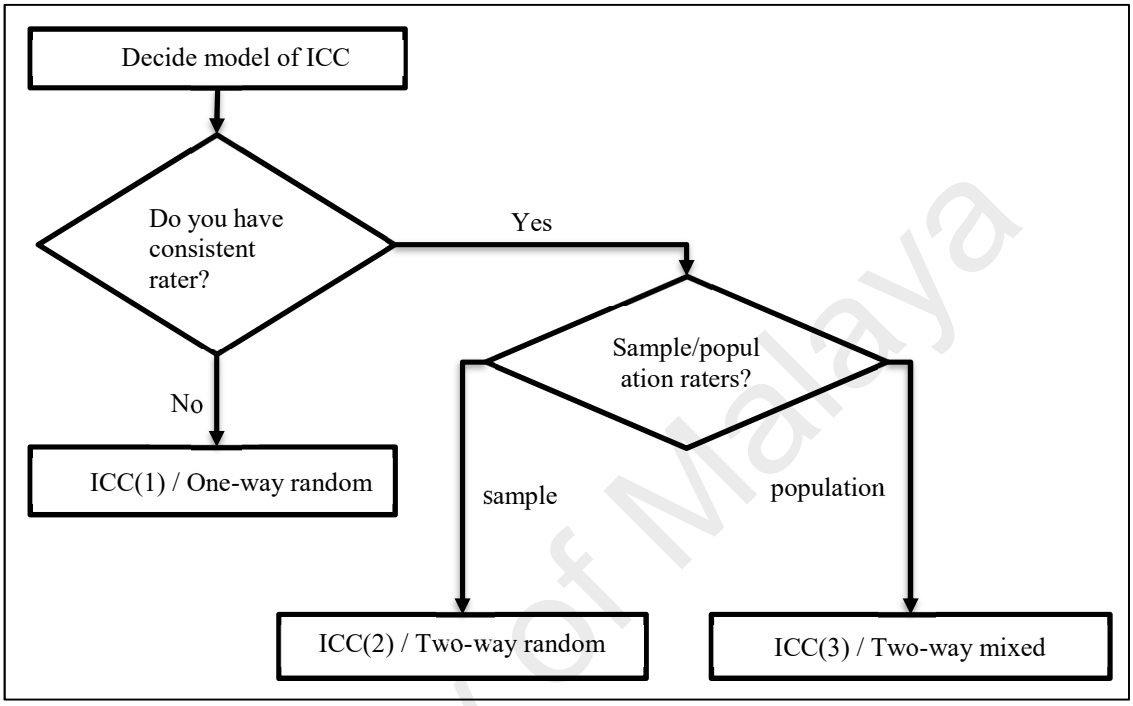

Figure 4.2: Deciding the ICC model (Adapted from (Landers, 2015))

The second decision is to determine if the raters' effect is random or fixed. The Model 3 treats the raters as fixed (Landers, 2015; Shrout & Fleiss, 1979) while removing the between raters variance (Bartko, 1966; Shrout & Fleiss, 1979). When the raters variance is ignored, the correlation coefficient is interpreted as rater consistency rather than rater agreement. The Model 2 treats the raters as random and allows generalization to other raters within the population. Meanwhile, the fixed raters in Model 3 indicate interest in a single rater or a fixed set raters and there are no other raters of interest (Shrout & Fleiss, 1979).

The final decision is related to the analysis unit. The analysis unit could take single or mean measures. The single measure applies to single measurements such as individual scores, and mean measure applies to average measurements such as the average score for a *k*-item test (Barrett, 2001) or an average of 2 or more measurements taken by different raters.

The ICC model and unit of measurement is shown by ICC(*digit1, digit2*), whereby *digit1* represents the model while *digit2* represents the measurement unit. For example, if the selected model is 2 and the unit of measurement is 1, the reliability index is represented by ICC(2,1). This experiment will focus on using single measurement unit, and further detail the models and the equation for each.

The ICC(1,1) is defined as below where $MSB_{targets}$ is mean square between targets, *MSW* is mean square within targets and *k* is the number of raters rating each target (Shrout & Fleiss, 1979).

Equation 4.2

$$ICC(1,1) = \frac{MSB_{targets} - MSW}{MSB_{targets} + (k-1)MSW}$$

The Model 2 has both raters and target effects while assuming both raters and targets are drawn randomly from a larger population (Landers, 2015). The two-way random model using single measurement unit is defined in Equation 4.3, whereby $MSB_{targets}$ is mean square between targets, MSE is mean square error, MSB_{rater} is mean square between raters, *k* is the number of raters rating each target and *n* is the number of targets.

Equation 4.3

$$ICC(2,1) = \frac{MSB_{targets} - MSE}{MSB_{targets} + (k-1)MSE + \frac{k(MSB_{raters} - MSE)}{n'}}$$

The ICC(3,1) is similar to ICC(2,1) but treats the raters as *fixed* (Landers, 2015) while removing the between raters variance (Bartko, 1966; Shrout & Fleiss, 1979). The ICC(3,1) is as below.

Equation 4.4

$$ICC(3,1) = \frac{MSB_{targets} - MSE}{MSB_{targets} + (k-1)MSE}$$

It was pointed out that computing two-way analysis of variance using one-way analysis of variance is not appropriate while ignoring raters variance (Bartko, 1966). As stated earlier, the denominator for ICC should be the total variance of the observation to be interpreted as the correlation coefficient (Bartko, 1966).

4.2.4.1 Manual computation of the ICC(2,1)

A manual computation of ICC may be lengthy compared to using the predefined ICC computation in the R language. But knowing the manual computation may provide a clear understanding of the ICC as a measure of the correlation coefficient. Table 4.5 shows the sample ratings by four raters and six targets as shown in (Shrout & Fleiss, 1979). The various computation to construct the ANOVA consists of MSB (mean square between), MSW (mean square within), MSE (mean square error), and the degree of freedom (df). Note that between targets represent the 6 rows, while between raters represents the 4 columns.

Table 4.5: Sample ratings for four raters and six targets (taken from (Shrout & Fleiss, 1979))

		Rater (k)				
		1	2	3	4	Mean, $\bar{x}_i$
Targets (n)	1	9	2	5	8	6
	2	6	1	3	2	3
	3	8	4	6	8	6.5
	4	7	1	2	6	4
	5	10	5	6	9	7.5
	6	6	2	4	7	4.75
	Mean	7.666667	2.5	4.333333	6.666667	$\bar{\bar{x}}$=5.291667

1. Firstly, calculate the degree of freedom (*df*) for raters, targets, within targets and errors using the defined formula below.

k = number of raters

n = number of targets

N = total number of ratings ($n \times k$)

$df_{raters} = k - 1$

$df_{targets} = n - 1$

$df_{within\ targets} = n(k - 1)$

$df_{error} = (n - 1)(k - 1)$

2. Next, notice that Table 4.5 contains a mean column and row. These values are the mean of the respective rows and columns. The overall mean of the ratings, $\bar{\bar{x}}$ is

$\bar{\bar{x}} = sum\ of\ ratings / total\ number\ of\ ratings$

$\bar{\bar{x}} = 127/24 = 5.291667$

3. Next is the calculation of the sum of squares between targets ($SSB_{targets}$) and mean square between targets ($MSB_{targets}$).

$$SSB_{targets} = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2$$

$$SSB_{targets} = 4 \times (6 - 5.291667)^2 + 4 \times (3 - 5.291667)^2 + 4 \times (6.5 - 5.291667)^2 + 4 \times (4 - 5.291667)^2 + 4 \times (7.5 - 5.291667)^2 + 4 \times (4.75 - 5.291667)^2$$

$$SSB_{targets} = 2.006944 + 21.00694 + 5.840278 + 6.673611 + 19.50694 + 1.173611 = 56.20833$$

$$MSB_{targets} = SSB_{targets}/df_{targets}$$

$$MSB_{targets} = 56.20833/(6 - 1) = 11.24$$

4. Following that is the calculation of the sum of squares between raters (SSB_{raters}) and mean square between raters (MSB_{raters}).

$$SSB_{raters} = \sum_{j=1}^{n} k_j\left(\bar{x}_j - \bar{\bar{x}}\right)^2$$

$$SSB_{raters} = 6 \times (7.666667 - 5.291667)^2 + 6 \times (2.5 - 5.291667)^2 + 6 \times (4.333333 - 5.291667)^2 + 6 \times (6.666667 - 5.291667)^2$$

$$SSB_{raters} = 33.84375 + 46.76042 + 5.510417 + 11.34375 = 97.45833$$

$$MSB_{raters} = SSB_{raters}/df_{raters}$$

$$MSB_{raters} = 97.45833/(4-1) = 32.49$$

5. Then, let's calculate the sum of squares within targets ($SSW_{targets}$) and mean square within targets ($MSW_{targets}$). This computes the difference of each rating and its mean within a target.

$$SSW_{target} = \sum_{j=1}^{n}\sum_{i=1}^{k}\left(x_{ij} - \bar{x_j}\right)^2$$

$$SSW_{target} = (9-6)^2 + (2-6)^2 + (5-6)^2 + (8-6)^2 + (6-3)^2 + (1-3)^2 + (3-3)^2 + (2-3)^2 + (8-6.5)^2 + (4-6.5)^2 + (6-6.5)^2 + (8-6.5)^2 + (7-4)^2 + (1-4)^2 + (2-4)^2 + (6-4)^2 + (10-7.5)^2 + (5-7.5)^2 + (6-7.5)^2 + (9-7.5)^2 + (6-4.75)^2 + (2-4.75)^2 + (4-4.75)^2 + (7-4.75)^2$$

$$SSW_{target} = 9 + 16 + 1 + 4 + 9 + 4 + 0 + 1 + 2.25 + 6.25 + 0.25 + 2.25 + 9 + 9 + 4 + 4 + 6.25 + 6.25 + 2.25 + 2.25 + 1.5625 + 7.5625 + 0.5625 + 5.0625$$

$$SSW_{target} = 30 + 14 + 11 + 26 + 17 + 14.75 = 112.75$$

$$MSW_{target} = SSW_{target}/df_{within\ targets}$$

$$MSW_{target} = 112.75/6(4-1) = 6.26$$

6. Next, calculate the sum of squares total (SST) which is calculated using each rating and the overall mean as shown in Table 4.6.

Table 4.6: Calculation of the sum of squares total

Ratings, x_{ij}	Overall mean, $\bar{\bar{x}}$	$(x_{ij} - \bar{\bar{x}})^2$
9	5.291667	13.75174
6	5.291667	0.501736
8	5.291667	7.335069
7	5.291667	2.918403
10	5.291667	22.1684
6	5.291667	0.501736
2	5.291667	10.83507
1	5.291667	18.4184
4	5.291667	1.668403
1	5.291667	18.4184
5	5.291667	0.085069
2	5.291667	10.83507
5	5.291667	0.085069
3	5.291667	5.251736
6	5.291667	0.501736
2	5.291667	10.83507
6	5.291667	0.501736
4	5.291667	1.668403
8	5.291667	7.335069
2	5.291667	10.83507
8	5.291667	7.335069
6	5.291667	0.501736
9	5.291667	13.75174
7	5.291667	2.918403

$$SST = \sum_{j=1}^{n}\sum_{i=1}^{k}\left(x_{ij} - \bar{\bar{x}}\right)^2 = 168.9583$$

7. Lastly, calculate the sum of squares error (SSE) and mean square error (MSE).

$$SSE = SST - SSB_{targets} - SSB_{raters}$$

$$SSE = 168.9583 - 56.20833 - 97.45833 = 15.29167$$

$$MSE = SSE/df_{error}$$

$$MSE = 15.29167/(6-1)(4-1) = 1.02$$

8. Based on all the above calculations, the ANOVA components are constructed in Table 4.7. These values can then be used in computing the various ICC models.

Table 4.7: ANOVA components for the data in Table 4.6.

Source of variance	df	MS
Between targets	5	11.24
Within targets	18	6.26
Between raters	3	32.49
Error	15	1.02

For example, the model ICC(2,1) has the following formula and the reliability value can be calculated as below.

$$ICC(2,1) = \frac{MSB_{targets} - MSE}{MSB_{targets} + (k-1)MSE + \frac{k(MSB_{raters} - MSE)}{n'}}$$

$$ICC(2,1) = \frac{11.24 - 1.02}{11.24 + (4-1)(1.02) + \frac{4(32.49 - 1.02)}{6}}$$

$$ICC(2,1) = \frac{10.22}{11.24 + 3.06 + 20.98}$$

$$ICC(2,1) = \frac{10.22}{35.28} = 0.29$$

An ICC can take values between 1 and -1/(k-1) (Bartko, 1976). However, a negative intraclass correlation coefficient is usually taken to be zero reliability(Algina, 1978).

4.3 Methodology

This study aims to evaluate the reliability of individual retrieval system's ranks based on their relative rankings. The proposed approach would use a combination of metrics to determine the reliability of individual system's ranking. The approach is proposed based on few possible scenarios that could affect the reliability of a retrieval system.

In the first scenario, a user expects a system performing well with a set of topic to perform well with another set of topic. In other words, a system should be able to perform consistently in retrieving documents that are relevant to the user each time or most times. For example, when comparing two systems, system A is able to retrieve 80% of the relevant documents while system B is able to retrieve 30% of the relevant documents. So a user may prefer to use system A since it has a better performance. However, does this system A maintain its performance in comparison to system B for another query? If system A is able to perform better than system B for various other queries, system A is likely to be a favorite choice of the user. As the performance of system A is consistent and reliable, and provides confidence to the user regarding its performance in relation to other systems.

In another scenario, a metric measures a system's performance according to its user model and underlying probabilistic interpretation. A system may perform well in one aspect of the user model but not the other or may perform well or not in both the user models. This can be understood through its effectiveness scores from both the metrics. Although the system performs well when measured using one metric but not the other, the rank of the system may very well be same and better than the other systems for both metrics. Even though the system's effectiveness score for the second metric is poor, the rank could still be better than the other systems. The user still gains from the system despite the lower effectiveness from the second metric since it still has a better

performance comparatively. The variation in effectiveness scores could have occurred from the user model employed for the metric.

It is inevitable that a system would perform differently for different topics or metrics. But, their ranking from the different effectiveness scores provides a clearer picture of the systems' performance compared to the other systems. Therefore, the need for evaluation of system performance in terms of reliability of ranking is further emphasized. A reliability score could indicate the level of confidence a user can expect from the retrieval system in satisfying the user's information need. The reliability score is beyond the variations of topics and metrics, and is not specific to good or poor performing system. Hence, this approach focuses on measuring the reliability of the individual system ranks using combination of metrics and topic sizes.

4.3.1 Test Collections

The experimentation uses data collection from the TREC-2004 Robust track, consisting of 110 systems and 249 topics. Topic 672 was removed because there were no relevant documents. Some of the topics from this test collection are from previous TREC. Only the 50 new topics use graded relevance judgment while the rest uses binary relevance. For this experimentation, the graded relevance judgment of the documents was translated into binary relevance. The relevant and highly relevant documents were both categorized as relevant.

A test collection from the TREC-2005 Robust track will be used for the validation experiment. This collection contains 74 system runs and 50 topics. These topics are said to be difficult topics taken from another collection and they also appeared in the TREC-2004 Robust track. However, the systems are different between both the test collections.

Compared to the earlier test collection, the topic size is much smaller for TREC-2005 Robust track collection. Nevertheless, the test collection could still be of use with the assumption a person evaluating the systems would prefer to utilize smaller numbers of topics.

4.3.2 Intraclass Correlation (ICC) Fitting for the Experimentation

Based on the various correlation coefficient and ICC discussed in the above sections, the ICC is most suited for this experimentation. Although ICC is not common in information retrieval evaluation, the concept of ICC can be adapted to IR evaluation. Before the experimentation, there's a need to make three important decisions in determining the ICC model.

The first decision is to determine the ANOVA model as one-way or two-way. For this experimentation, the effectiveness metrics are considered as the raters. More than one metrics will be used, which fits the two-way ANOVA model. Following this, there are various metrics for evaluation but only a few metrics will be used in the experimentation. Therefore, the metrics are a sample from a population. Hence, the ANOVA model is a two-way random.

The second decision is to determine the random or fixed effect of the raters, which is the effectiveness metrics. Model 2 treats the metrics as a random effect and the correlation coefficient is measured as agreement. The Model 2 also allows generalization to other raters within the population (Shrout & Fleiss, 1979). Model 3 treats the metrics as fixed or the only metrics of interest and measures the correlation coefficient as consistency. As shown earlier in Section 4.2.3, consistency and agreement are very different. This experimentation is inclined to measure the agreement of the ranks from the effectiveness

metrics to determine the reliability of individual retrieval system ranks. Besides, the study also attempts to determine if similar results could be obtained with the use of other similar metrics within the population. Therefore, Model 2 best suits the criteria of random effects, generalization and measurement of correlation coefficient as agreement.

The third decision is determining the measurement unit of the ratings. The ratings are represented by the rank of topics per system in relation to other systems. These individual ranks are the measurement unit of reliability, which is single measurements. If an average of the ranks is used, then the measurement unit would be the mean.

Table 4.8 shows an example of the matrix formed from the ranks of two different metrics. These ranks will be used for the correlation coefficient as a measure of agreement between metrics using ICC(2,1).

Table 4.8: Sample matrix of ranks from two metrics for random selection of topics

	Ranks	
Topics	**AP@1000**	**RBP@1000**
686	62	59
333	62	34
354	42	49
654	33	30
660	52	64
379	20	17
362	35	34
308	45	25
427	39	38
449	71	68

4.3.3 The Proposed Method for Measuring the Reliability of Individual Retrieval System Rankings

Figure 4.3 shows the approach undertaken to measure the reliability of individual system rankings and the correlation coefficient with the existing evaluation approach. The

lighter, dotted lines represent the existing approach while the darker, bolder lines represent the proposed approach. The numbers in the figure is a guide to explain the stages of the proposed approach.

The first stage is to select the various system runs from the test collection. The second stage is to split each system run according to their topics and identify the relevancy from the relevance judgment. At stage 3, use metrics to calculate the effectiveness scores of each topic, per system run. Then, in the existing method, these topic scores are aggregated to obtain an overall system score.

Stage 4 is the starting point of the proposed approach. Rank the systems for topic T1 using the effectiveness scores from metrics A. Next, rank the systems for topic T2 using the effectiveness scores from metrics A and continue for the rest of the topics, Tn. Repeat the same for metrics B. Such that, rank the systems for topic T1 using effectiveness scores from metrics B and repeat the same for the remaining topics, Tn.

Continuing with the proposed approach at stage 5, select *R* topics randomly. The topic size, *R* is 10, 20, 30, 40 or 50. The topic ranks are extracted from both the metric A and B for a particular system. The reliability test is for individual systems, therefore the ranks from the randomly selected topics are from a specific system run, let's say system A. Then at stage 6, calculate the mean rank for the randomly selected topics from both the metrics for system A. The mean score will be used for the evaluation at stage 8. Still at stage 6, perform the intraclass correlation coefficient reliability test using the ranks from both the metrics A and B for system A. Here, metrics A and B represent two different raters. For system A, the ICC will measure the agreement of the ranks between metrics A and B.

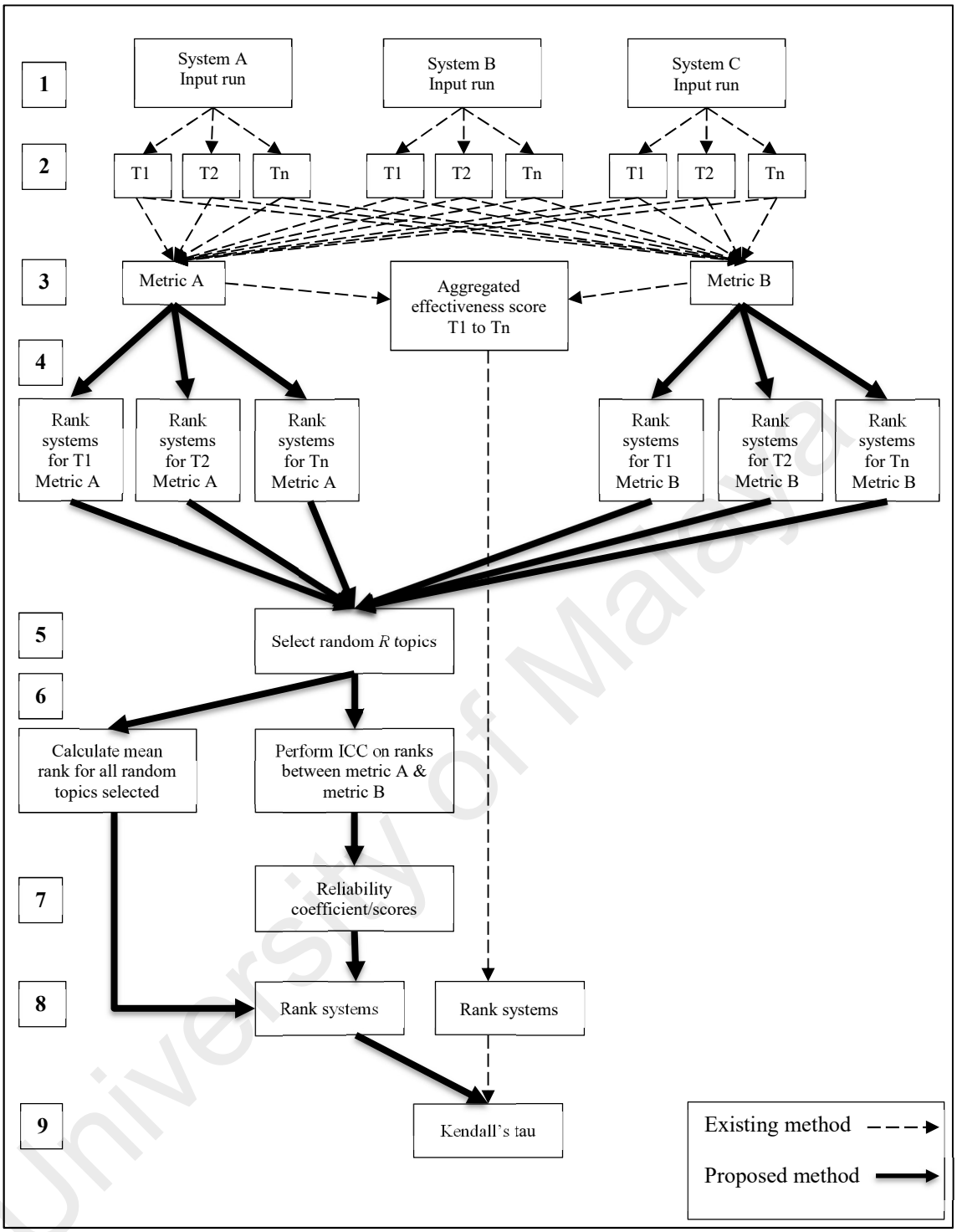

Figure 4.3: The proposed approach to measure the reliability of individual system ranks

In order to measure the reliability of other systems, the same randomly selected topics will be used to extract the respective ranks from another system, system B. The mean and reliability test is performed for system B similar to that done for system A. The same is repeated for all system runs. At this point, each system has its reliability coefficient scores

that would indicate the confidence a user can gain from the system. However, the user will not be able to determine the performance of the systems despite knowing their reliability coefficient. The mean rank will provide the information about the performance of a particular system in relation to other systems.

At stage 8, the systems are ranked using the aggregated effectiveness scores for the existing method. Similarly, overall system rank is needed for the proposed approach. The overall rank for each system is obtained from the mean rank. Any systems with identical ranks will be ordered by their reliability coefficient score in a descending manner. The mean rank best suits to represent the reliability score since the analysis of variance components calculate variance mainly from the mean rank. A similar tie breaker has been used for the existing method, whereby any systems with identical aggregated effectiveness scores will be sorted by the system name. The process of tie breaking will provide us with overall system rank without any identical ranks. These ranks can be used for Kendall's tau correlation coefficient.

At stage 9, Kendall's tau correlation coefficient is performed between the system ranks from existing method and overall system ranks from the proposed method. The Kendall's tau value would validate the system ranks from the proposed method with the gold standard. Therefore, the reliability coefficient for each system could be a representative of the original system rankings. The process of selecting random topics, calculating mean, performing reliability test, obtaining overall system ranks and correlation coefficient (stages 5 – 9) is repeated 100 times to allow for various combinations of topics.

4.3.4 Evaluation of Individual Retrieval System Rankings' Reliability

The proposed method measures the reliability of individual retrieval system rankings with combination of metrics. Each metric has its underlying user model. A reliable system measured from a combination of metrics showcases its capabilities to satisfy more than

one user model. The metrics selection are average precision, rank-biased precision, and precision at cut-off *k*. These metrics could be used to measure the effectiveness of the retrieval systems at different depth of evaluation. Some are evaluated shallowly, while others at depth 1000.

Randomly selected topics will be used and they could be 10, 20, 30, 40, and 50. A total of 249 topics are available for evaluation in the TREC2004 Robust track. Topics highly influence the comparison between retrieval systems. It is likely for a different set of topics of the same size to produce different results (Voorhees & Buckley, 2002) and some topics or topic sets to predict the actual effectiveness of systems better than others (Guiver et al., 2009). Hence, 100 iterations of the same topic size should minimize the effect of topic variation.

The evaluation of retrieval systems based on their reliability measures includes identifying the number of retrieval systems with high reliability score. A high reliability score indicates the system is capable of sustaining its performance in relation to other systems. Meanwhile, a poor reliability score indicates the system rankings have been varied, sometimes performing poorly or sometimes performing well. A system with poor reliability could affect the confidence of the user on the retrieval system due to inconsistent results when compared with other systems. Also, recall that the reliability score does not indicate the performance of the system with regards to other systems but rather the internal performance of the system itself. If a system has a reliability score of 0.8 and another has 0.7, it doesn't mean the performance of the first system is better than the latter. It just indicates the first system is capable of producing similar results 80% of the time, maybe at an effectiveness measure of precision of 0.5. It translates to having 5 relevant documents in the top 10 ranked list, and the system is likely to give the user 5 relevant documents about 80% of the time. This experimentation limits to measuring the

reliability and not the overall performance of the system. But the study validates the results with the gold standard performance to state the reliability score could be the representative of the original system performance.

The evaluation tackles the generalization of the results to other similar metrics. From a previous study (Baccini et al., 2012), metrics have been grouped together. This clustering or grouping is used as a way to identify similar metrics. Hence, the combination of metrics used initially will be replaced with another pair of similar metrics from the same cluster. And with the use of the two-way random model, ICC(2,1), the results obtained should allow for generalization. This evaluation is to understand the possibilities of generalization besides knowing the results are consistent.

The third evaluation is to identify the systems from the gold standard that are measured to be reliable. Although the experimentation has now shown the reliability of individual system rankings, it is not clear as to which systems have reliable rankings. Evaluating the results with graph plotting will provide an understanding of the reliable system with regards to their original system rankings.

The fourth evaluation undertaken is to validate the rankings from the proposed method. The reliability measurement provides the reliability coefficient of the individual systems, and it has been compared with the original system rank but it is still not known if the inferred mean rank from the proposed method correlates with the gold standard. A strong correlation coefficient would indicate the similarity of the system ranks between the proposed and gold standard. Since in the earlier evaluation the reliability score had been represented for the original system rank, it is necessary to also know if the rank from proposed method tallies with the gold standard. The evaluation output will provide certainty of the reliability score representation to the original system ranks when the correlation coefficient is strong.

The final evaluation is the repetition of the experimentation with another test collection just as if it is a one-time execution to determine the reliability of the individual system rankings. It will not include multiple selections of the same topic size but just one random topic selection. If the results are comparable with that from the initial experimentation, it can be accepted as a reliable and consistent approach for measuring the reliability of the individual system rankings.

Random errors in the current experiment could occur due to topic variation, relevance judgments or the way the reliability measure is constructed. The experiment is not impacted by observer error but could be affected by judges from relevance judgments. However, this effect should not impact the measurement of the systems' ranks.

4.4 Results and Discussions

This section consists of the experimental results and discussions presented in different subsections. Firstly, Section 4.4.1 presents the results of reliability measurement of individual system rankings for different combination of metrics. Next, Section 4.4.2 details the generalization of reliability measurement results using other similar metrics. Then, Section 4.4.3 identifies the original systems with reliable system rankings. Section 4.4.4 measures the correlation coefficient between the original system ranks and the mean rank inferred from proposed method. Lastly, Section 4.4.5 validates the findings of the proposed approach with another test collection in a one-time retrieval system evaluation.

4.4.1 Reliability of Individual Retrieval System's Ranking Measured by Intraclass Correlation Coefficient

This section focuses on detailing the reliability of individual retrieval system rankings and identifying suitable combination of metrics in measuring reliability of individual

system rankings and thus, corresponds to RQ1. The reliability of a retrieval system is measured by intraclass correlation coefficient from a combination of different evaluation metrics. The ICC(2,1) measures the agreement between two metrics using a single measurement of the topic ranks. The reliability will be higher when the ranks between the metrics are more uniform (Landers, 2015). A reliability coefficient of 0.8 and above indicates high reliability.

Reliability evaluation was conducted on groups of metrics. The combinations include within metric groups and outside metric groups as shown in Figure 4.4. In the figure, metrics A and C are assumed to be from the same group while metrics B and D are assumed to be from another metric group. The reliability evaluation is measured for both the outside metrics group and within metrics group. The outside metric groups consist of initial and generalization experimentation.

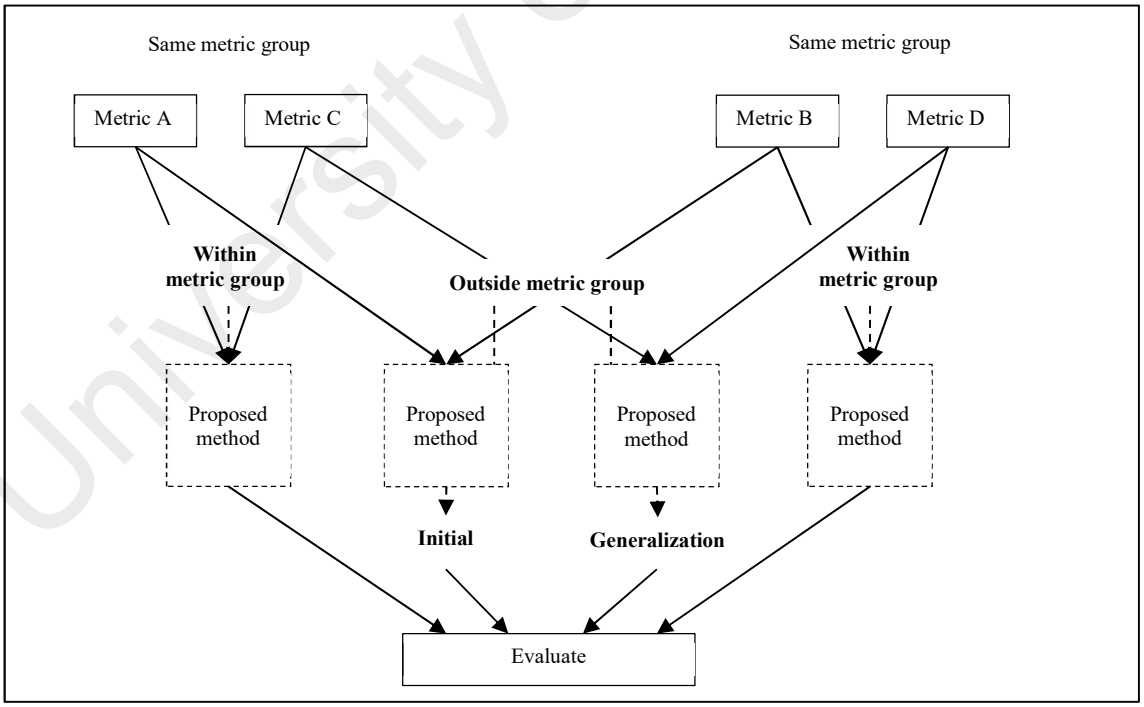

Figure 4.4: Outside metrics group and within metrics group combinations

The results from these different combinations of metrics from outside the metric group are shown in Table 4.9. The table shows results of various numbers of topic sizes used for each combination of metrics and the number of systems within each reliability score from 0.1 to 0.9 with an interval of 0.1. The data from Table 4.9 detailed in terms of total number of highly reliable system rankings for each combination of metrics and topic sizes is in Appendix J.

Based on the results shown in Table 4.9, two different metrics combinations AP@100-RBP@100 and RBP@100-P@30 have large numbers of systems with highly reliable rankings. These systems have reliability scores above 0.8. The results suggest consistent individual system rankings. All of the topic sizes also tend to have large numbers of systems with reliable rankings. If looked into specifically, for metric combination of AP@100-RBP@100, topic size 10 yields the highest numbers of systems with highly reliable rankings. Meanwhile, for metrics combination RBP@100-P@30, topic size 30 yields the highest numbers of reliable system rankings. These data suggest that small topic size is sufficient to measure the reliability of a system's rank. It is beneficial to use smaller topic sizes in evaluating the information retrieval systems as they provide the opportunity to researchers in measuring the system ranking reliability with reduced effort in terms of judging fewer topics.

The other combination of metrics has large numbers of systems within the moderately reliable system rankings. The reliability scores range from 0.4 to 0.8. These systems' rankings may be accurate 40% to 79% of the time upon repeated measure. For the three combinations of metrics, namely the AP@100-P@30, AP@100-P@100, and RBP100-P@100, highest numbers of systems with reliable rankings are for topic size 10. Again, it appears to be sufficient to measure individual system rankings' reliability with small topic size using the proposed method.

Table 4.9: Number of systems within each reliability scores for different topic sizes and combination of metrics

Metrics combination	Topic size	Reliability scores								
		0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
AP@100-RBP@100	10	0	0	0	0	0	2	7	**67**	**34**
	20	0	0	0	1	5	1	4	**73**	**26**
	30	0	0	1	4	2	2	14	**67**	**20**
	40	0	0	3	2	2	2	14	**64**	**23**
	50	0	0	5	0	4	12	4	**62**	**23**
AP@100-P@30	10	0	1	0	3	21	53	26	6	0
	20	0	1	3	5	20	48	28	5	0
	30	0	1	3	7	22	45	27	5	0
	40	0	2	5	2	19	46	30	6	0
	50	0	2	5	5	16	46	30	6	0
AP@100-P@100	10	1	0	1	3	25	61	19	0	0
	20	1	0	1	6	23	53	26	0	0
	30	1	0	3	5	22	49	30	0	0
	40	1	1	5	2	24	51	25	1	0
	50	1	2	4	3	23	53	24	0	0
RBP@100-P@30	10	0	0	0	1	0	27	18	**55**	**9**
	20	0	0	0	1	0	27	11	**60**	**11**
	30	0	0	0	1	0	17	19	**63**	**10**
	40	0	0	0	1	0	6	34	**58**	**11**
	50	0	0	0	1	0	21	16	**60**	**12**
RBP100-P@100	10	1	0	1	2	27	26	52	1	0
	20	1	0	1	2	24	27	50	5	0
	30	1	0	1	2	17	31	35	23	0
	40	1	0	1	2	21	27	22	36	0
	50	1	0	1	1	21	28	25	33	0

Note: Bold cells indicate high numbers of reliable rankings

In addition to measuring the reliability of system rankings using a combination of metrics from different groups, the reliability of individual system rankings was also measured using metrics combinations within the same groups. Table 4.10 shows the numbers of systems that have highly reliable system rankings when measured using metrics from the same groups. The table also displays the numbers for various topic sizes experimented.

Table 4.10: Number of systems with highly reliable system rankings measured by within the same group metrics combination

Topic size	10	20	30	40	50
AP@1000-AP@100	67	71	73	73	74
RBP@1000-RBP@100	83	83	83	83	83
P@30-P@10	0	0	0	0	0
P@100-P@200	109	109	109	109	109

The results from Table 4.10 shows that all within metric combinations have large numbers of highly reliable system rankings except for metrics combinations of P@30-P@10. These numbers suggest that the metrics within the group appear to measure the system rankings in a similar manner such that their system rankings are consistent between the metrics. Similarities in rankings could have resulted from little variation in effectiveness scores that translated to the ranks. The number of systems with highly reliable system rankings holds true for small topic sizes as well as large topic sizes. It can be reiterated that small topic sizes appear to be sufficient in measuring the reliability of individual retrieval system rankings.

However, the combination of P@30 and P@10 shows no reliable system ranks. One obvious difference of these metrics combination is the depth of evaluation. The evaluation depth is rather shallow compared to 100, 200 and 1000. Due to this shallow evaluation, the effectiveness scores of the systems may have been somewhat similar, causing system ranks to vary drastically. The rankings are obtained by ordering the systems by their effectiveness scores followed by alphabetically. Possibly, queries with fewer relevant documents could have caused the effectiveness metrics to become unstable, thus causing variation in the system performance and rankings (Voorhees, 2000). The P@10 and P@30 belong to the shallow depth evaluation metric suggested by Baccini et al. (2012). Hence, using P@k where k is shallow is not suitable for measuring the reliability of individual system rankings.

The ICC approach is capable of measuring the reliability of individual system rankings using various combination of metrics; outside metrics groups and within metrics groups. Measuring reliability of individual retrieval system rankings with small topic size generates most numbers of reliable system rankings. Nonetheless, shallow depth of evaluation using P@*k* is not a suitable combination of metrics for measuring the reliability of individual system rankings.

4.4.2 Generalization of the Reliability Measure to Other Similar Metrics

This section details the generalization of the retrieval system's reliability measurement using other similar metrics and thus answers RQ2. The proposed method uses the ICC(2,1) for measuring the reliability of the individual retrieval system rankings. The ICC model allows for generalization of its results to another similar metrics combination (refer Figure 4.4). Generalization of metrics means a specific result can be replicated for the same systems using similar metrics.

The generalization is measured for metrics combination outside the cluster only. If a similar metric combination is able to identify somewhat the same number of systems, regardless the systems are highly reliable or not reliable, then the generalization is valid. However, if the numbers are very different, then there is no generalization to other similar metrics. It means the usage of another combination of similar metrics will not yield comparable results as the initial. Notice that the generalization evaluation here is not focused on the highly reliable systems but whether the results from the initial and generalized metrics are similar.

Figure 4.5 shows the generalization results for measuring the reliability of individual system ranking with another similar metrics combination. The figure contains five

separate graphs labeled 1 to 5 for the outside cluster metrics combinations; AP-RBP, AP-P@*k* (*k*=small), AP-P@*k* (*k*=large), RBP-P@*k* (*k*=small), and RBP-P@*k* (*k*=large). Within each graph, there are 5 panels, one panel for each topic size. The x-axis represents the ICC(2,1) results for the primary or initial metrics combinations, and the y-axis represents the ICC(2,1) results for the generalization metrics combination. The dotted lines within the graphs and panels are marked at 0.4 and 0.8 on both axes. The reliability coefficient values below 0.4 are low in reliability, within 0.4 and 0.8 are moderately reliable, and those above 0.8 are highly reliable.

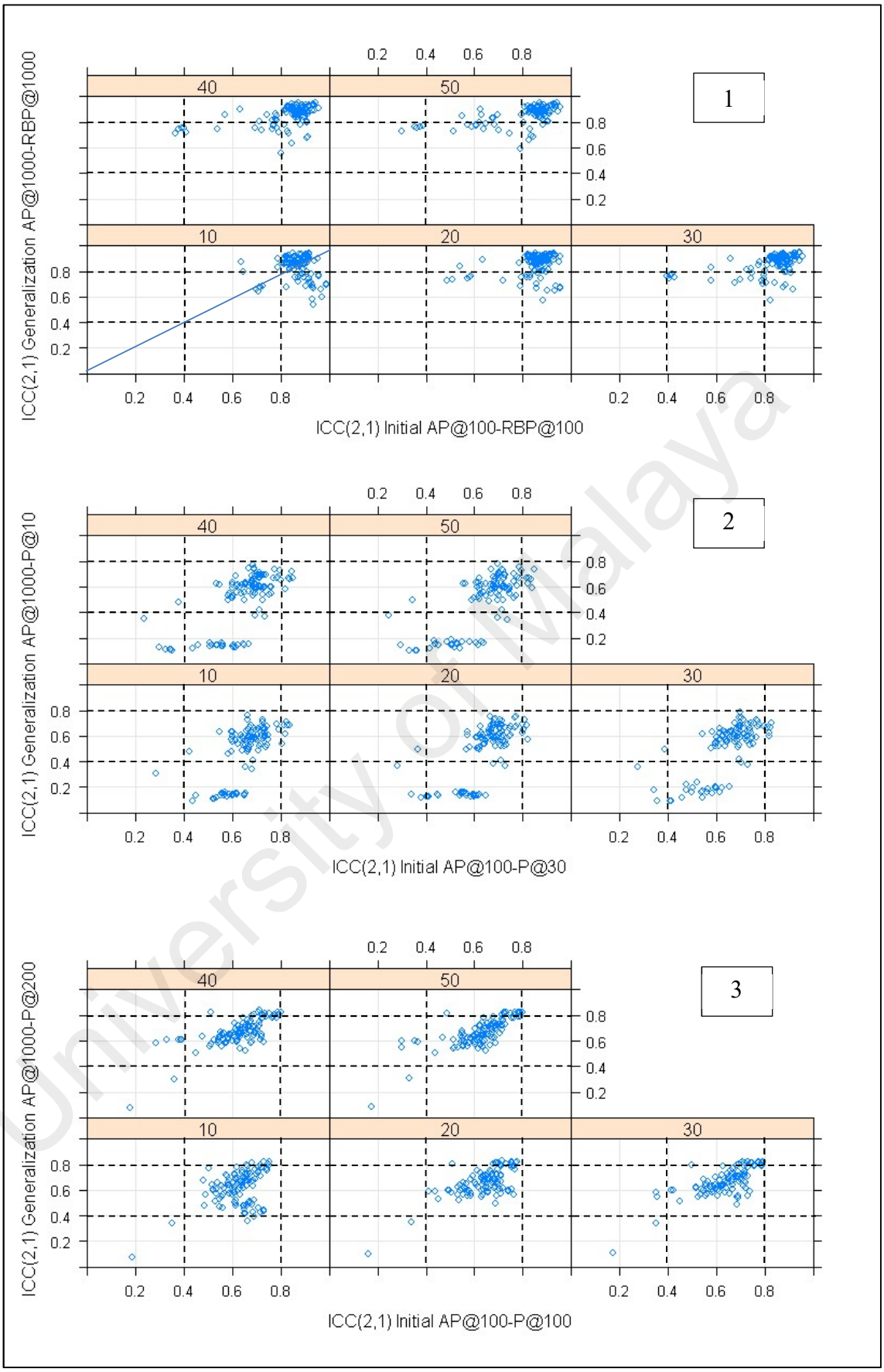

Figure 4.5: TREC-2004 Robust Track – Generalization results for various metrics combination

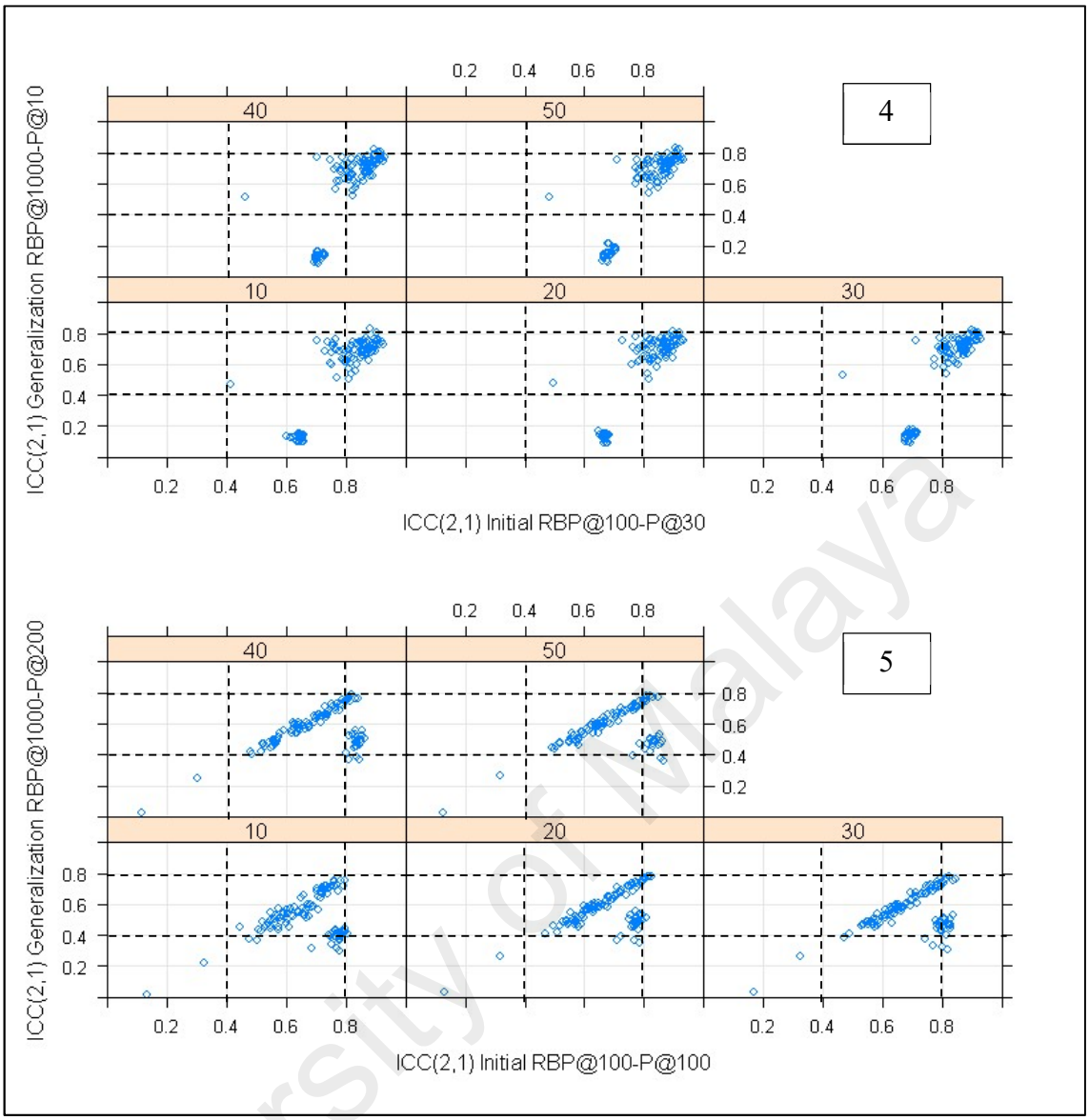

Figure 4.5, continued: TREC-2004 Robust Track ─ Generalization results for various metrics combination

If the plots fall within the same reliability coefficient for both primary and generalization metrics, the results would indicate the metrics combination produces consistent results. An easy way to recognize the region is to focus on the diagonal cubes formed from the dotted guidelines (a diagonal line is shown in the graph labelled 1). If the plots fall within this diagonal region, the results from primary metrics replicate to other similar metrics. If the plots fall in different regions, it means the generalization is

not valid. Thus, the metrics combination is not a suitable option for measuring system ranking reliability due to possible variation in results.

From Figure 4.5, four out of five metrics combinations have reproduced similar results with their initial metrics combination. The only metric combination that does not generalize well is the RBP-small P@k (graph 4). The remaining all have their plots mostly within the diagonal region which indicates similarities of reliabilities scores between the initial and generalized metrics combination. The results are consistent across all topic sizes.

However, highly reliable system rankings are favorable for user satisfaction. Previously it was highlighted that AP@100-RBP@100 and RBP@100-P@30 metrics combinations were able to measure large numbers of systems that are reliable in their rankings. The generalization metrics combination for these metrics is AP@1000-RBP@1000 (graph 1) and RBP@1000-P@10 (graph 4).

For graph 1, the generalization results are similar to that of its initial metrics combination AP@100-RBP@100. It also has large numbers of systems which are highly reliable in their system rankings. The results indicate that using similar metrics within this group of metrics would yield similar results in measuring the reliability of the individual system rankings and yet measure high reliability.

However, for graph 4 (RBP@1000-P@10 & RBP@100-P@30), the generalization results are not similar to that of the initial metrics combination. The initial metrics combination RBP@100-P@30 had large numbers of highly reliable system rankings, but the generalization metrics have measured them as moderately reliable systems only. Chances are the results may vary when measuring the reliability of system rankings using metrics combination within these groups of metrics.

As with graph 2, most plots fall into the category of moderately reliable systems for both the initial and generalization. However, generalization measured some systems as poor reliability but was initially classified as moderately reliable systems. Although this metrics combination and its generalization mostly agree, approximately 20% of the systems tend to have varied results.

On a positive note, graph 3 (AP@100-P@100 & AP@1000-P@200) shows that initial and generalization metrics combination have similar results. These combinations measured large numbers of moderately reliable system rankings. Although these metrics produce consistent results between initial and generalization, the individual system rankings are not reliable.

As for graph 5, the initial and generalized metrics combination has most systems with moderate reliability. However, for the primary metrics evaluated with topic size 30, 40 and 50, there are about 20 to 30 systems that were measured as highly reliable, but the generalization only measured it moderately. These metrics groups combinations also may produce varying reliability results when similar metrics are used for future evaluation.

The various metrics combinations generalize well except RBP-P@k (k=small). Generalization disregards the reliability coefficient values but focuses on obtaining similar results between the initial and similar metrics combinations. The usage of similar metrics within these groups and combinations should produce comparable results in future evaluations. However, it is favourable to identify retrieval systems that are highly reliable in their rankings. Therefore, AP-RBP metrics combination is the best option.

4.4.3 Retrieval Systems from Original System Ranks that are Highly Reliable in Their Rankings

Up to this point, the metrics combinations which produce highly reliable system rankings and those that produce similar output with its generalized metrics are known. However, it is not known which systems from the original system ranks do these reliability scores represent. Hence, this section focuses on identifying the original system ranks that are reliable and answers to RQ3. To determine the original system ranks and its corresponding reliability score, the average reliability scores from 100 iterations are plotted against the original MAP system ranks (refer to Figure 4.6).

Figure 4.6 only consists of data from topic size 30 from the various metrics combinations, including the initial and generalization. Each panel represents a metric combination and is labeled 1 to 5. The plots represent the average reliability scores for initial and generalization metrics combination. The horizontal dotted lines represent the correlation coefficient value of 0.8. Topic size 30 is a random selection to show the various plots since the data distribution is not very different from each topic size. The detailed plots of each topic size per metrics combination are included in the Appendix A till E.

From Figure 4.6, graph 1 representing metrics AP-RBP combination shows that highly reliable system rankings belong to systems that are top and middle ranked. These first 80 systems from the total 110 systems have strong reliability coefficient. Meanwhile, the remaining bottom ranked systems tend to have moderate reliability scores. The similar trend is observed for both the initial and generalized metrics combinations.

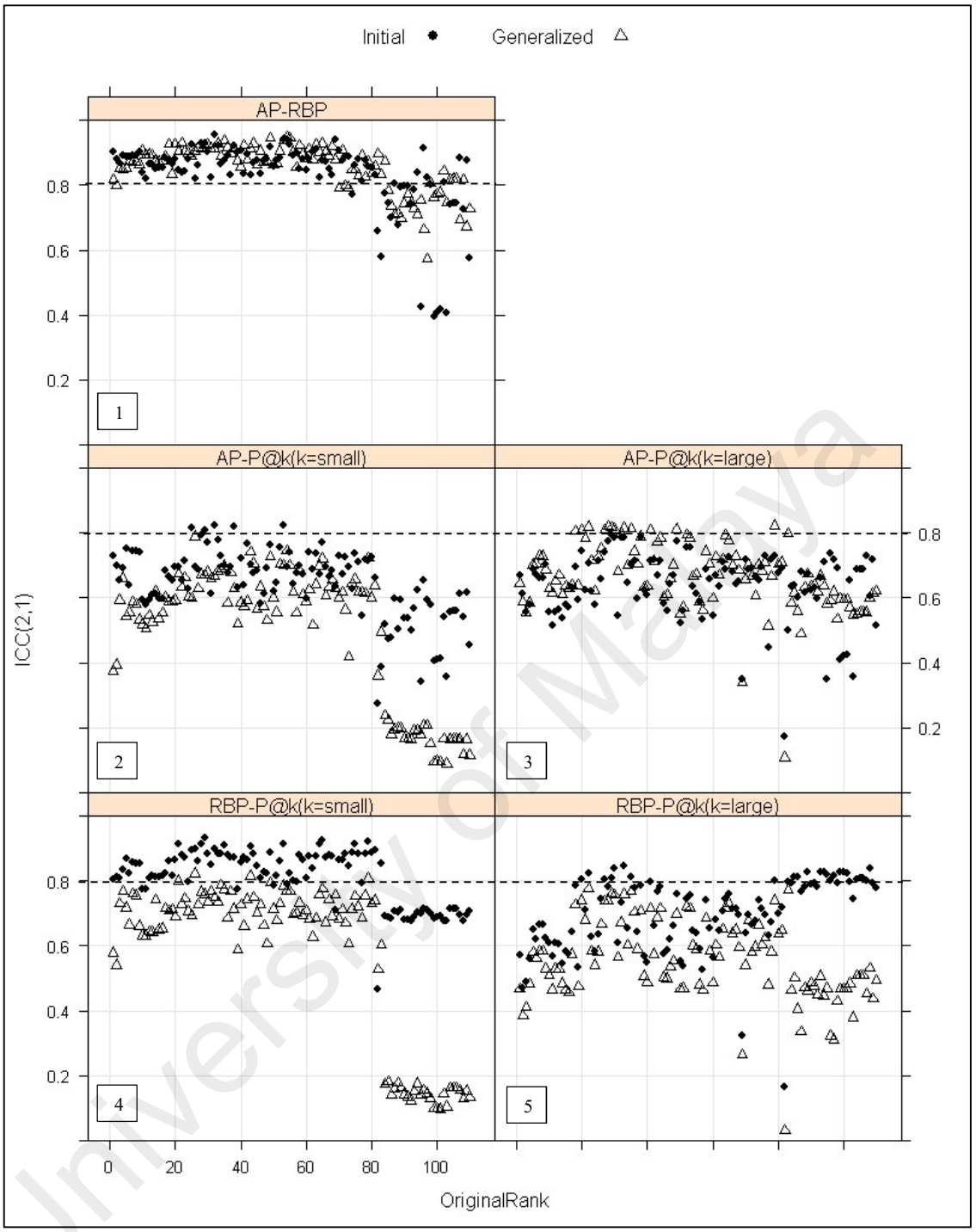

Figure 4.6: TREC-2004 Robust track ─ Average reliability scores against the original MAP system rankings for various metrics combinations

As for graphs 2 and 3, representing metrics AP-P@*k* (*k*=small) and AP-P@*k* (*k*=large) respectively, some middle rank systems are highly reliable. However, most systems ranging from top to bottom tend to have moderate ranking reliability. These results mean

a top ranked system might not necessarily perform consistently when compared to some of the middle ranked systems. A user may face inconsistency in retrieval results with these top ranked systems. Meanwhile, a user may observe better consistency in performance by some middle ranked systems. It is a trade a user has to decide about using top performing systems or consistent performing systems in regards to their ability in sustaining their ranking if the metrics used are AP and P@*k* (*k*=small, large).

Also, notice the bottom ranked systems (80 to 110) from graph 2. The reliability coefficient of the bottom ranked systems are moderate for initial metrics combination but it has become poor when using the generalized metrics combination. The similar occurs with graphs 4 and 5. These bottom ranked systems tend to have varied reliability when evaluated using different set of metrics of the same group since the generalization does not compare well with the initial metrics combination. The bottom ranked systems evaluated using metrics combinations AP-P@*k* (*k*=small), RBP-P@*k* (*k*=small), and RBP-P@*k* (*k*=large) show poor reliability.

The graph 4 also shows the top and middle ranked systems have strong reliability coefficient. They are able to sustain their rankings in relation to other systems. However, the generalization metrics measures these top and middle ranked systems as moderately reliable. As for graph 5, only a few middle-ranked systems have strong reliability scores. The bottom ranked systems also have strong reliability when the initial metrics combination is used. But their generalization does not show a similar pattern. Therefore, even this metrics combination may produce varied results when measuring the reliability of individual systems rankings.

The AP-RBP metrics combinations indicate top and middle ranked systems as having better reliability compared to the bottom ranked systems. The AP-P@*k* (*k*=small) metrics combination suggests top and middle ranked systems have better reliability compared to

the bottom ranked systems while a limited few middle ranked systems have better reliability. As for AP-P@*k* (*k*=large), there is no clear indication of which ranked systems are better in terms of their reliability. Then, metrics combination RBP-P@*k* (*k*=small) indicate top and middle ranked systems to be better in their reliability compared to bottom ranked systems. However, for RBP-P@*k* (*k*=large), the bottom ranked systems have better reliability compared to the other ranked systems.

In an attempt to understand the reliability of the systems from the original ranked systems, it appears that top and middle ranked systems are mostly reliable. The bottom ranked systems generally have poor reliability. The evaluation implies the capabilities of the top and middle ranked systems in producing consistent results. But the bottom ranked systems do not produce consistent results and may cause inconsistencies in retrieval results to the user.

Figure 4.7 shows the reliability scores from within metrics combinations against the original system ranks. The figure only includes graphs for topic size 30 as the remaining topic sizes are similar. The complete graphs with all topic sizes is available in Appendix F till I. For within cluster metrics combinations, mostly top-ranked systems have highly reliable system rankings. The results indicate the metrics within a population could produce a similar ranking for top performing systems. Some combinations of metrics show high ranking reliability for middle and low ranked systems. Therefore, reliable system rankings can be achieved by low ranked systems as well although their performance is not as good as other top performing systems.

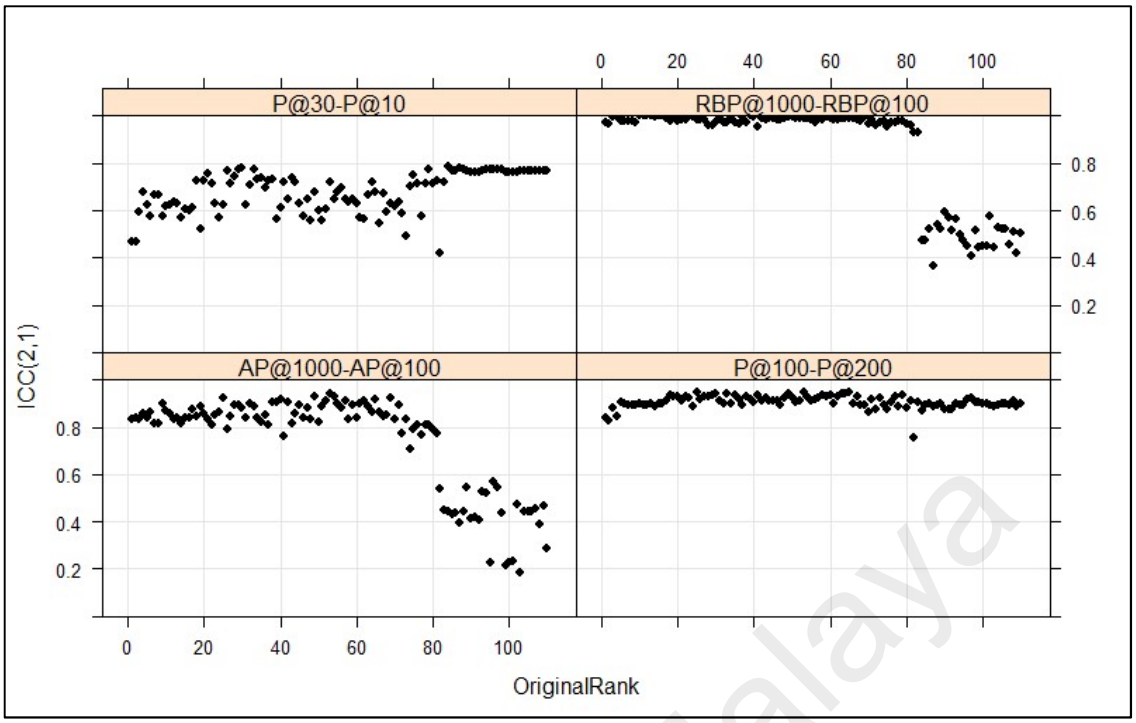

Figure 4.7: TREC-2004 Robust track – Reliability scores against the original MAP system rankings for various within cluster metrics combinations

4.4.4 Kendall's Tau Correlation Coefficient with Gold Standard

In the previous subsection, graphs were plotted to show the original system ranks and their possible reliability coefficient. However, the reliability scores from the proposed method represent a different set of ranking as per ICC computation (the mean rank). Therefore, to determine if the reliability scores can represent the original system rank, Kendall's tau correlation coefficient between the mean rank from the proposed method and original system rankings is performed to answer RQ4.

The evaluation outcome can assist in determining if these reliability coefficients truly represent the original system ranks. The Kendall's tau correlation uses the mean ranks calculated from the randomly selected topics for each system. The ICC(2,1) equation largely involves the use of mean rank to compute the $MSB_{targets}$, MSB_{judge} and SST to calculate the MSE. Due to the usage of mean rank in constructing the ANOVA

components, the mean rank is suitable as a representation of the proposed method's system ranks. Now, those mean ranks will be correlated to the original system ranks to determine the similarities of the mean ranks to the gold standard. The gold standard system ranks are obtained from the original MAP scores. If the correlation coefficient is strong, the mean ranks are close to the gold standard ranks. Thus, the reliability score can be accepted as a representation to the original system ranks.

Figure 4.8 and Figure 4.9 show the density plot of Kendall's tau correlation of system ranks between the proposed method (100 iterations) and the gold standard for metrics combination outside cluster and within cluster respectively. The plots within Figure 4.8 are arranged according to their initial (left) and generalization (right) metrics combination. The pair of plots is numbered from 1 to 5. There is a baseline Kendall's tau correlation coefficient between system ranks generated from MAP scores of 50 random topics and the original system ranks from MAP scores of 249 topics. This baseline density plot allows comparison of the shape and has a mean value of 0.88 while the standard deviation is 0.03.

Plots 1 from Figure 4.8 appears to have the same shape for the initial and generalization metrics combinations of AP-RBP. These metrics combinations also have strong mean correlation coefficient for most of the topic sizes except topic size 10. The ranks generated from the proposed method is mostly similar to that of the original system ranks. Hence, the reliability scores obtained from the proposed method can be accepted as the true representation of the original systems' reliability.

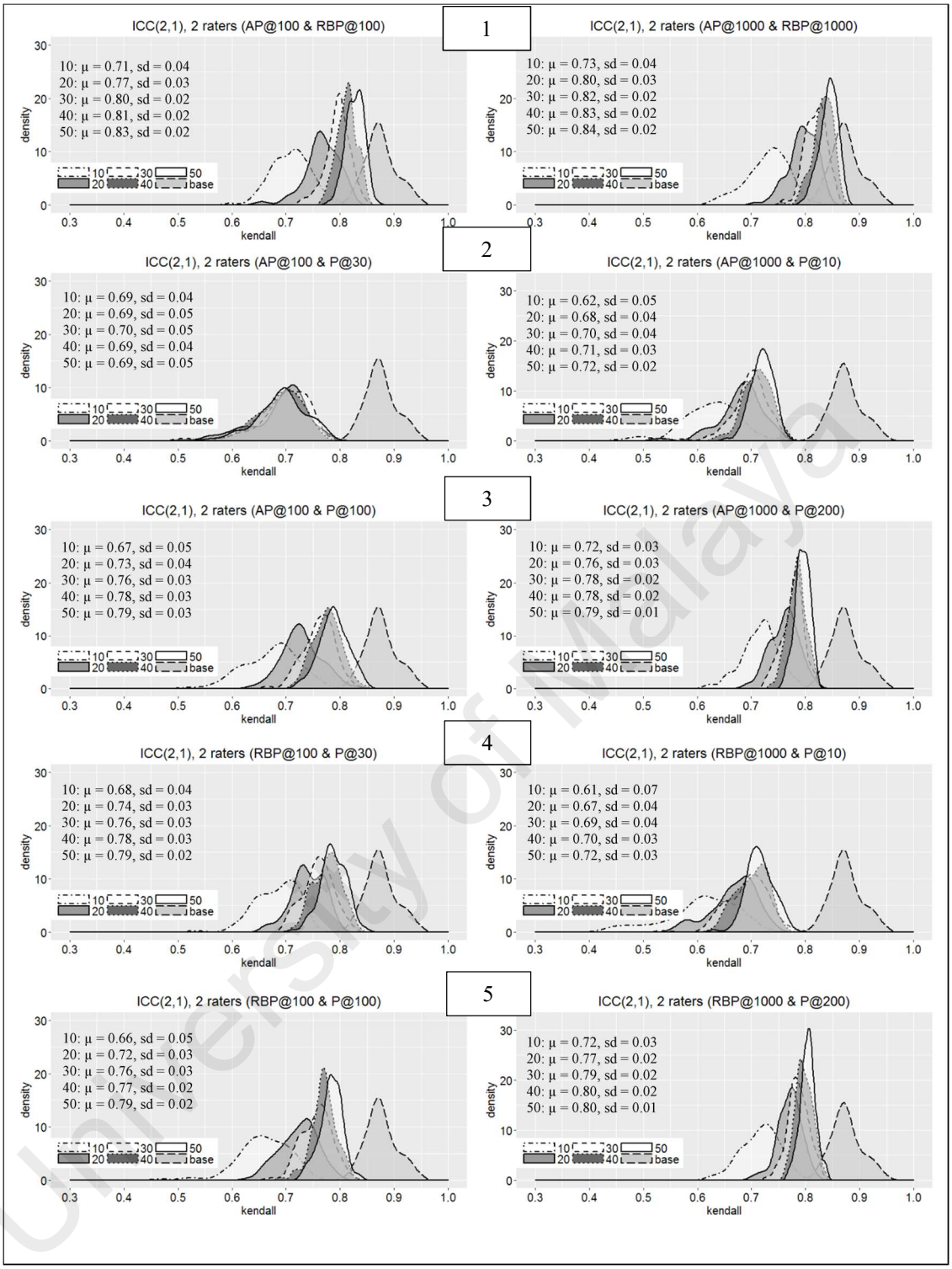

Figure 4.8: Density plot of Kendall's tau correlation coefficient between original system ranks and proposed method's system ranks for 100 iterations for outside group metrics combinations

Plots 2 combines AP and P@*k* (*k*=small), and shows the density plot is further away from the baseline plot. Their mean is moderate with larger standard deviations. The standard deviation spread shows the Kendall's tau correlation coefficient varies widely for the 100 iterations. Their reliability score representations of the original system rankings may be inaccurate due to smaller mean tau and larger standard deviations.

The plots 3, AP and P@*k* (*k*=large) has Kendall's tau correlation coefficient close to 0.8. The distribution is similar to the baseline as well. In addition, the standard deviation of the plots are narrower for the generalized metrics combination but slightly more spread out for the initial metrics. The reliability scores from these metrics combinations also appear to be a suitable representation of the original system ranks.

As for plots 4, RBP and P@*k* (*k*=small) have their mean closer to 0.8 for topic sizes 30 to 50 for the initial metrics combinations. The distribution of the Kendall's tau is similar to the baseline for both initial and generalized metrics except for topic size 10 in the generalized metrics. The shape of the density plot is flatter with a standard deviation of 0.07. But the mean Kendall's tau value for the generalized metric is not as good as the initial metrics. The reliability scores may not truly represent the original system ranks due to moderate mean tau and rather spread out standard deviation.

The plots 5, RBP and P@*k* (*k* = large) have mean values close to 0.8 for topic size 30 to 50 for both initial and generalized metrics combinations. Also, notice that their standard deviation is smaller than the baseline. It means the Kendall's tau values from the 100 iterations are close to the mean tau. Topic size 10 has smaller mean tau value and larger standard deviation, making it unsuitable for representing the reliability scores to the original system rank. Nevertheless, these metrics combinations tend to have moderately good mean tau and standard deviation for larger topic sizes. The reliability scores could still represent the original system ranks.

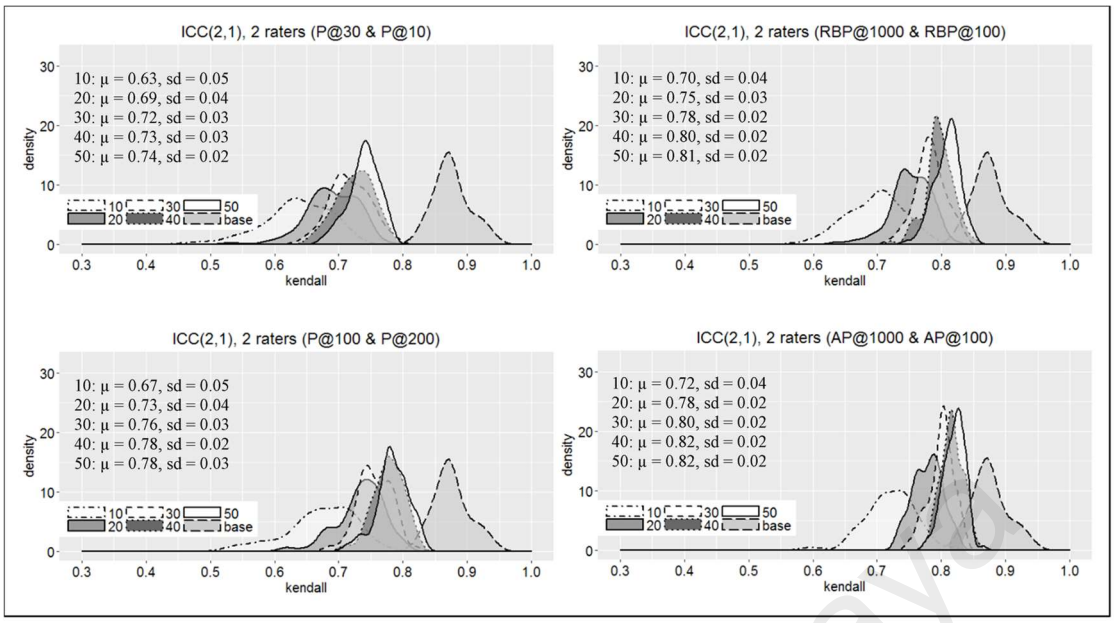

Figure 4.9: Density plot of Kendall's tau correlation coefficient between original system ranks and proposed method's system ranks for 100 iterations for within group metrics combinations

Density plots for metric combinations from the same group (Figure 4.9) show moderate to strong correlation coefficient with the original system rankings, and the shape of the plots are similar to the baseline plot as well. But, topic size 10 has the poorest mean tau compared to other topic sizes' mean tau. The standard deviation is also larger than the baseline, indicating a more varied Kendall's tau values from the 100 iterations. The effect of P@*k* (*k*=small) can be observed in the metric combination P@10 and P@30. The other metrics combinations tend to have mean tau that is close to 0.8 and narrow standard deviation, as good as and sometimes better than the baseline. The reliability scores from these metrics combinations is suitable representations of the original system ranks. But, the reliability scores from P@*k* (*k*=small) is not a suitable representation of the original system ranks.

Based on the experimental analyses, the mean tau is close to or above 0.8 with some metrics combinations. The standard deviations are also narrow. The reliability scores

from metrics combinations AP-RBP, AP-P@*k* (*k*=large), RBP-P@*k* (*k*=large) for outside cluster, and P@*k* (*k*=large), AP and RBP for within cluster can be accepted as a representation of the original system rankings.

4.4.5 Consistency of the Proposed Method in Measuring Individual Retrieval System Reliability

This study aimed at measuring the reliability of individual system rankings as opposed to the evaluation of a set of system rankings between a before and after experiment. The initial experiment measured reliability using ICC(2,1) using the ranks from two different metrics. The experimentation also explored various combinations of metrics pairs outside cluster and within the cluster. Based on the initial experimental result, an extension of the experiment is conducted on another test collection to determine the consistency of the proposed approach. The experiment is mainly to run a one-time evaluation of the systems' reliability.

The extension experiment is conducted with a perception that an individual is now interested to utilize the proposed method to understand the level of reliability of a retrieval system. Due to that, the experimentation will only select random topics once (not 100 times as in initial experimentation) to compute the reliability scores of the systems. Then, the number of systems that have high reliability (≥0.8) will be counted to determine if a similar trend is observed for the TREC-2005 Robust track.

Table 4.11 shows the number of highly reliable systems measured with ICC(2,1) using the topic ranks. The table also shows the various combination of metrics in pairs, the initial and the generalization. Since the test collection from TREC-2005 Robust track only contains 50 topics, the last column (50) in Table 4.11 is using all topics instead of random

selection. The remaining topics were randomly selected once to compute the reliability scores.

Table 4.11: TREC-2005 Robust track – Number of highly reliable systems measured using different pairs of metrics

Cluster	Topic size	10	20	30	40	50
Outside cluster	AP@100-RBP@100	55	41	45	42	50
	AP@1000-RBP@1000	47	45	39	36	34
	AP@100-P@30	37	23	27	23	25
	AP@1000-P@10	7	3	0	1	0
	AP@100-P@100	55	46	45	40	41
	AP@1000-P@200	53	56	53	47	49
	RBP@100-P@30	69	72	73	73	73
	RBP@1000-P@10	23	37	26	24	22
	RBP@100-P@100	47	36	50	56	49
	RBP@1000-P@200	20	26	20	18	11
Within cluster	P@30-P@10	12	14	16	12	10
	P@100-P@200	65	60	69	68	67
	AP@1000-AP@100	38	51	44	39	38
	RBP@1000-RBP@100	73	73	73	73	73

The TREC-2005 Robust track (dataset2) has 74 system runs that were evaluated. The metrics pairs AP-RBP and AP-P@*k* (*k*=large) from the outside cluster have more than 50% system runs that produce reliable rankings. The same results are obtained for within metrics combinations P@*k* (*k*=large), AP and RBP. The metrics combination AP-P@*k* (*k*=small) and P@*k* (*k*=small) have lesser than 50% system runs that have high reliability rankings. Again, the shallow evaluation depth using P@*k* is not suitable for measuring the reliability of individual system rankings. As for the RBP-P@*k* (*k*=small) and RBP-P@*k* (*k*=large) produces an inconsistent number of reliable system rankings. The initial metrics results in more number of systems that are reliable compared to the generalization metrics.

The within metrics results using this test collection is similar to that obtained from the TREC-2004 Robust track (dataset1) test collection. However, the outside cluster metrics combinations are varied between the test collections. The differences in results between the test collections could be due to the number of times random topics were selected. Dataset1 used an average of 100 iterations to compute the number of system runs with high reliability but dataset2 only used 1 random selection. Nevertheless, the result could also mean the system runs in the TREC-2005 Robust track has more consistent performance while fulfilling different user model compared to those in TREC-2004 Robust track.

The reliability results from initial metrics combinations are known. Now, a generalization metrics pair is also used to measure the reliability of the individual system rankings to determine how well do these results from the metrics combinations produce similar results with other pairs of metrics of the same cluster. Figure 4.10 shows the generalization results for measuring individual system rank reliability with other similar metrics combination. The figure contains five separate graphs labeled 1 to 5 for the outside cluster metrics combinations; AP-RBP, AP-P@*k* (*k*=small), AP-P@*k* (*k*=large), RBP-P@*k* (*k*=small), and RBP-P@*k* (*k*=large). Within each graph, there are 5 panels, one panel for each topic size. The x-axis represents the ICC(2,1) results for the initial metrics combinations and the y-axis represents the ICC(2,1) results for the generalization metrics combination. The dotted lines within the graphs and panels are marked at 0.4 and 0.8 on both axes. The correlation coefficient values below 0.4 are low in reliability, within 0.4 and 0.8 are moderately reliable, and those above 0.8 are highly reliable.

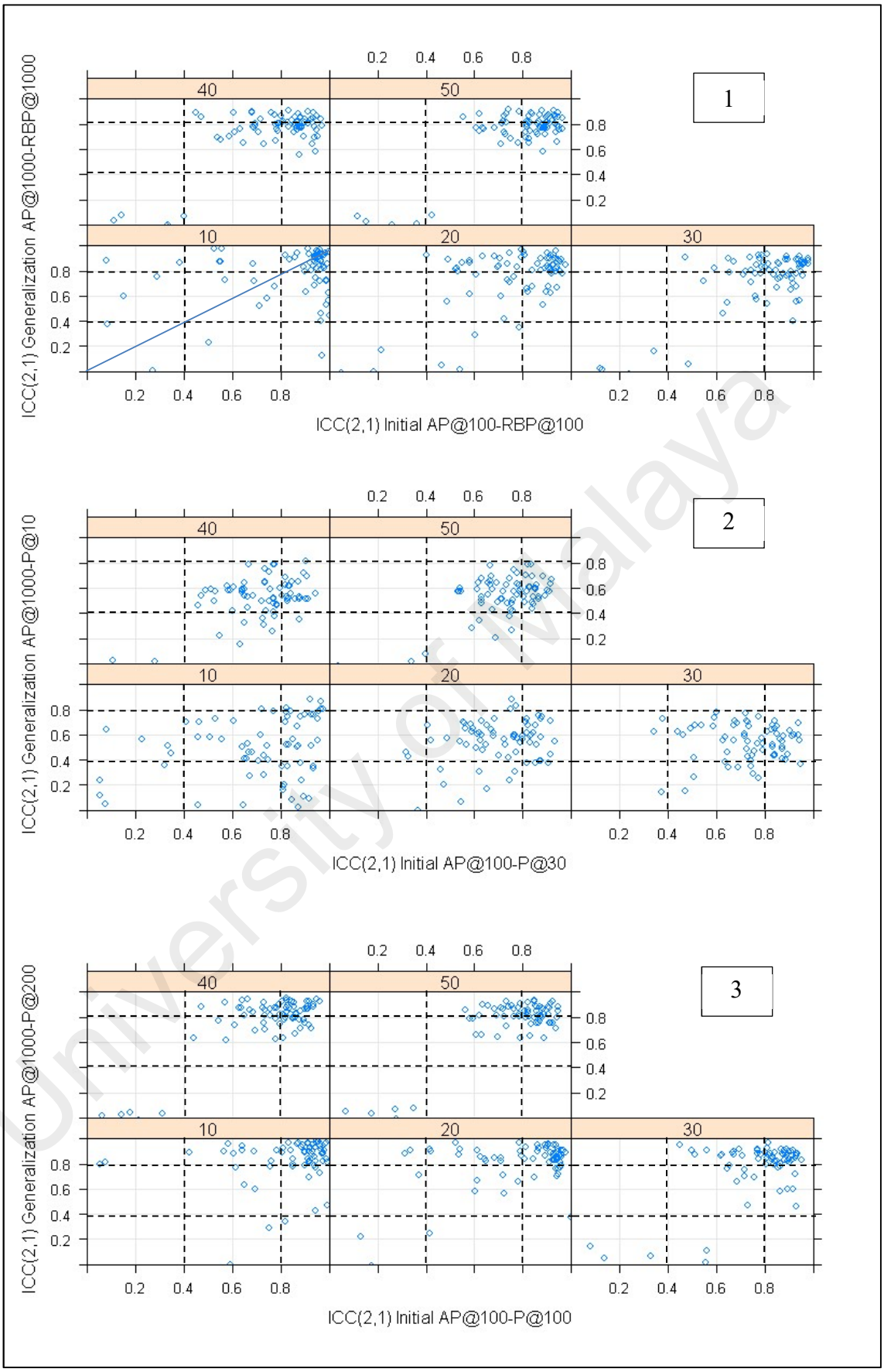

Figure 4.10: TREC-2005 Robust Track ─ Generalization results for various metrics combination

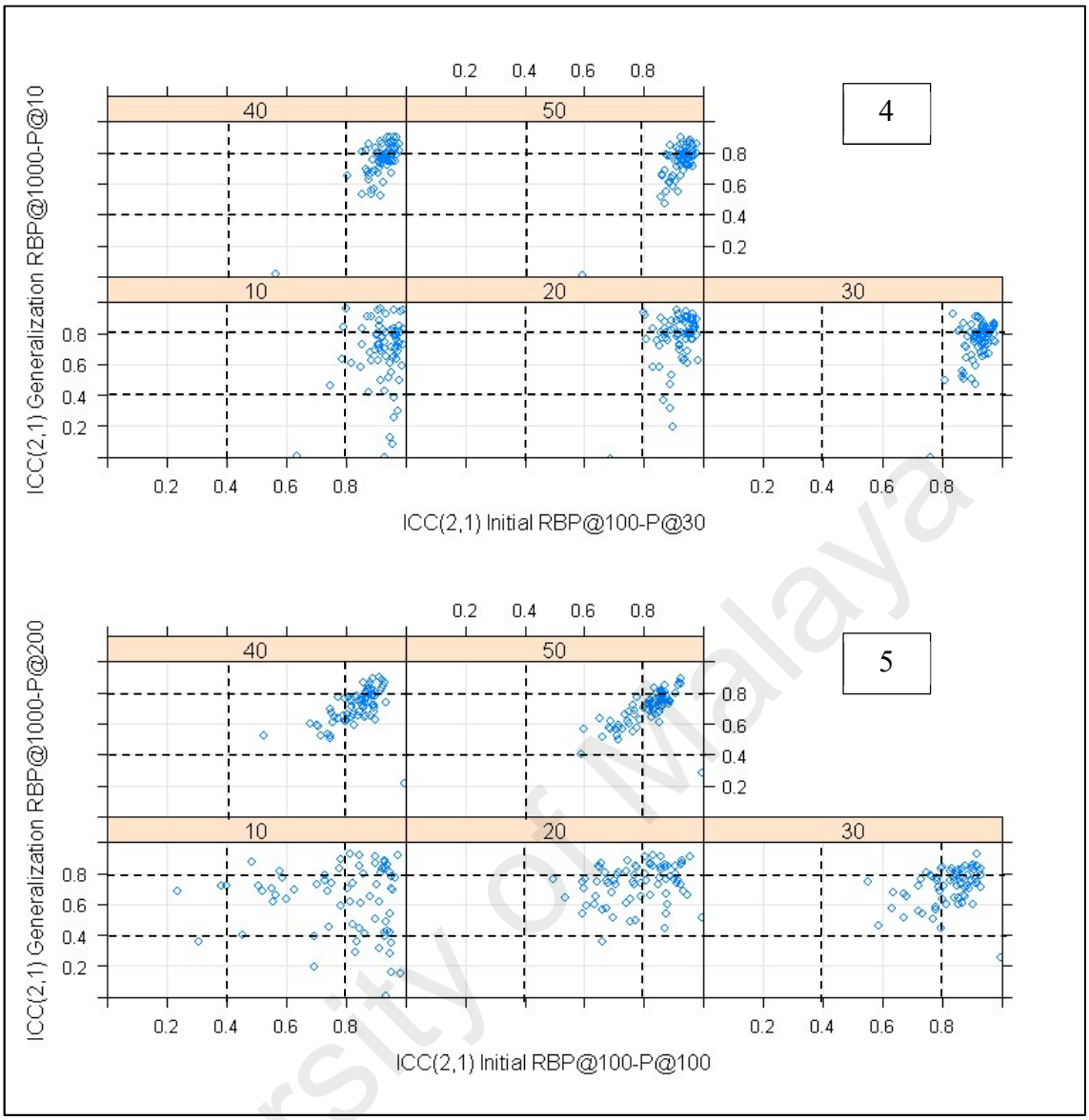

Figure 4.10, continued: TREC-2005 Robust Track ─ Generalization results for various metrics combination

If the plots fall within the same reliability coefficient for both initial and generalization metrics, the results would indicate the initial and generalized metrics combination produces results consistently. An easy way to recognize the region is to focus on the diagonal cubes formed from the dotted guidelines. If the plots fall within this diagonal region, the results from initial metrics are replicated to other similar metrics. If the plots fall at different regions, it means the generalization is not valid for other similar metrics

of the same cluster. Thus, the metrics combination is not a suitable option for measuring system ranking reliability due to possible variation in results.

From the figure, four out of five outside cluster metrics combinations generalize well to another set of similar metrics pairs. These metrics are AP-RBP, AP-P@*k* (*k*=small), AP-P@*k* (*k*=large) and RBP-P@*k* (*k*=large). The results produced by these metrics are similar between the initial and generalization. However, only AP-P@*k* (*k*=large) consistently generalize with highly reliable system rankings since the plots mostly fall within the diagonal region. The AP-RBP and RBP-P@*k* (*k*=large) generalize well for topic sizes 10 to 30 with highly and moderately reliable system rankings but not for the others. Notice that for topic sizes 40 and 50, more plots are dispersed outside the diagonal region.

The remaining pair of metrics, RBP-P@*k* (*k*=small) does not generalize well since the initial and another pair of similar metrics do not result similarly. Graph 4 clearly shows the plots do not fall within the diagonal region, which means they do not produce a similar outcome. Hence, poor generalization. The same metrics pair evaluation displayed poor generalization for the TREC-2004 Robust track.

Similar to the earlier test collection, the reliability scores from topic size 30 are matched with the original system rankings using MAP scores from 50 topics. The purpose is to identify which ranked systems from the original have strong, moderate or poor reliability. Figure 4.11 shows the reliability scores of the initial and generalized metrics pairs against the original system rankings.

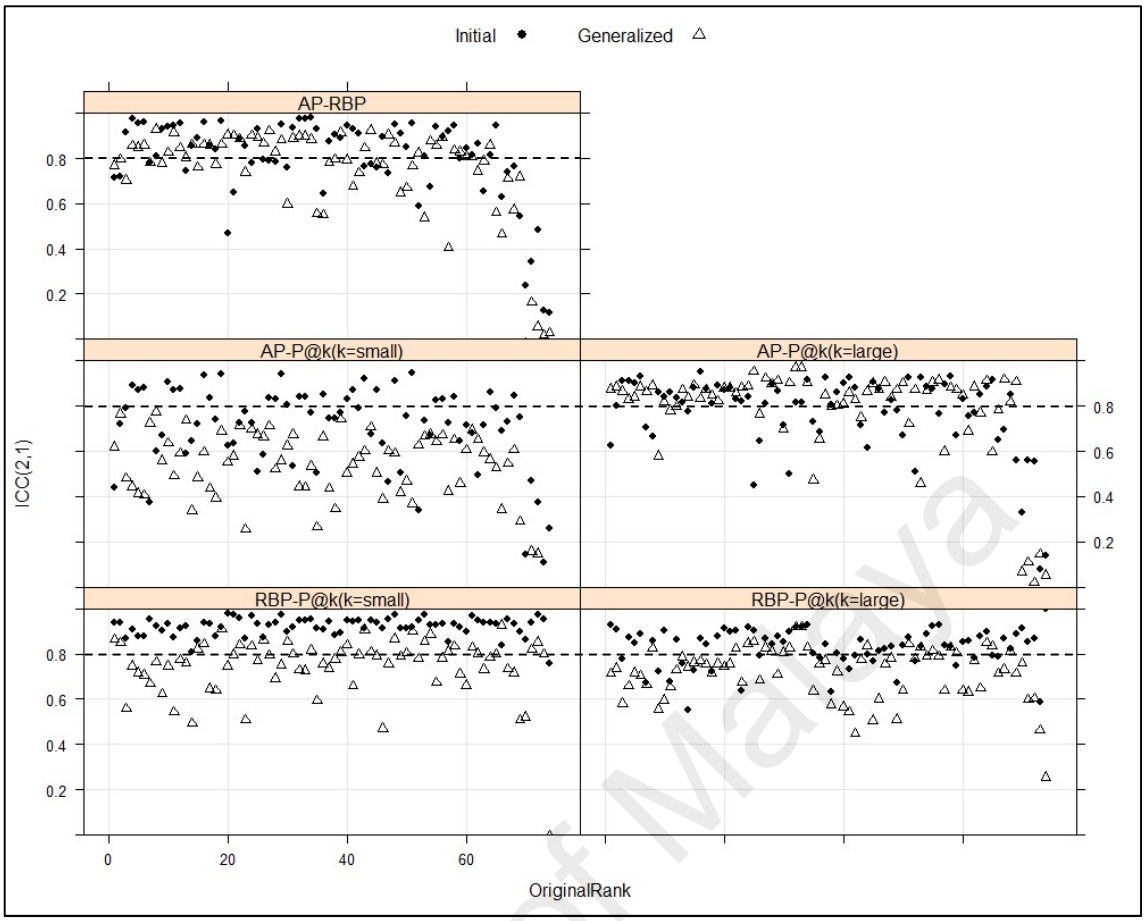

Figure 4.11: TREC-2005 Robust track ─ Reliability scores against the original MAP system rankings for various metrics combinations

All the graphs indicate top to middle ranked systems having high or moderate reliability scores. Some of the bottom ranked systems show poor ranking reliability for metrics pairs AP-RBP, AP-P@*k* (*k*=large) and RBP-P@*k* (*k*=large). There is no clear pattern as to which ranked systems have better reliability. Since the plots are dispersed vertically without specific connection with the system ranks. The outcome could have occurred because of consistent systems in the test collection regardless of their rankings. These patterns imply the systems from TREC-2005 Robust track may satisfy users better than those in TREC-2004 Robust track test collections.

For dataset2, Kendall's tau correlation coefficient was measured between the system ranks of the proposed method and the gold standard for metrics combination outside cluster and within cluster respectively. The proposed method system ranks were obtained from the mean rank, and tied ranks were resolved using their reliability scores, similar to dataset1.

Table 4.12 shows Kendall's tau correlation for the randomly selected topics sizes (topic size 50 uses all available topics for the collection). The tau values are mostly above 0.6, parallel to dataset1. Only 6 of the tau values in dataset2 are below 0.6. The Kendall's tau values are mostly moderate while some strong correlation coefficients were observed when the topic sizes are above 30. Then again, the tau values shown in Table 4.12 are from single random selection. There are possibilities of obtaining better or worse correlation coefficients with other random topic selections. Nonetheless, both datasets show comparable results for the proposed method.

Table 4.12: TREC-2005 Robust track – Kendall's tau correlation coefficient between mean rank of proposed method with original MAP system rank

	Topic size	**10**	**20**	**30**	**40**	**50**
Outside cluster	AP@100-RBP@100	0.67	0.71	0.75	0.76	0.76
	AP@1000-RBP@1000	0.69	0.79	0.77	0.77	0.81
	AP@100-P@30	0.68	0.58	0.74	0.72	0.76
	AP@1000-P@10	0.57	0.72	0.68	0.71	0.74
	AP@100-P@100	0.46	0.78	0.74	0.80	0.81
	AP@1000-P@200	0.71	0.75	0.81	0.85	0.83
	RBP@100-P@30	0.70	0.60	0.60	0.67	0.73
	RBP1000-P@10	0.51	0.63	0.65	0.67	0.65
	RBP100-P@100	0.78	0.65	0.79	0.83	0.82
	RBP@1000-P@200	0.64	0.81	0.79	0.78	0.82
Within cluster	P@30-P@10	0.51	0.57	0.56	0.65	0.65
	P@100-P@200	0.66	0.71	0.80	0.82	0.85
	AP@1000-AP@100	0.42	0.71	0.73	0.81	0.82
	RBP@1000-RBP@100	0.63	0.68	0.70	0.72	0.72

4.5 Summary

A retrieval system may be effective when measured using certain effectiveness metrics but the reliability of the individual system ranking is not known through the evaluation such as Kendall's tau correlation coefficient. A reliable retrieval system is crucial in satisfying users' need. This study fulfilled OBJ1 and proved experimentally that reliability of individual system rankings can be measured using intraclass correlation coefficient approach. The study suggests the use of metrics, AP and RBP to measure the reliability of individual system ranks. The metrics pair when combined or used individually with different depth of evaluation identifies highly reliable system ranks. It means the retrieval systems are equally reliable when measured using AP and RBP.

As for OBJ2, the various metrics combinations generalize well except RBP-P@*k* (*k*=small). Reliability of individual retrieval system rankings could be consistently measured using topic size 30 to achieve a sufficiently large number of reliable systems and equally good generalization. Smaller topic sizes reduced the amount of effort needed for relevance judgments and yet retain the quality of the approach.

The OBJ3 is met through the visual plotting of original system ranks and the reliability coefficients from ICC, and reveals that top and middle ranked systems have better reliability compared to bottom ranked systems. Therefore, a user can be confident in receiving good retrieval results most of the time from these ranked systems. Again, topic sizes 30 or more results in strong correlation coefficient. A strong positive correlation coefficient indicates the true representation of the reliability score to the gold standard system ranks and thus fulfilling OBJ4. Meanwhile, users can benefit from this retrieval systems due to their consistent performance.

However, P@*k* (*k*=small) should not be used to measure reliability of individual system rankings as it does not generalize well and produce inconsistent retrieval results. It also fails to represent the reliability coefficient to the original system rankings. The shallow depth of evaluation has caused variations in rankings for the topics, such that reliability measure is poor.

The proposed method is a reliable approach to measuring the reliability of individual retrieval systems using their ranks. The approach is capable of determining the versatility of the systems in satisfying multiple user needs. The study highlights the combination of effectiveness metrics which are suitable for measuring the reliability of ranking in information retrieval systems. Also, the reliability score is measured for individual systems from their topic ranks but the evaluation is dependable on relative ranking data among the systems. Nevertheless, ranking only makes sense if it is relatively measured.

The study can be easily replicated and suited to other test collections for it utilizes relative ranking in measuring reliability. There lies an opportunity to investigate more than two metrics combination that would determine the diversity of a system's performance in satisfying multiple user need.

CHAPTER 5: DOCUMENT LEVEL ASSESSMENT IN A PAIRWISE SYSTEM EVALUATION

This chapter covers the details of an experimentation focusing on identifying an alternative way to perform statistical significance testing without the use of averaged or cut-off topic scores. Section 5.1 details the background, problem statement, research questions and the objectives of this experimentation. Section 5.2 details the literature review on the various metrics, statistical significance tests and the past work related to this experiment. Section 5.3 details the selection of test collection, fulfillment of the dependent test criteria, p-values aggregation, the proposed method using document-level assessment, and the steps undertaken to conduct the experiment. Section 5.4 consists of the results obtained from the experimentation and discussions pertaining the results. Lastly, the summary of the chapter follows.

5.1 Background

The main focus of the retrieval system is to provide information as accurate as possible based on the user's query and that are relevant to the user. Information on the Web grows continuously making it impossible to access the information without the help of search engines (Tsytsarau & Palpanas, 2012) or retrieval systems. When information aggregation is from multiple sources such as virtual documents (Watters, 1999), retrieval gets tough, while solutions such as sentiment analysis and opinion mining need to be incorporated in traditional retrieval systems (Tsytsarau & Palpanas, 2012).

In the system-oriented evaluation, the effectiveness of information retrieval systems is measured using relevant documents obtained by the retrieval systems to meet users' queries. The relevancy of the retrieved documents is unknown until they are assessed or judged by experts or crowd-sourced judges. However, the retrieval system itself should be evaluated to determine its performance. System effectiveness can be evaluated by using evaluation measures, including—but not limited to—the following: precision at cut-off k ($P@k$) (Joachims, Granka, Pan, Hembrooke, & Gay, 2005), average precision (AP) (Buckley & Voorhees, 2000), normalized discounted cumulative gain (NDCG) (Järvelin & Kekäläinen, 2002) and rank-biased precision (RBP) (Moffat & Zobel, 2008).

Ranking the retrieval systems based on their effectiveness scores allow us to determine the superiority of the performance of one system over those of other systems. There may be only slight difference in the effectiveness scores between the systems in determining their ranking. An alternative way to determine the true difference in the performances of the systems is to do pairwise system evaluation.

Figure 5.1 shows an example of IR system evaluation using effectiveness scores in a pairwise system evaluation. The example demonstrates the evaluation of three different systems. An evaluation metric is used to measure the effectiveness of each system. Based on these effectiveness scores, the systems can be ranked. The comparison using effectiveness scores indicate System B is better than System C, and system C is better than System A. System B is ranked 1, System C is ranked 2, and System A is ranked 3.

Although these systems are ranked, their differences could have occurred by chance. Significance testing can measure their true differences. Significance tests are done to investigate whether the observed differences between pairs of systems are likely to be intrinsic or by chance. In Figure 5.1, a one-sided significance test determines if one system is truly better than the other system. A typical one-sided hypothesis suggests the

two systems are equally good while the alternate hypotheses could indicate one system is better than the other system.

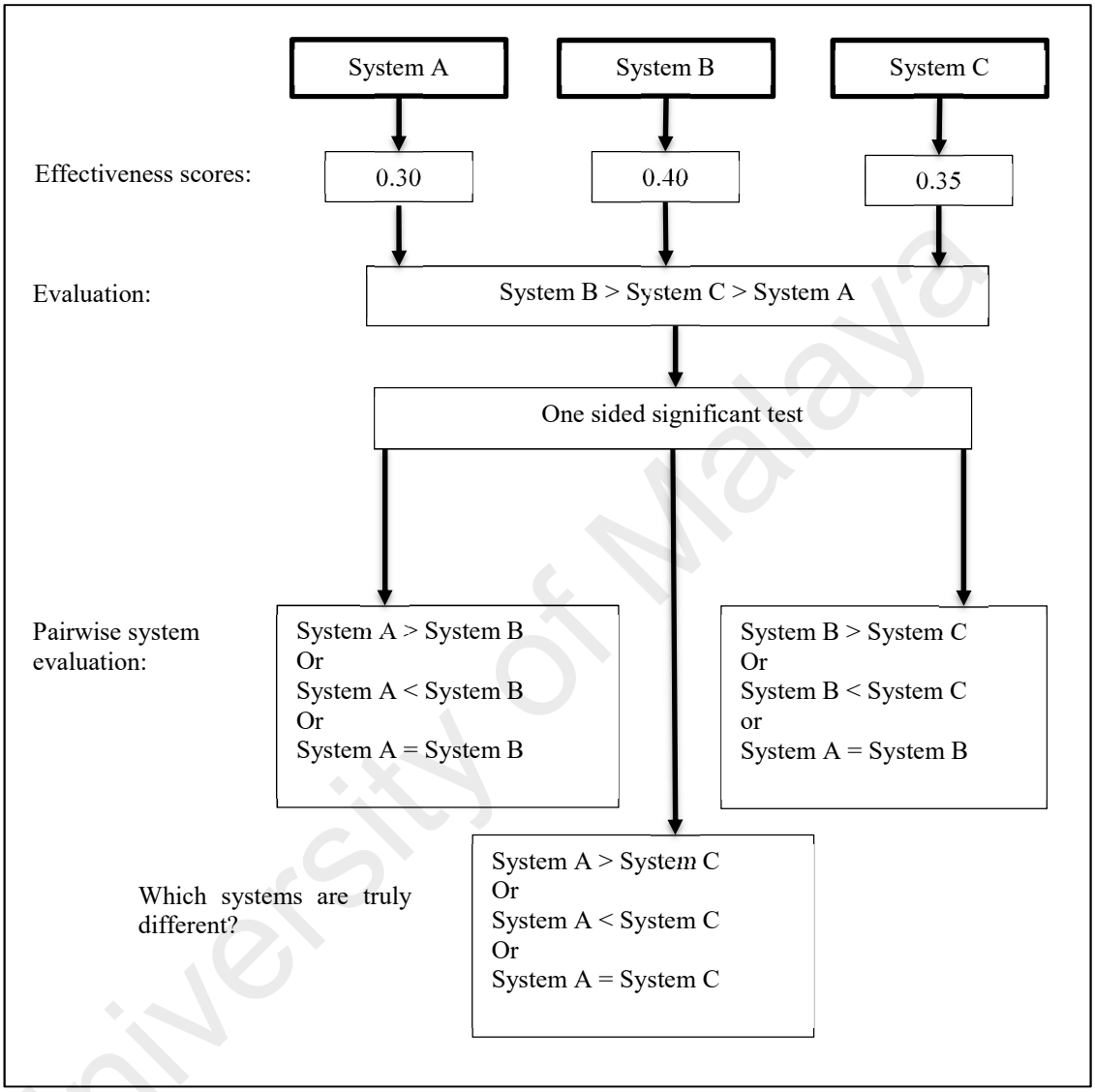

Figure 5.1: Evaluating retrieval systems with significance testing

For example, a dependent t-test or paired t-test compares the means between two related samples on the same continuous, dependent variable ("Dependent T-Test using SPSS," 2015). A dependent variable is measurable on a continuous scale such as interval or ratio. Table 5.1 shows an example of mean difference calculation for a paired t-test

between the topics of two systems. The mean difference, together with standard deviation will be used to calculate the *t* statistics and the corresponding *p*-value.

Table 5.1: Example of paired t-test mean difference calculation

Topic	System A score (P@10)	System B score (P@10)	Difference between system A & system B scores
1	0.2	0.5	-0.3
2	0.7	0.6	0.1
3	0.2	0.2	0
4	0.5	0.3	0.2
5	0.6	0.7	-0.1
Mean	0.44	0.46	-0.02

The average topic scores using mean average precision (MAP) from fifty topics are used to quantify the overall system effectiveness in TREC (Smucker et al., 2007, 2009; Urbano et al., 2013b). It is also common to use the effectiveness scores of individual topics as the unit of measurement in statistical testing (Jayasinghe, Webber, Sanderson, Dharmasena, & Culpepper, 2015; Robertson & Kanoulas, 2012; Scholer, Moffat, & Thomas, 2013). These topic scores are usually average or cut-off scores such as average precision (AP) (Smucker et al., 2007), expected reciprocal rank (ERR@20) (Dinçer, Macdonald, & Ounis, 2014) and P@*k* (Lewandowski, 2008; Sanderson & Zobel, 2005).

When closely examined, although the topic scores could be suitable in representing the performance of systems, measuring the statistical significance could produce inaccurate results. The P@*k* metric provides diverse user experience depending on the number of relevant documents within a specific cut-off rank. For example, let's assume a system is having two relevant documents ranked 1 and 2, the P@10 score of 0.2. Another system having two relevant documents ranked 9 and 10 also has a P@10 score of 0.2. These scores are similar to the observations for Topic3 in Table 5.1. The system which

retrieved and ranked the relevant document earlier is not identified as a better system, due to the use of topic score. The difference in the relevant documents' ranking is not considered in P@*k*. If these P@*k* scores are used in measuring the significant difference between the pair of systems, the mean difference for Topic3 is 0. The result obtained could be distorted for not taking into account the variation in relevant documents' ranking in the statistical test.

As stated earlier, AP is also a common metric to represent a system performance from each topic and MAP is the mean of AP. However, the AP requires a total relevant document per topic among the compared systems to measure the topic effectiveness of a system. In real Web, it is not feasible to determine the total relevant documents per topic due to the constantly changing nature of the retrieval systems. Conversely, it is possible in a laboratory experiment whereby the retrieved documents are no longer changing over time. As such, measuring the significant difference between two Web retrieval systems will be a challenge with the usage of AP scores.

An alternate approach to overcome these negatives is needed to assess and evaluate not only the systems in a laboratory environment but also an approach suitable for use in real Web. An approach that is not affected by the diverse user experience at specific cut-off rank, and not dependent on the number of relevant documents.

5.1.1 Problem Statement

The metric precision at a cut-off rank *k* (P@*k*) gives a differing user experience depending on the number of relevant documents within a specific cut-off rank (Webber et al., 2010). It is a drawback measuring the significant difference between a system pair with their P@*k* performance scores which could produce incorrect results despite the

difference observed in the ranked relevant documents (As mentioned in the example from Table 5.1).

On the other hand, the average precision is a widely utilized metric in measuring information retrieval system performance. However, the system performance score can only be measured with a known total relevant documents for a topic. It is feasible to obtain the total relevant documents per topic from the relevance judgment from static test collections. Conversely, in real Web experience, the total relevant documents for a topic is unknown. The actual Web is constantly changing, making it difficult to obtain a static total relevant document per topic for existing or new systems under evaluation. Due to this downside, it is not suitable to evaluate the significant difference between systems in a real Web environment.

From the mentioned drawbacks, the usage of these averaged, or cut-off topic scores are not suitable for statistical significance testing. Although the use of averaged or cut-off scores is common in statistical significance testing, the results could be inaccurate in determining the true difference between systems.

In addition, different sample sizes used in the statistical tests tend to produce varying outcome. The use of 50 topics (Smucker et al., 2007) and 10 topics (Smucker et al., 2009) in statistical tests recommend the use of Student's t-test. However, disagreements among the statistical tests increased with 10 topics. Therefore, it is crucial to explore different sample sizes in statistical tests and analyze the results before conclusions can be drawn.

5.1.2 Research Questions

RQ1. What is an alternate approach to measure statistically significant system pairs to overcome the drawbacks of using averaged or cut-off topic scores?

RQ2. What is the suitable sample size for statistical testing using the alternate approach to achieve reliable results?

RQ3. Can the alternate approach effectively measure statistically significant system pairs?

5.1.3 Objectives

This study aims

OBJ1. To propose an approach suitable for evaluating retrieval systems by overcoming the drawbacks of inaccuracy using averaged or cut-off topic scores in statistical significance test.

OBJ2. To identify a suitable sample size in pairwise retrieval systems evaluation using the proposed approach that produces reliable results.

OBJ3. To validate the effectiveness of the proposed method in measuring statistically significant system pairs.

5.2 Literature Review

This section describes the various effectiveness metrics to evaluate the retrieval systems, the details on statistical significance testing to determine the true difference between pairs of systems, and the related past studies.

5.2.1 Effectiveness Metrics

Metrics can be utilized to measure the effectiveness or performance of retrieval systems according to its underlying user model or probabilistic measurement. A metric is needed to assess the successful completion or the computable abstraction of a task for a particular system (Moffat & Zobel, 2008). There are various metrics which could be used to measure the effectiveness of retrieval systems. The effectiveness metrics selected for use in this experimentation are precision, average precision, and rank-biased precision.

The precision and recall are two basic information retrieval evaluation metrics. The precision measures the fraction of documents that are relevant to the query among all the returned documents, whereas recall measures the ratio of relevant items retrieved to all relevant items in the file. The performance of the systems can be measured with the retrieved documents, evaluation metrics, and relevance judgment. Due to a large number of documents, TREC incorporates pooling before generating the relevance judgments. A system that returns relevant documents earlier in the retrieval process will have a better performance compared with a system that retrieves relevant documents later or at lower rankings.

The effectiveness of a system could also be measured using precision at a certain cut-off rank k ($P@k$), in which the total number of relevant documents is not required. However, among the commonly used evaluation measures, this is the least stable (Manning et al., 2009; Webber et al., 2010) and does not average very well across topics. The $P@k$ is regarded as an unstable measure because small changes in the ranking can cause significant influence in the score, and big variations in the ranking can cause no difference in the score. Another reason for regarding $P@k$ as unstable is that a constant cut-off represents widely varying user experiences depending on the number of relevant documents for the query.

Average precision (AP) is computed by averaging the precision scores of each document per topic. The irrelevant documents contribute to lower the precision scores at the ranks of the relevant documents. These irrelevant documents contribute a score of 0, instigating to reduce the effectiveness rating. For example, a topic has the following ranked documents' relevancy, {R R R NR NR}. Up to rank 3, the average precision is 1. At rank 5, the two non-relevant documents contribute to lower the average precision to 0.6.

The AP is top-weighted because a relevant document in position 1 contributes more to the effectiveness score than one at position 2 and so on down the ranking. The top-weightiness is considered an advantage of AP (Sakai & Kando, 2008). Also, stability and discriminative power of AP are considered as one of the best after normalized discounted cumulative gain (Shi, Tan, Zhu, & Wu, 2013). Average precision is also widely accepted in TREC (Webber et al., 2010) and is a commonly used metric for system effectiveness (Robertson, Kanoulas, & Yilmaz, 2010). However, AP computation requires the total number of relevant documents (refer to Equation 2.4 in Section 2.2.4.2) which are not readily available in real Web experience.

On the other hand, rank-biased precision (RBP) is a rank-sensitive metric that uses parameter p as a measure of user persistence. Persistence is the probability that a user, having reached any given point in the ranked document list returned by a system, will proceed to the next rank (Moffat & Zobel, 2008). When $p = 0.0$, it is assumed that the user is either satisfied or dissatisfied with the top-ranked document and would not look further down the list of retrieved documents. RBP measure assumes that the user would look through many documents before ending the search task as p approaches 1.0. The RBP@k metric is monotonic as the depth of evaluation k is increased (Moffat et al., 2012). The irrelevant documents also contribute 0 scores in RBP computation similar to average

precision. The RBP score at a specific cut-off rank is obtained by adding all the individual relevant document scores up to the cut-off rank.

An example of RBP and RBP@*k* calculation is shown in Table 5.2 for persistence value of 0.8. The RBP equation is shown in Section 2.2.4.3. The irrelevant documents contribute 0 scores to the RBP@5 while the relevant documents each have their respective document scores. Notice that although there is an irrelevant document at rank 2 for system B, the document score at rank 3 is similar to that of rank 3 scores of system A. Meanwhile, the RBP@5 shows the difference in the effectiveness scores between the two systems. Based on this characteristic of RBP, it appears that statistical significance testing using RBP cut-off scores might have different effect than P@*k* and average precision. Nonetheless, the RBP@*k* metric will be experimented within this study. These persistence values, $p = 0.95$ and $p = 0.8$ are common in the information retrieval field (Jones, Thomas, Scholer, & Sanderson, 2015; Shokouhi, Craswell, & Robertson, 2009; Webber, Moffat, Zobel, & Sakai, 2008).

Table 5.2: Example of RBP and RBP@5 calculation for persistence 0.8

Document rank	Relevancy system A	Contribution System A	Relevancy system B	Contribution System B
1	1	0.2	1	0.2
2	1	0.16	0	0
3	1	0.128	1	0.128
4	1	0.1024	0	0
5	1	0.08192	1	0.08192
RBP@5		0.67232		0.40992

5.2.2 Significance Testing

This experimentation is tackling the issue of inaccuracy in statistical significance test using average or cut-off scores. However, various statistical significance tests may be suitable in different scenarios. The three commonly used statistical significance tests are Student's paired t-test, Wilcoxon signed-rank test, and Sign test.

A significance test is a statistical method based on experimental data that aims at testing a hypothesis (Kulinskaya et al., 2014). It shows whether the outcome attained between a pair of system could have arisen by chance rather than intrinsically. The statistical test also shows confidence in the results obtained (Baccini et al., 2012). A previous study suggested to discontinue the utilization of the Wilcoxon signed-rank test and Sign test for measuring the significance of system means due to their poor ability to detect significance (Smucker et al., 2007). They also tend to steer toward false detection of significance.

The most widespread method of determining the significant difference is through the *p*-value, by which a specific null hypothesis can be rejected. The *p*-value is the probability of obtaining equivalent or more evidence against the null hypothesis with the assumption that the null hypothesis is true (Fisher, 1995). A *p*-value larger than 0.1 is not small enough to be significant, a *p*-value as small as 0.05 can seldom be disregarded, and a *p*-value less than 0.01 indicates it is highly unlikely to occur by chance (Fisher, 1995). The possibilities of obtaining a significant result are higher with the use of more data in significance testing.

In this hypothesis tests, the relationship between 2 systems' effectiveness is measured. Both system pair's effectiveness is thought to be equal, where the null hypothesis is defined as H_0: A=B. Two-sided *p*-values do not provide directions of deviations from H_0,

but a one-sided *p*-value is directional. Under the null hypothesis, the density of a *p*-value from a continuously distributed test statistic is uniform on the interval, 0 to 1.

A dependent t-test or Student's paired t-test compares the means between two related samples on the same continuous, dependent variable. The below assumptions are associated with dependent or paired t-test ("Dependent T-Test using SPSS," 2015) and needs to be met before experimenting.

1. Dependent variable is measurable on a continuous scale such as interval or ratio.
2. Independent variable should consist of related groups or matched pairs. Related groups indicate the same subjects are present in both groups.
3. The distribution of the differences in the dependent variable between the two related groups should be approximately normally distributed.

When using dependent or paired t-test, independence of observations does not apply as opposed to the independent t-test. Independence of observation means there is no relationship between the observations between the groups ("Independent t-test in SPSS Statistics | Laerd Statistics," 2017). Mean of differences is used when testing with dependent sample t-test. The 95% confidence interval is derived from the difference between the two sets of paired observations (t-test (paired and unpaired), 2015).

In the paired t-test, *t* statistics is defined as

Equation 5.1

$$t = \frac{\bar{x}_d}{s/\sqrt{n}}$$

Where *n* is the sample size, $\bar{x}_d$ is the mean difference of the *n* observations, and *s* is the sample difference standard deviation computed using Equation 5.2

Equation 5.2

$$s = \sqrt{\frac{1}{n-1}\sum(x_d - \overline{x_d})^2}$$

The computed *t* statistics are then compared to an appropriate *p* threshold determined by the Student's *t* distribution. An example of computing the t-statistics is shown below for data from Table 5.3. The mean calculation for each column is the last row of the table.

Table 5.3: Example of t-statistics calculation and determining the significant difference for a specific p-value

System	x_1	x_2	Difference (x_1-x_2), x_d	$(x_d - \overline{x_d})^2$
1	0.5	0.4	+0.1	0.0049
2	0.7	0.3	+0.4	0.1369
3	0.2	0.3	-0.1	0.0169
4	0.7	0.7	0	0.0009
5	0.9	0.9	0	0.0009
6	0.6	0.8	-0.2	0.0529
7	0.4	0.3	+0.1	0.0049
8	0.7	0.6	+0.1	0.0049
9	0.6	0.7	-0.1	0.0169
10	0.1	0.1	0	0.0009
Mean	0.54 ($\overline{x_1}$)	0.51 ($\overline{x_2}$)	0.03 ($\overline{x_d}$)	Sum = 0.24

Firstly, calculate the sample difference standard deviation.

$$s = \sqrt{\frac{1}{n-1}\sum(x_d - \overline{x_d})^2} = \sqrt{\frac{1}{10-1} \times 0.24} = \sqrt{0.027} = 0.16$$

Then, calculate the t-statistics as below.

$$t = \frac{\bar{x}_d}{s/\sqrt{n}} = \frac{0.03}{0.16/\sqrt{10}} = \frac{0.03}{0.05} = 0.6$$

To determine if the paired t-test is significantly different (1-tailed test, df = 9, $p < 0.01$), the critical t-value is obtained from the T Distribution chart (Appendix M). If calculated t-value (0.6) is smaller than the critical t-value (2.821), accept the null hypothesis. Therefore, the system pair is not significantly different at p-value < 0.01.

In the existing methods, the significance tests use the average scores such as AP or MAP scores (Aslam, Pavlu, & Yilmaz, 2006; Sanderson & Zobel, 2005; Smucker et al., 2007, 2009), and cut-off scores such as P@*k* (Lewandowski, 2008; Sanderson & Zobel, 2005) to measure the actual difference observed between a pair of system. These averaged, or cut-off scores will be used to measure the mean difference between the system pair. The TREC test collection can be easily fitted into the paired t-test assumptions.

A test collection from TREC have the same topics and document corpus for each retrieval systems. The independent variable is a match between the paired systems, which are the topics. The dependent variable is the effectiveness scores for each topic. These topic scores are usually measurable in ratio scale. The distribution of differences between the topic scores is assumed to be approximately normally distributed, suggesting that parametric Student's t-test is suitable for testing the significance of paired systems. Meanwhile, the use of either parametric or nonparametric statistical tests has little impact because both evaluations result in the same conclusions (Sheskin, 2011).

The fulfillment of the paired t-test assumptions for the existing method has been detailed above while the fulfillment the assumptions for the proposed method will be explained in the Methodology section.

5.2.3 Previous Related Studies

Statistical significant tests are uncommon in the information retrieval evaluation. In this literature review section, some of the related studies are detailed. A previous study has used the mean average precision to determine the statistical difference between system pairs (Smucker et al., 2007). Five different tests of statistical difference, namely, Student's t-test, Wilcoxon signed-rank test, sign test, bootstrapping and Fisher's randomization (permutation) were done by using 50 topics. Smucker et al. (2007) attempted to find agreement among these tests with the use of *p*-values. They concluded that Student's t-test, bootstrapping and randomization had little practical difference, whereas the use of Wilcoxon and Sign tests for measuring significant differences between means should be discontinued (Smucker et al., 2007). In another study, only 10 topics were used for statistical tests of randomization, Student's t-test, and bootstrapping (Smucker et al., 2009). Disagreements among these tests were found to increase with 10 topics, but the recommendation to use randomization and Student's t-tests remained.

In another study, significance tests were done with the metrics P@10 and MAP. In the study (Sanderson & Zobel, 2005), an expansion to Voorhees and Buckley's experiment was done. Previously, Voorhees and Buckley (2002) examined the significance by measuring the absolute difference in MAP between pairs of systems. The 50 topics were split into two disjoint sets of 25 topics each to determine if the system ordering in the second set was similar to that in the first set based on error rates in MAP differences. Instead, Sanderson and Zobel extended the research to determine the impact of significance tests on error rates. They concluded that significance increased the reliability of retrieval effectiveness measures (Sanderson & Zobel, 2005).

The authors (Dinçer et al., 2014) aimed to establish a theory of statistical hypothesis testing for risk-sensitive evaluations with the use of a new risk measure known as TRisk.

The testing was done to shift from a descriptive analysis to an inferential analysis of risk-sensitive evaluation. The TREC 2012 Web track was used in a two-sided statistical testing that applied topic scores, ERR@20.

Five different real-life information retrieval systems, namely, Google, Yahoo, MSN, Ask.com, and Seekport, were evaluated to determine their effectiveness based on the list of results and the results description (Lewandowski, 2008). The significant difference between these systems was obtained using chi-square test. System performance was determined with the use of precision at a cut-off of 20 for the top 20 retrieved results. A total of 40 queries were analyzed. The results showed that Google and Yahoo had the best performance of the five systems, although the performances were not significantly different ($p < 0.01$) based on the list of results. However, the P@*k* depending on the results description showed that the differences among all five systems were highly significant (Lewandowski, 2008).

Statistical significance has been measured for system runs using various metrics to examine their ability in differentiating between the systems. The results have been measured in regards to confidence indicators from a test for statistical significance (Moffat et al., 2012). For example, if the agreement of statistical significance between AP and another metric is 81%, 19% of the time the statistical significance results could have been wrongly rejected if a metric other than AP was used (Moffat et al., 2012).

The above studies mentioned used significance testing for various information retrieval system evaluations. The studies utilized averaged or cut-off scores, but none attempted the use of document scores in the statistical tests. It appears that Student's t-test and randomization are most suitable statistical significance test for measuring the mean difference compared to few other statistical tests. Also, incorrect conclusion of

paired significance test needs to be taken into consideration when selecting the significance level.

5.3 Methodology

This study attempts to overcome the problem of using the average or cut-off scores in statistical tests by utilizing unit document scores. The unit document scores such as precision scores of each document will be used as an input to the statistical test to evaluate systems in pairs. The document scores are the base to other metrics such as P@*k* and average precision and do not require the knowledge of total relevant documents to compute the scores. Besides, the proposed approach using document scores could be implemented on actual web retrieval systems, since only the documents scores at each rank will be used in the statistical tests. The mean difference of a pair of system is now dependent on the relevancy of the documents within the evaluation depth instead of the cut-off score at the specified rank.

For example, Table 5.4 shows precision document scores and average precision topic scores for two different systems. The existing method of using topic scores, average precision (SysA: 0.09,0.2267, and SysB: 0.0347, 0.01) in one-sided paired significant test results in a *p*-value of 0.17. The result indicates the two systems are not significantly different (assuming *p*-value = 0.01). In another one-sided paired significant test using P@10 scores between SysA and SysB, the *p*-value is 0.25. The result also indicates the systems are no different. The AP is itself an average of precision scores from relevant documents, and it is expressed in terms of the last document's probability (Moffat et al., 2012). On top of this, the significance test further measures the mean differences of these scores.

However, the proposed method attempts to measure the significant difference between the two systems using the top 10 individual document scores. The *p*-value from a one-sided paired significant test for Topic1 between document scores of SysA and SysB is 0.001. Similarly, the *p*-value for Topic2 is 0.00002. From these significant tests' output, both the systems can be concluded as significantly different (assuming *p*-value = 0.01). The mean differences for the paired t-test are now calculated using the individual document scores. If the two systems are equally good in retrieving relevant documents, regardless they are same or different documents, their effectiveness will be equal, and their mean difference will be 0.

Table 5.4: Example precision and AP scores for different topics and systems (NR = not relevant, R = relevant)

	Topic 1 (total relevant 10)				**Topic 2 (total relevant 10)**			
Rank	Sys A	Precision	Sys B	Precision	Sys A	Precision	Sys B	Precision
1	NR	0.0000	NR	0.0000	NR	0.0000	NR	0.0000
2	R	0.5000	NR	0.0000	R	0.5000	NR	0.0000
3	NR	0.3333	NR	0.0000	NR	0.3333	NR	0.0000
4	NR	0.2500	NR	0.0000	R	0.5000	NR	0.0000
5	R	0.4000	NR	0.0000	R	0.6000	NR	0.0000
6	NR	0.3333	NR	0.0000	R	0.6667	NR	0.0000
7	NR	0.2857	NR	0.0000	NR	0.5714	NR	0.0000
8	NR	0.2500	R	0.1250	NR	0.5000	NR	0.0000
9	NR	0.2222	R	0.2222	NR	0.4444	NR	0.0000
10	NR	0.2000	NR	0.2000	NR	0.4000	R	0.1000
P@10		0.2		0.2		0.4		0.1
AP		0.09		0.0347		0.2267		0.01

For example, observe the relevance of Topic1 documents for both systems. They have each retrieved 2 out of 10 relevant documents and have each ranked them within the top 10 retrieved documents, although the relevant documents may be the same or different. In a simple understanding, these systems are equally effective and cannot be differentiated

when evaluated using P@10. Nonetheless, comparing document scores indicate both systems are different, and SysA is better than SysB for Topic1 since SysA can retrieve relevant documents earlier in the ranks.

As for the RBP metrics, the RBP scores at each document rank are denoted as document-level scores while the RBP@*k* as the topic-level scores. This experimentation will utilize the RBP with persistence values, $p = 0.95$ and $p = 0.8$. These persistence values are common in the information retrieval field (Jones, Thomas, Scholer, & Sanderson, 2015; Shokouhi, Craswell, & Robertson, 2009; Webber, Moffat, Zobel, & Sakai, 2008).

It appears that the selection of unit document scores as an input to paired statistical significance test could be a suitable alternative to the drawbacks in average or cut-off scores in statistical tests. The valuable document scores could be utilized as a way of paired system evaluation.

5.3.1 Test Collection Selection and Cleanup

This study is a laboratory-based experimentation and uses test collections from TREC, including the dataset from TREC-8 ad hoc track and TREC-9 Web track. The ad hoc task evaluates the performance of retrieval systems that search static document corpus (Voorhees & Harman, 1999). The Web track mimics the Web retrieval environment although it used a traditional ad hoc retrieval task (Voorhees & Harman, 2000). The study primarily suits the Web environment; however, the ad hoc task collection is a standard retrieval which is suitable for other retrievals like in the library. Two different test collections will provide evidence if the proposed method is only applicable to Web environment or other ad hoc style retrieval.

The document corpus consists of 100GB of web-crawled documents (Hawking, Craswell, et al., 1999). Both the test collections have different sets of 50 topics available for the retrieval systems but using the same document corpus. The difference among these systems lies in using topics' title-only or combined with other descriptions in retrieving the relevant documents. A total of 129 system runs is available as part of TREC-8, and 105 system runs as part of TREC-9; however, one system run from TREC-9 was not accessible on the TREC website, resulting in 104 system runs.

From these submitted runs, 25% of the least effective runs are removed for this experimentation. This measure was undertaken to avoid distortion in the experimental results because they have poor effectiveness compared to other runs. The least effective systems were determined from their mean average precision scores using traditional TREC method. After the removal of these systems, 97 systems from TREC-8 and 78 systems from TREC-9 remained for consideration in the experimentation.

Each test collection consists of 50 topics, with up to 1,000 documents per topic in the submitted runs. The effectiveness of a system lies in the performance of the system on any query. Hence, all topics were considered in the experimentation regardless of the difficulty level or the number of known relevant documents (based on the relevance judgments from each test collection) for each topic. The TREC relevance judgments were constructed from a pooling depth of 100 based on the contributing systems for each test collection. Contributing systems are those system runs that were chosen for pooling. The TREC-8 relevance judgment uses binary relevancy. However, the relevance judgment from TREC-9 uses ternary relevancy. In this experiment, the ternary relevance is interpreted as binary relevance, with *relevant* (indicated as 1) and *highly relevant* (indicated as 2) as just *relevant* (indicated as 1 in this experiment) for standardization between both test collections.

Before the experimentation, standardization of document rankings was done, in which all documents were arranged in descending order based on their similarity scores; any two or more documents with the same similarity scores were ordered alphabetically by the document ID or identifier. The standardization was necessary because some submitted system runs had rankings that did not match the similarity scores, whereas others had the same similarity scores for many documents. The system runs are now ready for experimentation with the selection of test collection, complete standardization between the test collections and cleanup of the system runs.

5.3.2 Fulfillment of Dependent Test Assumptions as Part of the Experimentation

The proposed method will use the unit document scores in the context of a statistical significance test, precisely the paired Student's t-test, in evaluating system pairs. As stated earlier in Section 5.2.2, paired t-test has a few assumptions that are necessary. These assumptions take account of matched pairs of the independent variable, dependent variable and the distribution of the differences in the dependent variable.

At topic-level evaluation, each system consisted of the same topics per test collection and measures the distribution of relevant documents amongst non-relevant documents. The proposed method also measures the distribution of relevant documents within the top k documents per query with the use of document scores at top k ranks. It focuses on the document's relevance but disregards which of the relevant document was retrieved. This approach is similar to the evaluation at topic-level in regards to measuring the distribution of relevant documents.

To measure the distribution of relevant documents using document scores, the following variables have been selected for the paired t-test. The independent variable is

the document ranks while the dependent variable is the document scores measured on a continuous scale. Independent variable requires both samples to have matched pairs and present in both groups. The assumption on independent variable is met with the selection of document ranks that match between the same topic from different systems.

The third assumption on the distribution of differences in the dependent variable between the two samples should be approximately normally distributed. The metrics selected have values within [0,1]. Thus, the differences between these scores would also remain within the range of 0 and 1. The assumptions integrated for this experimentation facilitates the dependent test to measure the mean difference between the distributions of relevant documents from top *k* documents.

5.3.3 Method for Aggregating p-values

Most times in information retrieval evaluation, individual significant test results are sufficient. Occasionally, a combined significant test result may be necessary. A combined significance test provides an overall level of significance for a series of tests (Cooper & Hedges, 1994). The aggregation of *p*-values is necessary as the proposed method attempts to use document scores for each topic as an input for the paired t-test. Such statistical test would result in many *p*-values from the multiple topics. However, the existing method only results in a single *p*-value for a system. Therefore, a mechanism to enable the comparison between the existing method and the proposed method is essential.

Just like multiple options in significance testing, aggregation of the *p*-values also has numerous choices. Several approaches for summarizing significance level or *p*-values will be discussed in this section to determine a suitable method for the experimentation. Figure 5.2 shows some of these summarizing methods. The summarizing methods assume

p-value as a continuous variable whereby it is a continuous test statistic (Cooper & Hedges, 1994). The continuous *p*-values summaries can be divided into those based on uniform distribution and those based on the statistical theories of other random variables such as transformed uniform variables. A uniform distribution has a constant probability for each variable, where summarizing *p*-values from multiple independent topics per system pair has equal probability. No topic is superior to another. Hence it is safe to state each topic has equal probability. Uniform summaries include counting methods and a linear combination of *p*-values (Cooper & Hedges, 1994).

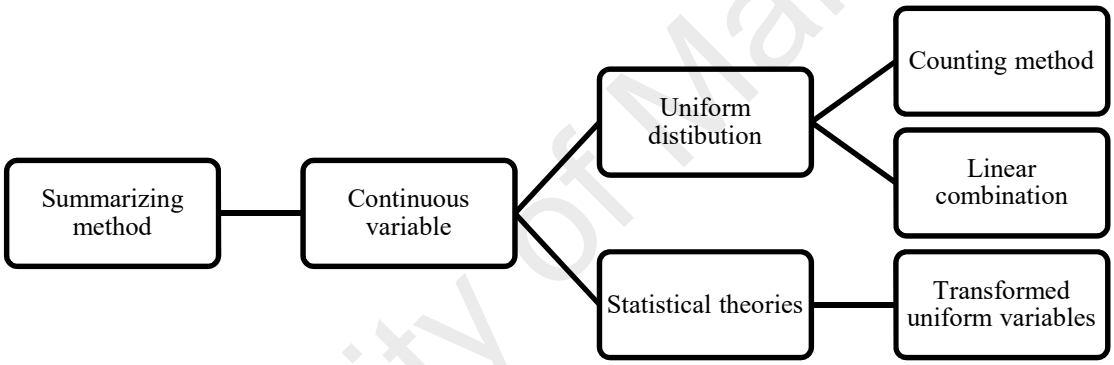

Figure 5.2: Summarizing methods of p-values (the list is not exhaustive)

Sometimes, *p*-value summaries could test a common statistical null hypothesis but do not have to be true for all tests (Cooper & Hedges, 1994). According to Cooper (1994), all of the null hypothesis from the combined *p*-values must be true. The alternative hypothesis, however, could imply that at least one of the population parameter provides evidence to reject the null hypothesis (Cooper & Hedges, 1994). Such alternative hypothesis compliments Fisher's claims (Fisher, 1969).

Fisher (1969) argued that sometimes only, a few or no individual probabilities are significant, but their combined probabilities can be lower than would have been obtained by chance. He also mentioned that occasionally it is necessary to take into account only the individual probabilities to get the aggregated probability, instead of the data from which the individual probabilities were derived, (Fisher, 1969). Rejecting the null hypothesis due to a single *p*-value is not suitable for this study. As mentioned earlier, all topics contribute equally to the effectiveness of a retrieval system despite the fact that some topics may be harder or easier. Hence, it is inadequate to have only one topic per system pair to be statistically different, to conclude for the system pair.

Alternatively, another summarizing method as part of the uniform distribution category known as *meanp* uses a linear combination. *Meanp* is defined as

Equation 5.3

$$z = \sqrt{12k}\left(0.5 - \sum_{i=1}^{k} p_i/k\right)$$

where k is total numbers of *p*-values and p_i are the individual *p*-values that need to be combined or summarized (Cooper & Hedges, 1994) and is a standard normal. If the defined *meanp* has a *z* value of more than $z(\alpha)$, the null hypothesis is rejected. The $z(\alpha)$ for 99% confidence is 2.58.

In combining *p*-values, selecting suitable *p*-value summarizing method is important to reject the null hypothesis only if most topics between the system pairs were significantly different. A system's performance is usually determined by the effectiveness of all topics through mean average precision. Similarly, a combined *p*-value should take into consideration the contribution of all topics per system pair. An analysis between Fisher's and *meanp* method could determine the suitability of a summarizing method. The analysis

determines the percentage of significantly different *p*-values needed from 50 topics to reject the combined null hypothesis.

Figure 5.3 and Figure 5.4 show the percentage of statistically significant *p*-values from 50 topics that reject the null hypothesis for Fisher's and *meanp* method respectively. The plots consist of 100 randomly selected *p*-values. The horizontal line within the graphs represents the *p*-value 0.01 needed to reject the combined null hypothesis.

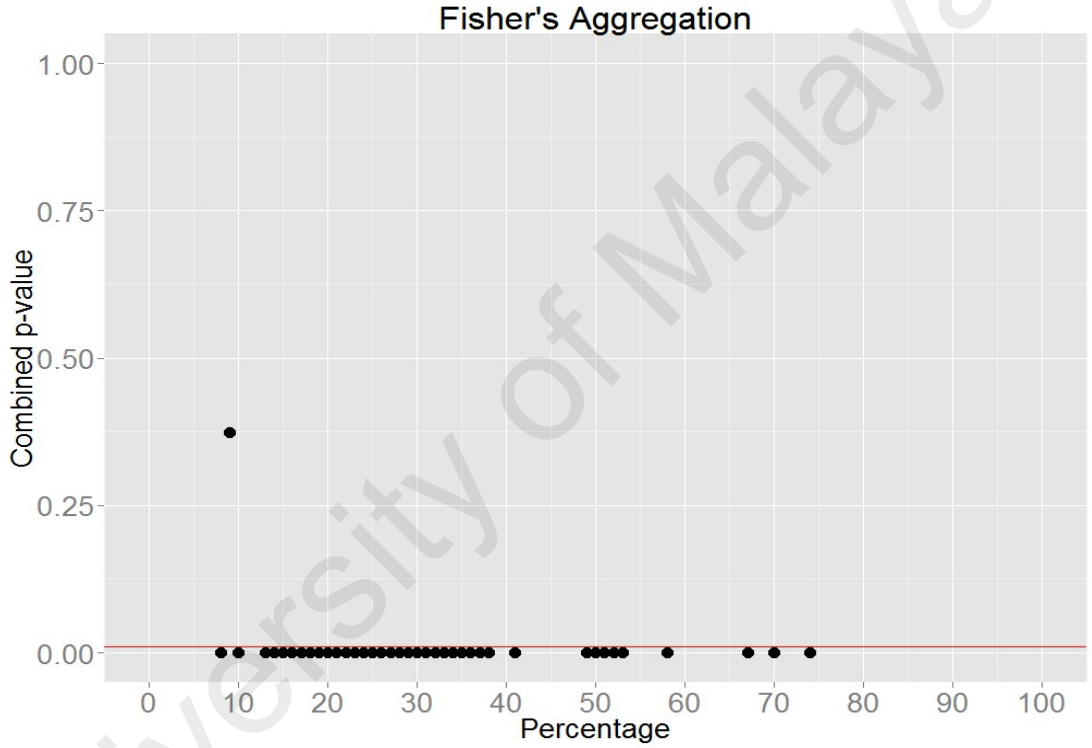

Figure 5.3: Percentage of statistically significant p-values against combined p-values using Fisher's method

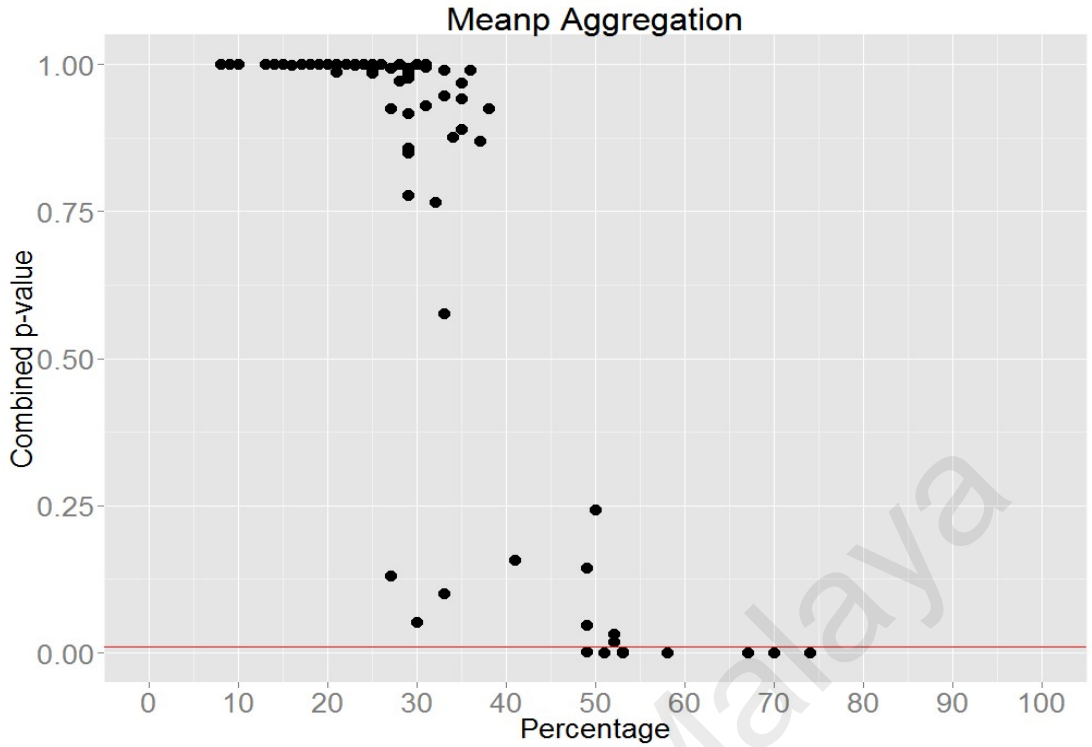

Figure 5.4: Percentage of statistically significant p-values against combined p-values using the *meanp* method

The Fisher's method could provide evidence to reject the null hypothesis even when the percentage of combined *p*-values is as low as 10%. However, the *meanp* method requires almost 50% of the individual *p*-values from 50 topics to be significantly different before providing evidence to reject the combined null hypothesis. Based on this analysis, it is appropriate to employ the use of *meanp* aggregation method in this experimentation to combine multiple *p*-values from the different topics.

5.3.4 The Document Level Assessment in Statistical Significant Test

The proposed method attempts to measure the statistical difference between pairs of systems using document-level scores instead of averaged topic-level scores. Topic-level scores are average or cut-off scores such as AP, RBP@*k,* and P@*k* while document-level

scores are indivisible units document scores such as precision and RBP. The document-level scores cannot be broken down into smaller units and are the basis of a metric. These terms; topic-level and document-level will be used in the remaining sections of this chapter. The total relevant documents do not impact the RBP@k metric but will be experimented to determine if there is a positive influence in using RBP document-level scores in the pairwise system evaluation.

The chosen statistical test is the paired Student's t-test as suggested by Smucker et al. (2007, 2009) for measuring the mean difference. The paired Student's t-test (one-sided, both ways) will be used to determine if one system is better than the other. The null hypothesis, H_o states that the two systems do not differ, whereas the alternative hypothesis, H_A suggests that system A is better than system B or vice versa (refer to Equation 5.4).

Equation 5.4

$$H_0: sysA \neq sysB$$

$$H_A = sysA > sysB \; or \; sysA < sysB$$

Figure 5.5 shows the proposed document-level assessment and the existing method of evaluating system pairs using a statistical test. The dark lines in the illustration represent the proposed method. Meanwhile, the dotted lines are the existing method. In the figure, there are two systems as an example for a pairwise system comparison. Each of these systems has topics, T_n.

The system run is a file that contains all the retrieved documents for each of the topics. The identification of the relevancy of each document for the specified topics is by the

relevance judgment. The relevancy information allows for measurement of effectiveness scores with selected metrics. Firstly, calculate the document scores using specified metrics for each of the retrieved ranked documents per topic. For System1, topic *T1* till *Tn*, select top *k* document scores for each of the topics (step 2). Perform the same for System2. On the other hand, the existing method proceeds to calculate the topic scores such as average precision using the document scores with regards to each topic (step 3).

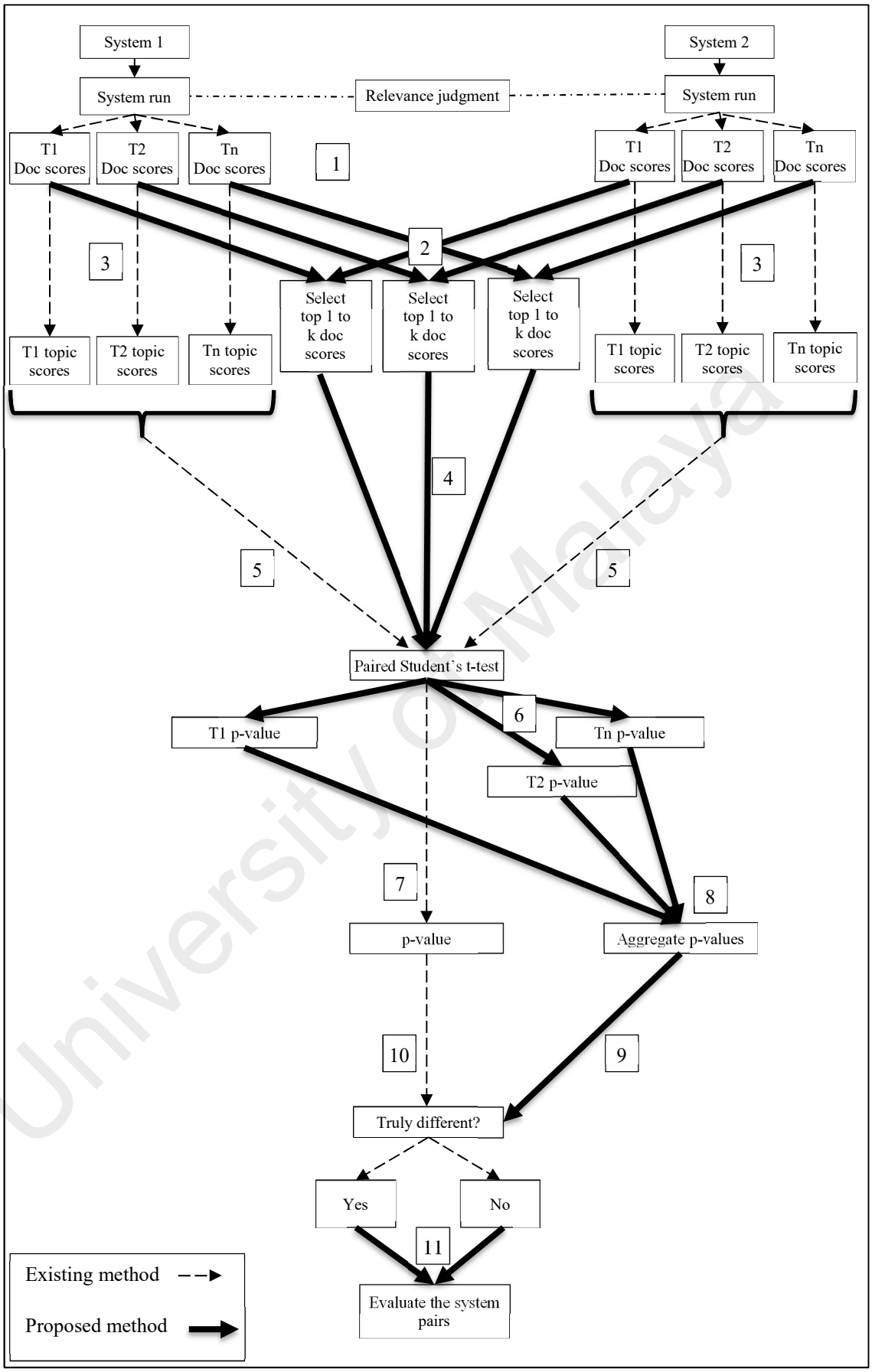

Figure 5.5: Proposed document-level significant test using document scores as opposed to existing method

Use the selected top *k* document scores for topic *T1* from System1 and System2 in the paired Student's t-test (one-sided, both ways) as part of the proposed method (step 4). Perform paired Student's t-test using selected document scores (one-sided, both ways) for all topics between the same pair of system. In the existing method (step 5), the topic scores, *T1* till *Tn* will be used in the Student's paired t-test instead. There are two differences here between the proposed and existing method. First is the type of score used in the paired Student's t-test. The second is that the proposed method calculates mean difference for each topic separately as opposed to the existing method which uses all topic scores in a single paired Student's t-test.

An aggregation method (as discussed in Section 5.3.3) can be used to summarize the multiple *p*-values (step 8). The aggregation provides evidence to reject or accept the null hypothesis, if the systems are truly different or not (step 9). With the aggregation of *p*-value from the proposed approach can now be compared with the existing method's *p*-value (step 11). The comparison is performed with the existing method because it is assumed that it is the gold standard method. Notice the difference between the proposed and existing method of determining the true difference between a pair of systems. The existing method directly produces a single *p*-value that provides evidence to reject or accept the null hypothesis. Conversely, the proposed method has two additional phases to determine the true difference between a pair of system. Nonetheless, it is believed these additional phases would overcome the drawbacks of average and cut-off scores in statistical testing, while still allowing for comparison with the existing method.

Figure 5.6 shows an example calculation of the existing and proposed method's pairwise system evaluation. The box on the left of the figure shows the existing method using topic-level scores while the box on the right is the proposed method. Firstly, using

the AP@1000 scores for all the 50 topics for SysA and SysB (step 1), perform paired Student's t-test (step 2). Due to space limitation, the figure only shows up to 10 topics.

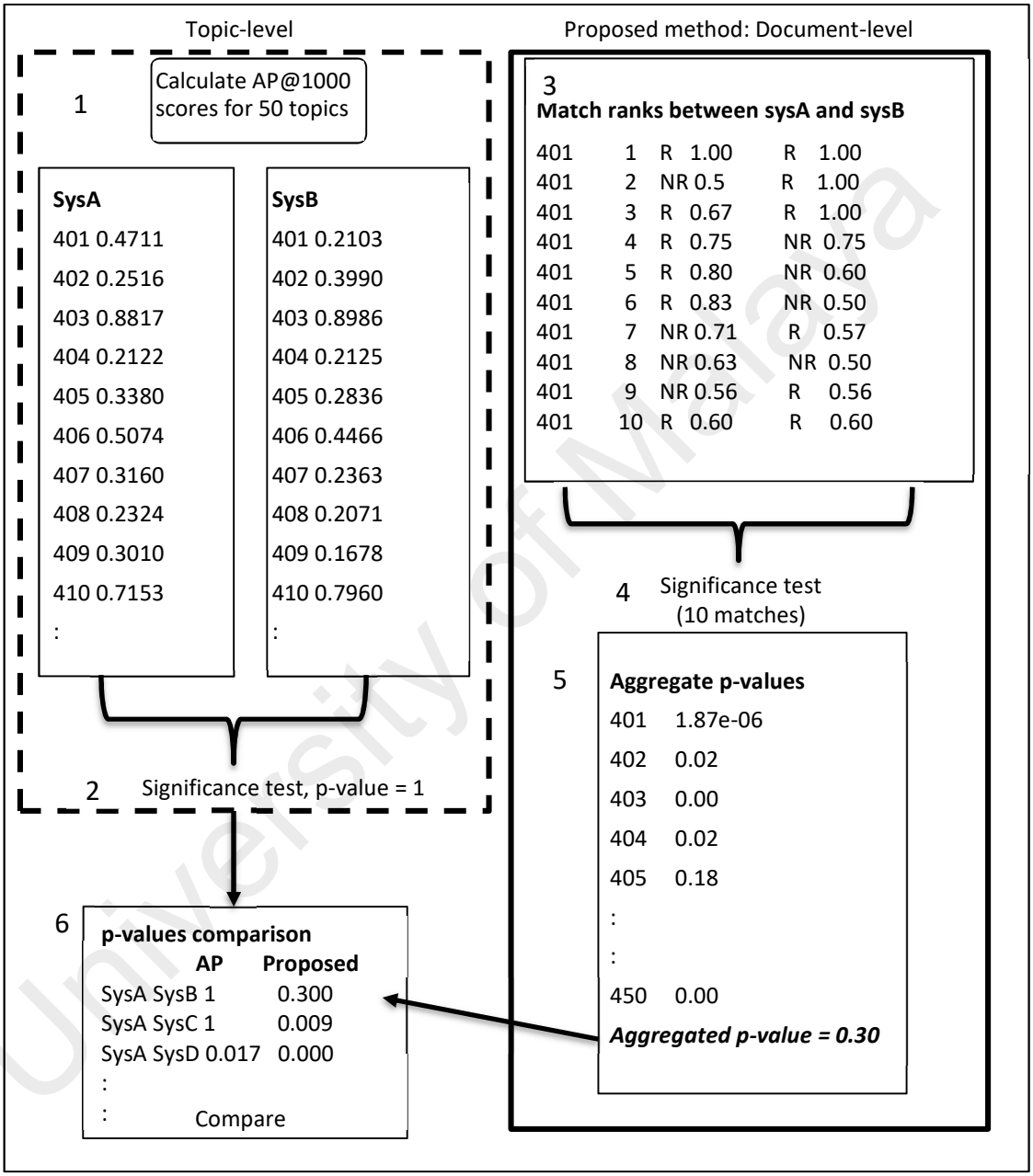

Figure 5.6: Example calculation of proposed method and existing topic-level pairwise system evaluation

On the right side of the figure is the proposed method using document-level scores (step 3). Perform paired Student's t-test using the selected sample size of document scores for Topic 401. The sample size here means the number of top k document scores to be used in the paired Student's t-test (Figure 5.6 shows k=10). In Figure 5.6, the precision document scores of Topic 401 between SysA and SysB is shown. The significant test for topic 401 produces a p-value (step 4). Similarly, proceed to complete the paired Student's t-test for the remaining topics of the same system pair. Then perform the aggregation of these 50 p-values (step 5). Finally, evaluate the results between the proposed document-level method and those from existing topic-level method (step 6).

The evaluation of the results from the significant test includes computing the number of system pairs that are significantly or not significantly different. The comparison of these system pair numbers would indicate the effectiveness of the proposed method in resolving the issue encountered with the usage of topic-level scores. A total of 9,312 for TREC-8 and 6,006 for TREC-9 possible system pairs are available for evaluation in each test collection. The following formula computes total numbers of system pairs for each test collection.

Equation 5.5

$$total\ system\ pairs = \ total\ systems^2 - total\ systems$$

However, there is a limitation in conducting this experiment that relies on the number of matching documents per topic for pairs of systems. A paired Student's t-test is not possible if there are insufficient documents per topic between pairs of systems. For example, a system pair with only five matching documents will not be suitable for evaluation when the sample size needed is 10.

This subsection has detailed the proposed method using document-level scores, the difference between the existing and proposed method, and an example of the existing and proposed method's pairwise system evaluation. The next two subsections, 5.3.5 and 5.3.6 will detail the depth of evaluation, the persistence values for RBP metrics and the sample sizes experimented.

5.3.5 Average Precision (topic-level) versus Precision (document-level)

Existing significance test evaluation by using topic-level AP scores is the basis for comparison with the proposed document-level method. An evaluation depth of 1,000 documents, similar to the standard evaluation depth in TREC (Webber et al., 2010) and 100, will be used in computing the AP scores for each topic. The AP is a commonly used metric in measuring system effectiveness (Robertson, Kanoulas, & Yilmaz, 2010; Moffat & Zobel, 2008). A paired Student's t-test (one-sided, both ways) is performed using these AP@1000 and AP@100 scores, between pairs of systems.

In the proposed method, the document-level scores are calculated using one of the basic evaluation metric, the precision. The precision score of each document per topic per system is computed. A heuristic selection of sample sizes is 10, 30, 50, 100, or 150. These sample sizes represent the number of top k documents that need to be selected for each topic. The paired Student's t-test (one-sided, both ways) uses these precision scores of top k documents per topic from the pair of systems.

System pairs with insufficient top k documents will be excluded from evaluation. The number of significant pairs obtained from the proposed method would be compared with those from the topic-level AP@k (k = 100 or 1000).

5.3.6 RBP@*k* (topic-level) versus RBP (document-level)

Another metric considered is the rank-biased precision. The topic-level metrics considered for evaluation are RBP@100 and RBP@1000. Paired Student's t-test (one-sided, both ways) is performed using these topic-level scores as a basis of comparison for the proposed method using RBP document scores. The RBP persistence values, $p = 0.8$ and $p = 0.95$ will be used since larger values of p are known to lead to deeper evaluation (Webber, Moffat, & Zobel, 2010).

For the proposed document-level method using RBP, the scores are computed for all documents per topic per system using both persistence values. The selection of sample sizes is 10, 30, 50, 100, or 150. These sample sizes represent the number of top *k* documents that need to be selected for each topic. Paired Student's t-test (one-sided, both ways) uses these RBP scores of top *k* documents per topic between pairs of system. The significance test will be performed for both persistence values and evaluated separately.

Similar to the document-level method using precision, any system pairs with insufficient top *k* documents will be excluded from evaluation. The number of significant pairs obtained from the proposed method would be compared with those from the topic-level RBP@*k*.

5.4 Results and Discussions

This section consists of the experimental results and discussions presented in different subsections. Firstly, Section 5.4.1 details the results through identification of statistically significant system pairs using the proposed document-level approach to overcome the drawbacks of topic-level approach. Section 5.4.2 discusses the suitable sample size for statistical testing using the document-level approach to achieve reliable results. Lastly,

Section 5.4.3 evaluates the effectiveness of document-level approach in measuring statistically significant system pairs.

5.4.1 Identification of Statistically Significant System Pairs Using the Document-level scores

This section focuses on evaluating retrieval systems in statistical significance test using document-level approach and corresponds to RQ1. The ability of the document-level approach to differentiate between systems that are statistically significant determines the effectiveness of the proposed method. The results are presented in two separate subsections detailing the number of statistically significant system pairs identified using the document-level approach.

In information retrieval field, higher numbers of statistical significance indicate a proposed method is better than the existing method. Hence, as in any experimentation, error is inevitable. The error that could occur due the to the significance test needs to be understood. A Type I error indicates 1% of the statistically significant system pairs could be wrongly rejecting the null hypothesis. In other words, 10 in 1000 system pairs could be wrongly rejecting the null hypothesis. However, the remaining 99% of the results of the significance tests are believed to be accurate.

5.4.1.1 Statistically Significant System Pairs based on Average Precision and Precision

This experimentation proposed the use of precision document-level scores as an alternate to using the topic-level average precision and precision at cut-off *k* scores. The

effectiveness of the proposed method is measured by comparing the number of statistically significant system pairs and with those from the existing method. The number of system pairs that are statistically significant from the proposed document-level method using precision scores is counted from the aggregated *p*-values (using *meanp* method). Higher numbers of statistically significant pairs indicate the effectiveness of the approach in differentiating system pairs.

Some system pairs had either insufficient ranked documents per topic or the data were constant, making it infeasible for paired Student's t-test. Due to this, there were only between 1 to 3 topics that had significant *p*-values out of the 50 topics. Therefore, there is no aggregated *p*-value for three pairs of systems (see Table 5.5), and the total system pairs that could be evaluated are lesser than the initial total numbers 9,312 system pairs for TREC-8 and 6,006 system pairs for TREC-9. Throughout the experimentation, the statistical significance level is 1%.

Table 5.5: System pairs without aggregated p-values

Topic sample size	Test collection	System Pair	Reason
10	TREC-8	READWARE READWARE2	Mean difference is 0 except two topics.
10	TREC-9	NENRtm NENRtmLpas	Mean difference is 0 except 3 topics
10	TREC-9	uwmt9w10g0 uwmt9w10g2	Mean difference is 0 except two topics.

Table 5.6 shows the number of statistically significant ($p = 0.01$) system pairs at topic-level using AP@1000, AP@100 and P@10 scores and the proposed document-level method using precision scores at various sample size. The percentage values are total

significantly different system pairs out of the total system pairs for the particular test collection. The table includes results from both test collections.

Table 5.6: Pairs of systems that are significantly different (p = 0.01) based on AP@1000, AP@100 and P@10 (topic-level) and precision (document-level) scores. The p-values for document-level method were aggregated by the meanp method

Metrics/Sample size	TREC-8		TREC-9	
	Total	%	Total	%
AP@1000	2166	23%	1297	22%
AP@100	1888	20%	1084	18%
P@100	1678	18%	1200	20%
P@10	2010	22%	1139	19%
10	2609	28%	1627	27%
30	2731	29%	1879	31%
50	2754	30%	1975	33%
100	2748	30%	2016	34%
150	**2772**	**30%**	**2021**	**34%**

Based on Table 5.6, all selected sample sizes of the proposed method at document-level using precision scores yield better results in identifying significantly different system pairs compared to using topic-level scores. Both test collections show similar results. This result could have been achievable from the usage of unit document scores in significance testing as opposed to averaged or cut-off topic scores.

There were no conflicting claims for any of the selected sample sizes, including that of topic-level measures. Conflicting claims here refers to results from paired Student's t-test (one-sided, both ways) suggesting system A is better than system B, and system B is better than system A. The breakdown details of the paired Student's t-test results (one-sided, both ways) is in Appendix K.

Among the topic-level, AP@1000 has identified higher numbers of statistically significant system pairs compared to AP@100 and P@10. Therefore, further analysis

between the existing and the proposed method uses the numbers from AP@1000 instead of AP@100 and P@10. However, the numbers by AP@100 and P@10 are only slightly lesser than that of AP@1000.

For TREC-8, the numbers of statistically significant system pairs identified by the document-level method are 20% to 28% more than that of AP@1000. Meanwhile, for TREC-9, the numbers are 26% to 56% more than AP@1000. These reveal the increase in numbers by the document-level method when the document score sample size increases.

Graphical representation provides an aesthetic of the pairwise system evaluation. The graphical illustration is suitable to quickly show the distribution of *p*-values between the topic-level and the proposed document-level method. Figure 5.7 is a scatterplot graph indicating the *p*-values of system pairs between the topic-level AP@1000 and the document-level precision for TREC-9 with sample size 150. The x-axis represents the topic-level *p*-values, whereas the y-axis denotes the document-level aggregated *p*-values using *meanp*. The vertical and horizontal lines within the graph mark the axes with a *p*-value of 0.01 based on the significance level.

The graph is divided into four different regions as stated below.

I) System pair *p*-values that were significantly different at document-level but not at topic-level;

II) System pair *p*-values that were significantly different at both topic- and document-level;

II) System pair *p*-values that were significantly different at topic-level but not at document-level;

IV) System pair *p*-values that were not significantly different at both topic- and document-level.

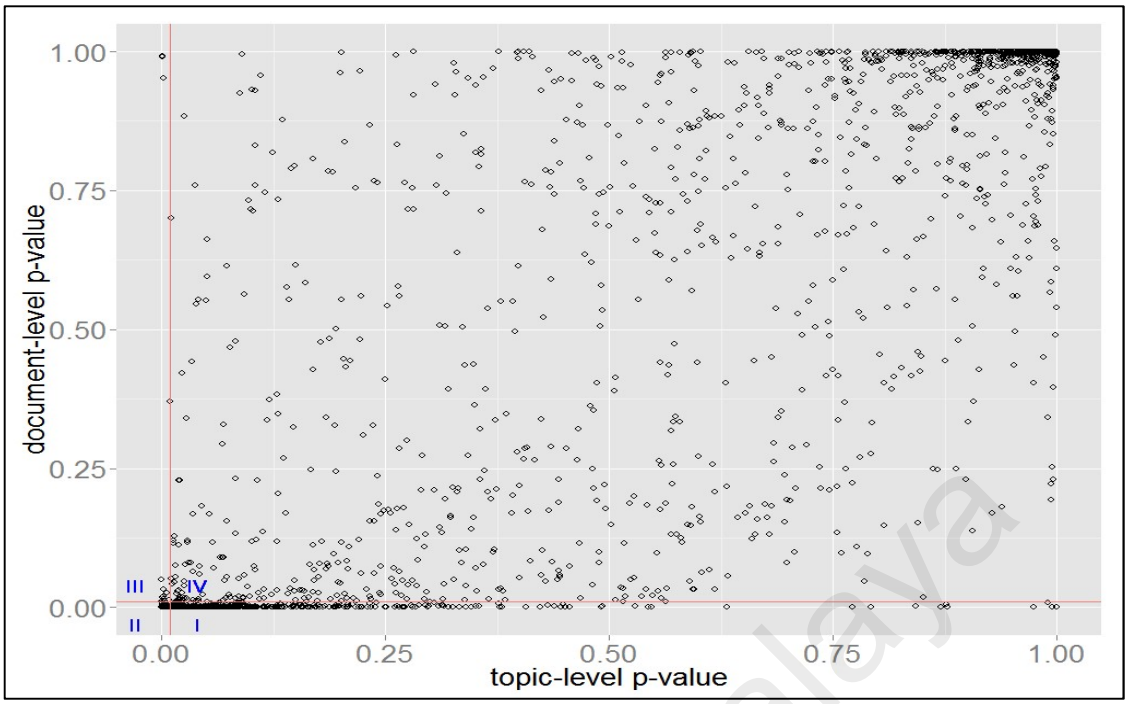

Figure 5.7: System pairs p-values between topic level AP@1000 against document level precision for TREC-9 using 150 sample size

Regions I and III are the most interesting because they include the system pairs that are reversed or in disagreements. Reversed here means the system pair that was initially statistically different in topic-level but is non-statistically different in document-level and vice versa. Region I show many plots compared to region III (see zoomed in areas in Figure 5.8).

More plots in the region I mean document-level method has been able to reverse the non-statistically significant system pairs from topic-level compared to statistically significant. Within region I and III the reversal that takes place suggest that document-level method is effective in using document-level scores to determine statistically or non-statistically significant systems. The higher numbers in the region I suggest that those systems previously not differentiated by the topic-level are now detected by using the

document-level method. A graphical representation conveniently provides a brief understanding of the results promptly without getting into the numerical details.

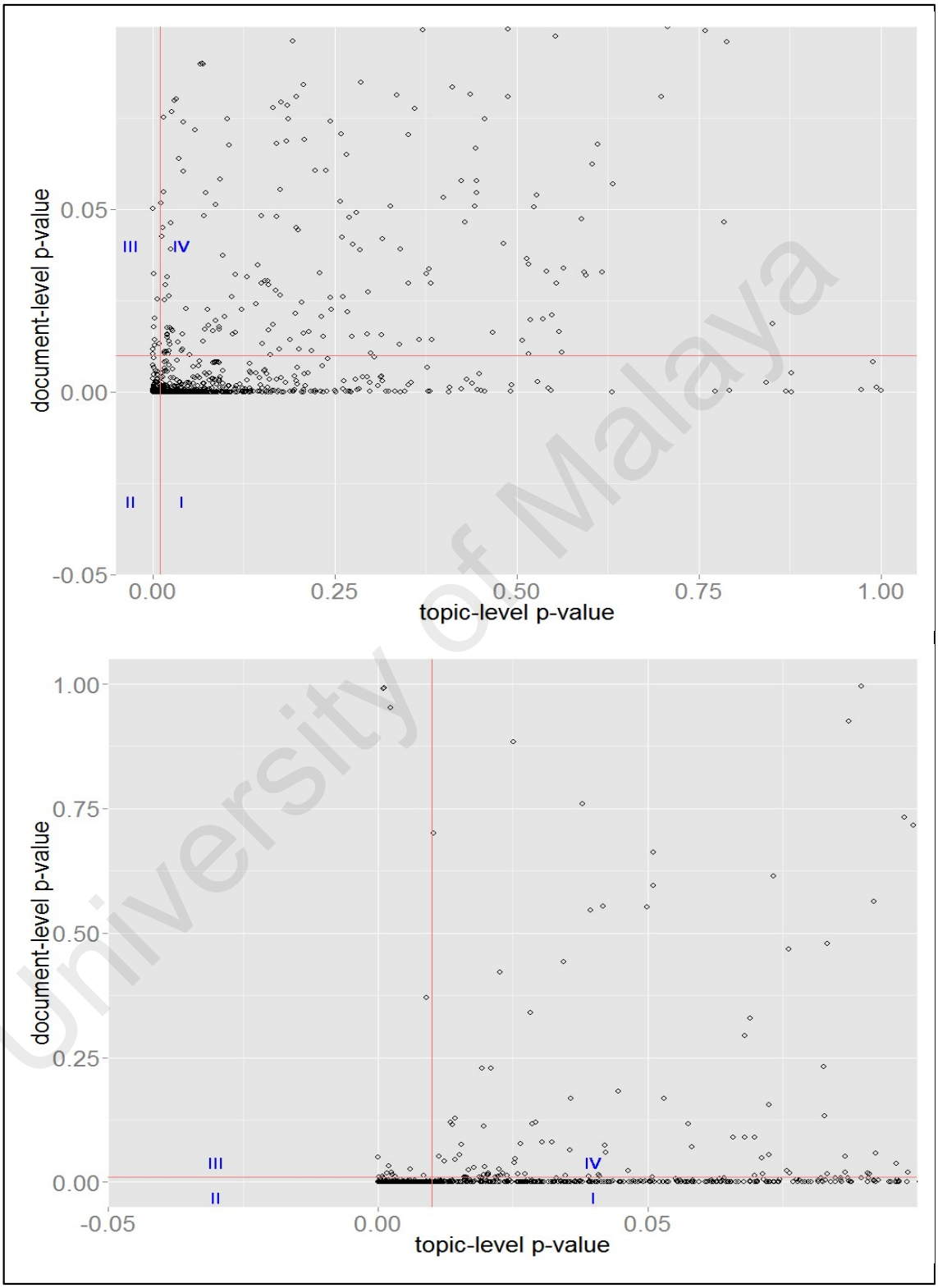

Figure 5.8: Zoomed in regions of graph in Figure 5.7

5.4.1.2 Statistically Significant System Pairs based on RBP@*k* and RBP

Besides precision scores, the proposed approach utilized RBP scores as an alternative to the topic-level cut-off scores in the statistical significance test. The experimental results using metrics RBP@1000, RBP@100 and RBP for p=0.95 and p=0.8 is detailed in Table 5.7. The table contains the number of statistically significant ($p=0.01$) system pairs using RBP metrics for p=0.95 and p=0.8. The total numbers constitute from paired Student's t-test (one-sided, both ways). The percentage values are total significant system pairs out of total system pairs for the particular test collection. The breakdown details of the paired Student's t-test (one-sided, both ways) is in the Appendix L.

Table 5.7: Number of statistically significant system pairs using document level RBP(p=0.95) for the various sample sizes. The document-level p-values are aggregated values

Metrics/Sample size	TREC-8		TREC-9	
	Total	(%)	Total	(%)
RBP@100(p=0.95)	1979	21%	1468	24%
RBP@1000(p=0.95)	1619	17%	1148	19%
10	1929	21%	1384	23%
30	2144	23%	1677	28%
50	**2241**	**24%**	1734	29%
100	2222	24%	1741	29%
150	2195	24%	**1749**	**29%**
RBP@100(p=0.8)	1700	18%	1226	20%
RBP@1000(p=0.8)	1653	18%	1072	18%
10	1978	21%	1366	23%
30	2125	23%	1501	25%
50	**2129**	**23%**	1521	25%
100	2041	22%	1526	25%
150	2002	21%	**1539**	**26%**

The proposed method using document-level RBP scores yield better results in identifying significantly different ($p=0.01$) system pairs compared to using topic-level RBP@1000 and RBP@100 scores. Better evaluation is expected for persistence 0.95 compared to persistence 0.8 due to the nature of rank-biased precision, in which a larger value of persistence leads to a deeper evaluation (Webber et al., 2010). There were no

conflicting claims for both document-level and topic-level method. The similar was observed for AP and precision metric.

The difference in numbers of statistically significant system pairs identified is small amongst the various sample sizes. The differences in percentage are not more than 5% between the sample sizes. The proposed method using RBP(p=0.95) document-level scores has identified approximately 8% to 19% more statistically significant system pairs compared to the topic-level RBP@100(p=0.95). Meanwhile, the document-level method using RBP(p=0.8) has identified approximately 18% to 26% more statistically significant system pairs compared to RBP@100(p=0.8).

The evaluation reveals statistical significance test using document-level scores is capable of differentiating system pairs. The document-level method successfully identified high numbers of statistically significant system pairs. Comparing the results between the document-level metrics show the usage of precision scores in statistical tests is better compared to using RBP scores. Unanimous results for all sample sizes and test collections signify the document-level approach is reliable in measuring statistical significance between system pairs.

5.4.2 Identifying Sample Size for Reliable Statistical Significant Test Using Document-level Scores

This section focuses on identifying a suitable sample size in pairwise retrieval systems evaluation using the document-level approach that produces reliable results and thus answering RQ2. The numbers of statistically significant system pairs using document scores are better than the topic-level. However, within the document-level, each document score sample sizes produced different numbers. An evaluation is performed to

determine the sample size suitable for reliable statistical significance test using document-level scores. Achieving higher numbers of statistical significance system pairs with smaller sample size is better compared to larger sample sizes. A small sample size would mean reliable results can be accomplished in a pairwise system evaluation with judging less ranked documents.

For precision document scores, a sample size of 150 identifies highest numbers of statistically significant system pairs for both test collections. This outcome could have resulted from the larger sample sizes producing smaller *p*-values (Cooper & Hedges, 1994). However, a comparison focusing on the number of data used in the significance testing indicates a possible different reason. The topic-level method used 50 sets of data in their paired Student's t-test. In other words, 50 pairs of scores from 50 topics between a pair of system. In the proposed method, sample size 50 also uses 50 sets of data in its paired Student's t-test, but this time it is the document scores. Isolating the case by the number of sample sizes in statistical test proofs the document-level method is better than topic-level method. Therefore, the underlying reason for larger numbers of statistically significant system pairs by the document-level method is the use of unit document scores.

The similar evaluation is performed for the RBP metric. The TREC-9 results also indicate sample size 150 to be better in identifying higher numbers of statistically significant system pairs. Nevertheless, the TREC-8 results suggest 50 as the optimum sample size. The number of statistically significant system pairs tend to decrease for RBP when the sample size is increased further for TREC-8. It is important to note that although same numbers of datasets were used in significance testing for the topic-level and document-level (sample size 50), the document-level method is better in evaluating pairwise systems.

Also, recall that the depth of evaluation of the topic-level metrics is 100 or 1000 except for P@10. In TREC, pooling gathers documents up to 100 per topic for relevance judgments. However, the proposed method uses between 10 to 150 documents and yet achieve comparably better results than the existing method. Even the statistical tests with sample size 10 attains better results using document-level scores compared to the topic-level method.

The document-level approach achieves sufficiently good results with sample sizes as low as 10. Lesser judgments facilitate fast turnaround of retrieval system results compared to pooling depth of 100. Larger sample sizes have better results in terms of number of statistical significance system pairs.

5.4.3 Evaluation of System Pairs Agreement and Disagreement between the Document-level and Topic-level Method

This section focuses on validating the effectiveness of the document-level approach in measuring statistically significant system pairs and corresponds to RQ3. It is known the document-level method is reliable is differentiating statistical significance test and requires fewer judgments to obtain sufficiently good results. However, it is not known if the proposed method agreed with the existing method to some extent. The analysis would reveal if the document-level method merely identified statistically different system pairs by disregarding the existing method. The existing method is assumed as a baseline for comparison. If the agreement between the document-level and topic-level method are high, it denotes the document-level method measured the system pairs fairly.

The classification of agreement and disagreement was adopted from (Moffat et al., 2012) and modified to suit the comparison between two methods in this experimentation.

The categorizations based on statistical significance are as below, where TLM refers to the topic-level method, and DLM refers to the document-level method:

1. *Active agreements*: where TLM and DLM both provide evidence that system A is significantly better than system B, or vice versa on systems;
2. *Active disagreements*: where TLM states that system A is significantly better than system B, but DLM states that system B is significantly better than system A, or vice versa on systems;
3. *Passive disagreements TLM*: where TLM provides evidence that system A is significantly better than system B (or vice versa on systems), but method DLM does not provide evidence in support of the same claim;
4. *Passive disagreements DLM*: where DLM provides evidence that system A is significantly better than system B (or vice versa on systems), but TLM does not provide evidence in support of the same claim;
5. *Passive agreements*: where TLM fails to provide sufficient evidence that system A is significantly better than system B, and so does DLM;

Based on the above categorization, the agreement and disagreement between the methods can be evaluated. Table 5.8 shows the number of system pairs in agreement or disagreement between AP@1000 and proposed method using precision scores, where the *p*-values are aggregated using the *meanp* summarizing method. The AP@1000 comparison is selected because this metric produced larger numbers of statistically significant system pairs compared to other topic-level metrics. The table shows the comparison for various sample sizes and test collections experimented.

The *active agreements* between topic-level and document-level method increase with the increase of sample sizes. The increase in the *active agreement*s between the methods indicates that larger sample sizes in the proposed method are better able to match the

numbers of statistically significant system pairs identified by the topic-level method. More than 90% of the statistically significant system pairs from the topic-level method has been recognized by the proposed method when using 150 sample sizes for both test collections. Possibilities exist that larger sample sizes produced better *active agreements* since the numbers are highest at sample size 150, the largest sample size experimented. For the smallest sample size 10, *active agreements* of the document-level with the topic-level matches more than 78% for both test collections.

Table 5.8: Number of system pair agreements or disagreements of p-values between topic-level AP@1000 and proposed method document-level precision scores. Aggregated p-values use *meanp* method

	Categorization	AP – 10	AP – 30	AP – 50	AP – 100	AP – 150
TREC-8	Active agreements	1683	1821	1878	1933	**1976**
	Active disagreements	0	2	1	1	1
	Passive disagreements TLM	482	344	288	233	190
	Passive disagreements DLM	926	910	876	815	796
	Passive agreements	6217	6233	6268	6329	6348
TREC-9	Active agreements	1039	1152	1192	1226	**1239**
	Active disagreements	0	1	5	5	4
	Passive disagreements TLM	258	145	105	71	58
	Passive disagreements DLM	591	732	789	793	789
	Passive agreements	4112	3975	3918	3914	3918

The *active disagreements* between the topic-level and the proposed method remain low for the various sample sizes. Only less than 1% of the total number of system pairs were in disagreement. Low numbers in *active disagreements* mean low numbers of conflicting claims by both the method. The low number is a good indication that the proposed method did not cause many conflicting claims and inflict doubts with the results from both methods. Probable reasons for doubts in both methods are because topic-level scores could have caused an error in the statistical test result or the document-level method could have wrongly rejected the null hypothesis as part of the type I error.

The *passive disagreements TLM* and *passive disagreements DLM* numbers decrease as sample size increases. However, for TREC-9 the *passive disagreements DLM* fluctuates. When disagreement decreases, it shows a positive indication that agreements between the topic-level and document-level method are improving. It is convincing to have higher agreements of the proposed method with the existing method in addition to having more statistically significant system pairs identified by the proposed method. Higher agreements mean the proposed method closely mimics the existing method although using only smaller sample size compared to topic-level's depth of 100 (from pooling).

The total numbers of *active agreements* and *passive disagreements TLM* from the various sample sizes are similar. It appears that as the numbers of *passive disagreements TLM* decreases, these numbers mostly add on to the *active agreements*. It shows that as the sample sizes increases, those system pairs that were not identified as statistically significant by the proposed method is now statistically significant. The trend is observed for both the test collections.

The topic-level method using AP@1000 has a total of 7,144 non-statistically significant system pair (1 pair had no t-test result – both ways. Otherwise there should have been 7,146) for TREC-8, and 4,709 for TREC-9. The percentage of *passive agreements* are more than 80% for all sample sizes in both test collections.

Evaluation of the agreement and disagreement between existing topic-level and proposed document-level method is also performed for the RBP metrics. Table 5.9 shows the number of system pair's agreements or disagreements between topic-level RBP@100(p=0.95) and proposed method using RBP(p=0.95) scores whereby the *p*-values are aggregated using *meanp* summarizing method. Meanwhile, Table 5.10 shows the number of system pairs' agreements or disagreements between traditional RBP@100

(p=0.8) and the proposed method using RBP (p=0.8) scores aggregated using *meanp* summarizing method. The RBP@100 was used because it had higher numbers of statistically significant system pairs compared to RBP@1000.

Table 5.9: Number of system pairs agreements and disagreements of p-values between RBP@100(p=0.95) and proposed method RBP (p=0.95). Aggregated p-values use *meanp* method

	Categorization	RBP – 10	RBP – 30	RBP – 50	RBP – 100	RBP – 150
TREC-8	Active agreements	1692	1868	**1907**	1880	1849
	Active disagreements	0	0	0	0	0
	Passive disagreements TLM	287	111	72	99	130
	Passive disagreements DLM	237	276	334	342	346
	Passive agreements	7092	7053	6997	6989	6985
TREC-9	Active agreements	1201	1411	1417	**1419**	1413
	Active disagreements	0	0	0	0	0
	Passive disagreements TLM	266	57	51	49	55
	Passive disagreements DLM	183	266	317	322	336
	Passive agreements	4346	4270	4219	4214	4200

Table 5.10: Number of system pairs agreements and disagreements of p-values between RBP@100 (p=0.8) and proposed method RBP (p=0.8). Aggregated p-values use *meanp* method

	Categorization	RBP – 10	RBP – 30	RBP – 50	RBP – 100	RBP – 150
TREC-8	Active agreements	1631	**1651**	1647	1638	1600
	Active disagreements	0	0	0	0	0
	Passive disagreements TLM	69	49	53	62	100
	Passive disagreements DLM	347	474	482	403	402
	Passive agreements	7261	7134	7128	7207	7208
TREC-9	Active agreements	1164	**1184**	1180	1181	1179
	Active disagreements	0	0	0	0	0
	Passive disagreements TLM	62	42	46	45	47
	Passive disagreements DLM	202	317	341	345	360
	Passive agreements	4568	4461	4437	4433	4418

The increasing sample size decreases *active agreements* for RBP metric. The decline in *active agreements* suggests that it is unlikely for further increase in active agreements with the growth in sample size. However, the percentage of *active agreements* for both

RBP persistence are equally good. The document-level method using RBP can identify more than 93% of the statistically significant system pairs of topic-level RBP@100 for both test collections, and different persistence values. The percentage is calculated by dividing the number of active agreements by the total number of statistically significant system pairs as shown in Table 5.7. When compared between the metrics for document-level method, RBP(p=0.95) is better than RBP(p=0.8) in terms of *active agreements*.

The RBP metric did not record any *active disagreements*, which are conflicting claims by both methods. The *passive disagreements TLM* for RBP (both persistence values) decreases slightly and then increases as the number of sample size increases. However, the trend of *passive disagreements DLM* is different as it continues to increase with the increase in sample size. In contrast, TREC-8 *passive disagreements DLM* for RBP(p=0.8) tend to decrease with the increase of sample size. The increase in disagreements also translates to decrease in the *active agreements* as can be observed by the reducing number of *active agreements*.

The *passive agreements* for both the RBP persistence values decrease when the number of sample size increase. The only difference lies with TREC-8 *passive agreements* for RBP(p=0.8) where the numbers increase with the sample size. The differences in numbers in *passive disagreements TLM* and *passive disagreements DLM* is widely influenced by the numbers of *active agreements* and *passive agreements*, and vice versa.

The overall agreements (active and passive) and disagreements (active and passive) is shown in Table 5.11. The total agreements include *active agreements* and *passive agreements*. These are the numbers that show the capabilities of the proposed method to match the existing method. A higher percentage means a better agreement of the document-level with the topic-level. The total disagreements include *active*

disagreements, and *passive disagreements TLM* and *DLM*. These numbers indicate the dissimilar claims by the proposed method with the existing method.

Table 5.11: Total number of agreements and disagreements of p-values between topic-level and document-level methods

Test collection	Metrics	AP – 10	AP – 30	AP – 50	AP – 100	AP – 150
TREC-8	Total agreement	7900 85%	8054 86%	8146 87%	8262 **89%**	8324 **89%**
	Total disagreement	1408 15%	1254 13%	1164 13%	1048 11%	986 11%
TREC-9	Total agreement	5151 **86%**	5127 85%	5110 85%	5140 **86%**	5157 **86%**
	Total disagreement	849 14%	877 15%	894 15%	864 14%	847 14%
Test collection	**Metrics**	**RBP – 10**	**RBP – 30**	**RBP – 50**	**RBP – 100**	**RBP – 150**
p=0.95						
TREC-8	Total agreement	8784 94%	8921 **96%**	8904 **96%**	8869 95%	8834 95%
	Total disagreement	524 6%	387 4%	406 4%	441 5%	476 5%
TREC-9	Total agreement	5547 92%	5681 **95%**	5636 94%	5633 94%	5613 93%
	Total disagreement	449 7%	323 5%	368 6%	371 6%	391 7%
p=0.8						
TREC-8	Total agreement	8892 **95%**	8785 94%	8775 94%	8845 **95%**	8808 **95%**
	Total disagreement	416 4%	523 6%	535 6%	465 5%	502 5%
TREC-9	Total agreement	5732 **95%**	5645 94%	5617 94%	5614 93%	5597 93%
	Total disagreement	264 4%	359 6%	387 6%	390 6%	407 7%

The total agreements exceed 80% for AP-precision combination while the total disagreements are below 15% for both test collections. As for RBP, both persistence values, the total agreements exceed 90% while the total disagreements remain below 7%.

The overall similarity in results for both test collections shows that the proposed method is reliable and consistent.

A high percentage also validates the results from the document-level because it is as capable as the topic-level. The results provide confidence in the proposed approach such that it did not simply evaluate system pairs as statistically significant. The disagreements could have occurred because the existing method had inaccurately evaluated the statistical significance of the system pairs due to the topic-level scores. Now with the usage of document-level scores, the proposed method is able to identify those missed statistically significant system pairs.

5.5 Summary

The topic-level scores tend to create diverse user experience when expressed in terms of the last document's probability despite the differences in the ranks of each relevant document. Some times the average scores also require total relevant documents for measuring the effectiveness scores used in statistical significance testing. The need for total relevant documents in evaluating the systems is not suitable for use in the real Web due to ever changing contents of the retrieval systems. Also, the use of topic-level scores causes inaccurate results in statistical testing.

An alternate approach using document-level scores is effective in evaluating system pairs in statistical significance testing. The use of indivisible document-level scores is versatile to the constant changes of Web retrieval systems, as it does not require the number of total relevant documents at the time of evaluation, and thus meets the OBJ1. However, repeated judgments may be needed from the user to determine the relevancy of the documents from the Web. Nevertheless, the proposed approach is capable of

differentiating system pairs statistically with just 10 document pairs. Expecting judgments for the top 10 documents per system is not a huge effort, though results may differ between expert judgments and crowdsourced judgments.

The document-level approach is an effective approach for evaluating system pairs in statistical significance testing. The approach is also reliable and produces consistent results with different sample sizes in statistical tests and test collections. However, evaluation of system pairs using document-level precision scores is better than RBP scores. The nature of the metrics could have caused such variation. Both metrics have different ways to regard the relevant and non-relevant documents at each document rank.

For OBJ2, it can be summarized that larger sample sizes identified higher numbers of statistically significant system pairs. But, evaluation by an equal number of sample size in the significance testing for both topic-level and document-level, suggests sample size is not the cause of higher numbers of statistically significant pair. Therefore, the underlying reason for high numbers of statistically significant system pairs in the document-level approach is the usage of document scores.

To fulfill OBJ3, evaluation based on the agreement and disagreement between the document-level and topic-level approach reveal the document-level approach has a high percentage of agreement with the results from the topic-level approach. Agreements validate the proposed approach as it matches the existing, additionally differentiating more system pairs missed by the topic-level method. The overall similarity in results for both test collections shows that the proposed method is reliable and consistent.

A simple variation in the document-level approach compared to the existing shows a huge difference in the statistical tests result. High numbers of significant pairs, percentage of agreements, and better results with an equally same number of sample sizes in

statistical tests definitely highlights the abilities of the document-level approach. The proposed approach is a suitable alternative to statistically evaluate Web retrieval systems besides the static test collections by overcoming the inaccuracy of score averaging and cut-offs in a pairwise system evaluation.

CHAPTER 6: CONCLUSION

In information retrieval evaluation, system-oriented evaluation uses test collections. In this thesis, several contributions are made in the system-oriented evaluation in a laboratory experimental methodology. Firstly, information retrieval systems can be evaluated reliably using effort based relevance judgments at different depth of evaluation. Second, the reliability of individual system rankings can be measured consistently with different combinations of metrics based on the relative topic ranks. Lastly, the study proposed the use of document-level scores to statistically evaluate retrieval systems in a pairwise manner. In the following paragraphs, the summaries of the three main contributions of the thesis are detailed. Finally, the thesis is concluded with future research in these areas.

6.1 Thesis contributions

6.1.1 Evaluating retrieval systems using effort based relevance judgments

System-oriented evaluation has always prioritized relevance as part of its test collection but the effort in terms of work needed by the user to retrieve relevant contents from the documents is equally important for measuring the effectiveness of retrieval systems.

The first problem addressed in this experimentation relates to the real users whom do not put in as much effort as experts in retrieving the relevant contents. They also tend to give up easily while trying to identify relevant contents from the documents. Therefore, retrieval systems evaluation should prioritize low effort to ensure user satisfaction.

The second problem tackled relates to the limited evaluation depth and retrieval systems that showed differences in system rankings due to low effort relevance

judgments. However, it is unknown if similar conclusions can be drawn with deeper evaluation depth and wider groups of retrieval systems within a test collection.

The third problem focuses on the advancement of reduced workload on relevance judgments without risking the quality of evaluation. Such reduction in workload is achievable by evaluation with reduced topic size.

The objectives of this experimentation addressed the mentioned problems. The three objectives and the achievement of the objectives are as below.

The first objective is to propose a systematic way of generating relevance judgments that incorporate effort and determine the variation in system rankings due to low effort relevance judgment in evaluating retrieval systems at different depth of evaluation. The objective was achieved through the generation of various relevance judgments for different effort features using the boxplot approach for simple document features and HTML features, and standardized grading for readability features. Deeper depth of evaluation using low effort relevance judgments show system rankings do not differ widely. But, it is crucial to use low effort relevance judgments in determining retrieval systems that satisfy users when evaluating.

The second objective is to measure the variation in system rankings due to low effort relevance judgment on groups of systems in information retrieval system evaluation. The objective was achieved by evaluating different groups of retrieval systems and the findings reveal differing system rankings when evaluated using low effort relevance judgments. The top performing systems are capable of retrieving relevant documents but not documents that could satisfy real users because they are likely high effort documents. An actual user could face difficulties in consuming the information from these retrieval

systems. However, it is not the case with the least performing systems since they have been retrieving low effort documents suitable for the utility of actual user.

The third objective is to explore the effectiveness in evaluating retrieval systems using low effort relevance judgment with reduced topic sizes. The objective was achieved through evaluation of retrieval systems using different numbers of reduced topic sizes and the created low effort relevance judgments. Limited effort features are effective in evaluating retrieval systems using low effort relevance judgment with reduced topic sizes. Also, there exists variation in suitable effort feature between test collections. The prediction of overall system effectiveness with reduced topic size using low effort relevance judgments is inconclusive due to the effect of topic variation.

Chapter 3 experimentation implies that information retrieval system evaluation should prioritize the use of low effort relevance judgments to recognize retrieval systems that are capable of satisfying real users. Evaluation focusing only on relevance highlights retrieval systems viewed by expert judges as effective but is not necessarily effective in satisfying real users.

Therefore, those retrieval systems that have been capable of retrieving high numbers of relevant documents should focus on retrieving and ranking low effort documents to ensure user satisfaction. Meanwhile, those retrieval systems already retrieving low effort documents should prioritize on increasing the number of low effort documents to fulfill user's need. As such, the retrieval mechanism from the systems of high retrieving relevant documents and systems retrieving low effort relevant documents could be merged to produce a strong retrieval system that mainly targets to satisfy the users.

6.1.2 Measuring the reliability of individual retrieval system rankings

A reliable system is crucial in satisfying users' need. In evaluating the retrieval systems, different metrics based on user models, fulfilling different user needs have measured reliability for a set of systems through experimentations based on the differences in scores or ranks obtained from a test-retest kind of setting. Evaluation of system rankings in a set of systems indicates the overall reliability of the systems but not for the individual systems. The reliability of the individual system performance measured by their ranks provides the confidence a user can gain from the retrieval system upon repeated queries.

The first problem addressed in this experimentation is that reliability is measured based on the differences in scores or ranks obtained from a test-retest setting for a set of system rankings but the reliability of individual system rankings have not been explored.

The second problem focuses on suitable metrics and combination of metrics for measuring reliability of individual system rankings. A large number of metrics were grouped based on their mathematical properties, although the metrics have different user models. One metric per group was highlighted to sufficiently evaluate retrieval systems. But, it is not known which of those metrics are suitable for measuring the reliability of system rankings.

A single evaluation metric measures the effectiveness of a retrieval system based on specific user model. An evaluation of the retrieval systems using a combination of metrics will indicate the versatality of the retrieval system in fulfilling different user models. Nonetheless, reliability of individual system rankings from a combination of metrics has not be explored.

The objectives of this experimentation addressed the mentioned problems and measure the effectiveness of the proposed approach. The objectives and the achievement of the objectives are as below.

The first objective is to propose a method to evaluate the reliability of individual retrieval systems and determine suitable combination of metrics for measuring reliability of individual system rankings. The objective was achieved using intraclass correlation coefficient approach which uses relative rankings from topics of the specific retrieval system to measure the reliability of individual system rankings. The findings suggest the metrics combination, AP and RBP as suitable measure for reliability of individual system ranks. The metrics pair is capable of measuring reliability of individual retrieval system ranking even with different depth of evaluation.

The second objective is to understand the generalization of the system ranking reliability to other similar metrics pairs. The objective was achieved using intraclass correlation coefficient with other similar metrics. The various metrics combinations, AP-RBP, AP-P@k (k=large), AP-P@k (k=small), and RBP-P@k (k=large) generalize well except RBP-P@k (k=small). Generalization disregards the reliability coefficient values but focuses on obtaining similar results between the initial and similar metrics combinations. However, it is favourable to identify retrieval systems that are highly reliable in their rankings.

The third objective is to identify the original systems with reliable system rankings. The objective is achieved by mapping the reliability coefficient scores to the original system ranks. The findings reveal the top and middle ranked systems are highly reliable compared to bottom ranked systems. Therefore, a user can be confident in receiving good retrieval results most of the time from these highly reliable ranked systems.

The fourth objective is to validate the reliability measurement of individual retrieval system rankings with the original system rankings. The objective is achieved through the correlation coefficient of system rankings between the original and inferred ranks. The inferred system ranks are obtained from the intraclass correlation coefficient approach. High correlation coefficient of system rankings between the original and inferred ranks validates the reliability measurement of individual retrieval system rankings. Therefore, the reliability coefficients could represent the original system ranks despite some variation in the inferred system ranks.

The chapter 4 experimentation highlights the proposed intraclass correlation coefficient approach using relative rankings of topics from the specific retrieval system as a reliable approach to measure the reliability of individual retrieval system ranks. Nevertheless, ranking only makes sense if it is relatively measured. The approach is capable of determining the versatility of the systems in satisfying multiple user needs since evaluation measures utilized combination of metric. Besides, the approach can be easily replicated and suited to other test collections for it utilizes relative ranking in measuring reliability.

However, P@*k* (*k*=small) should not be used to measure reliability of individual system ranks as it does not generalize well and produces inconsistent retrieval results. It also fails to represent the reliability coefficient to the original system rankings.

6.1.3 Document level assessment using document level scores

Ranking the retrieval systems based on their effectiveness scores allow us to determine the superiority of the performance of one system over those of other systems. However, the difference in performance could have occurred by chance. An alternative way to

determine the true difference in the performances of the systems is to do pairwise system evaluation.

The first problem addressed in this experimentation relates to the relevant document's rank. The metric precision at a cut-off rank *k* (P@*k*) gives a differing user experience depending on the number of relevant documents within a specific cut-off rank. It occurs because the rank of the relevant document is not taken into consideration. Therefore, measuring the significant difference between a system pair with their P@*k* performance scores produces incorrect results despite the difference observed in the ranked relevant documents.

The second problem addressed relates to the knowledge of total number of relevant documents prior to evaluation. The average precision (AP) is a widely utilized metric in measuring information retrieval system performance. However, the system performance score can only be measured with a known total relevant documents for a topic. In real Web experience, the total relevant documents for a topic is unknown due to the constant changes in the actual Web. Due to this, evaluating existing or new system pairs in real Web becomes infeasible. Therefore, the usage of these averaged or cut-off topic scores are not suitable for statistical significance testing although they are common evaluation measures.

The third problem addressed is the variation in outcome due to different sample sizes. Different sample sizes used in the statistical tests tend to produce varying outcome which raises concern on the conclusions drawn. To ensure consistent results, exploration of different sample sizes in retrieval system evaluation is necessary. Hence, it is crucial to explore different sample sizes in statistical tests and analyze the results before conclusions can be drawn.

The objectives of this experimentation addressed the mentioned problems and measure the effectiveness of the proposed approach. The objectives and the achievement of the objectives are as below.

The first objective is to propose an approach suitable for evaluating retrieval systems by overcoming the drawbacks of inaccuracy using averaged or cut-off scores in statistical significance test. The objective is achieved with an alternate approach using document-level scores for evaluating system pairs in statistical significance testing. The use of indivisible document-level scores overcomes the drawbacks of inaccuracy using averaged or cut-off scores and does not require total relevant documents in statistical significance test. The proposed approach identified higher numbers of statistical significant system pairs compared to the existing topic-level approach. The high numbers of statistical significant system pairs prove the effectiveness of the proposed document-level approach.

The second objective is to identify a suitable sample size in pairwise retrieval systems evaluation using the proposed approach that produces reliable results. The objective was achieved through exploration of different sample sizes on the proposed document-level approach. The document-level approach achieves sufficiently good results with small sample sizes. Lesser judgments facilitate fast turnaround of retrieval system evaluation compared to larger sample sizes. However, larger sample sizes produce better results in terms of number of statistical significance system pairs. The approach is reliable and produces consistent results with different sample sizes in statistical tests and test collections.

The third objective is to validate the effectiveness of the proposed method in measuring statistically significant system pairs. The objective was achieved by measuring the effectiveness of the document-level approach based on the agreements and disagreements of the statistical test results between the proposed and existing approach.

High percentage of agreements between the approaches validates the effectiveness of the document-level approach. The similarities of outcome from different test collections highlights the reliability and consistency of the proposed approach.

The chapter 5 experimentation signifies a simple variation in statistical test using document-level approach shows a huge difference in the statistical tests result. The document-level approach is versatile to the constant changes of Web retrieval systems, overcomes inaccuracy of using topic-scores, and identifies higher numbers of statistically significant pairs. In addition, the approach is consistent and reliable, besides being suitable for evaluation on the Web and laboratory experimentation.

6.2 Future works

This section highlights some of the interesting studies that can be extended from the work documented in this thesis.

In relation to chapter 3, the experimentation could include many other effort features such as summary features, document specific features, outlink oriented features, query specific features, and query term window specific features (Verma et al., 2016) in evaluating retrieval systems. It could also explore the combination of metrics to evaluate the impact of the combined features on the system rankings. Other than that, retrieval of documents could be personalized to the user's reading capabilities (Collins-Thompson, 2011) with the incorporation of reading effort of a document. This thesis has included selected effort features affecting the retrieval system's performance but future experimentations could also produce interesting results.

In chapter 4, the reliability of individual retrieval system rankings was measured using combinations of metrics from selected clusters. Further experimentations could include

other metrics such as NDCG, ERR, or metrics from other clusters (Baccini et al., 2012). This thesis explored two different metrics combinations in measuring reliability of retrieval system rankings but the ICC approach could cater to more than three metrics combinations. The expansion of number of metrics combinations could highlight versatile retrieval systems in future works.

In chapter 5, document level assessment revealed effective results compared to using topic-level scores. The experimentation could include additional effectiveness metrics measured at the document-level and cut-off ranks. The approach could be extended to the real Web with the assistance of crowdsourced judgments to obtain the relevancy of documents on-the-go. However, experiments involving users may have to be controlled to eliminate user errors, but it is an interesting area of study.

Finally, information retrieval is an ongoing process of improvisation through query formulation, indexing, and retrieval algorithm. Information retrieval through search engines in the Web continuously evolves to satisfy users. But the information retrieval evaluation should also improve equivalently. Therefore, new advancements in the system-oriented evaluation is definitely an important aspect to cater for the constant changes in the information retrieval field. And developing such approaches in the information retrieval evaluation is definitely an important aim.

REFERENCES

Al-Maskari, A., & Sanderson, M. (2010). A Review of Factors Influencing User Satisfaction in Information Retrieval. *Journal of the Association for Information Science and Technology*, *61*(5), 859–868.

Algina, J. (1978). Comment on Bartko's "On various intraclass correlation reliability coefficients." *Psychological Bulletin*.

Anderson, T. D. (2006). Studying human judgments of relevance. In *Proceedings of the 1st international conference on Information interaction in context - IIiX* (Vol. 176, pp. 6–14). Copenhagen: ACM. https://doi.org/10.1145/1164820.1164825

Aslam, J. A., Pavlu, V., & Yilmaz, E. (2006). A statistical method for system evaluation using incomplete judgments. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 541–548). https://doi.org/10.1145/1148170.1148263

Aslam, J. A., Yilmaz, E., & Pavlu, V. (2005). The maximum entropy method for analyzing retrieval measures. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '05* (p. 27). New York, USA: ACM Press. https://doi.org/https://doi.org/10.1145/1076034.1076042

Baccini, A., Déjean, S., Lafage, L., & Mothe, J. (2012). How many performance measures to evaluate information retrieval systems? *Knowledge and Information Systems*, *30*(3), 693–713. https://doi.org/https://doi.org/10.1007/s10115-011-0391-7

Bailey, P., Moffat, A., Scholer, F., & Thomas, P. (2015). User Variability and IR System Evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '15* (pp. 625–634). https://doi.org/10.1145/2766462.2767728

Barrett, P. (2001). Assessing the Reliability of Rating Data.

Bartko, J. . (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, *19*(1), 3–11. https://doi.org/https://doi.org/10.2466/pr0.1966.19.1.3

Bartko, J. J. (1976). On Various Intraclass Correlation Reliability Coefficients. *Psychological Bulletin*, *83*(5), 762–765. https://doi.org/https://doi.org/10.1037/0033-2909.83.5.762

Berto, A., Mizzaro, S., & Robertson, S. (2013). On Using Fewer Topics in Information Retrieval Evaluations. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval - ICTIR '13* (pp. 30–37). https://doi.org/10.1145/2499178.2499184

Blanco, R., Halpin, H., Herzig, D. M., Mika, P., Pound, J., & Thompson, H. S. (2013).

Repeatable and Reliable Search System Evaluation using Crowdsourcing. *Journal of Web Semantics*, *21*, 923–932. https://doi.org/10.1016/j.websem.2013.05.005

Borenstein, M. (2009). *The handbook of research synthesis and meta-analysis*. (H. Cooper, L. Hedges, & J. Valentine, Eds.) (2nd ed.). New York, USA: Russell Sage Foundation.

Borlund, P. (2009). *Information retrieval : searching in the 21st century*. (A. Goker & J. (N. J. Davies, Eds.). Wiley.

Box, G. E. ., Hunter, W. ., & Hunter, J. . (1978). *Statistics for Experimenters*. New York: John Wiley & Sons.

Braschler, M., & Peters, C. (2004). Cross-Language Evaluation Forum: Objectives, Results, Achievements. *Information Retrieval*, *7*(7), 7–31.

Buckley, C., & Voorhees, E. M. (2000). Evaluating Evaluation Measure Stability. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 33–40). https://doi.org/10.1145/345508.345543

Carterette, B. (2012). Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Transactions on Information Systems*, *30*(1), 1–34. https://doi.org/10.1145/2094072.2094076

Carterette, B., Allan, J., & Sitaraman, R. (2006). Minimal Test Collections for Retrieval Evaluation. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval-SIGIR'06* (p. 268). Washington: ACM.

Carterette, B., Kanoulas, E., Pavlu, V., & Fang, H. (2010). Reusable test collections through experimental design. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10* (pp. 547–554). Geneva: ACM. https://doi.org/10.1145/1835449.1835541

Carterette, B., Kanoulas, E., & Yilmaz, E. (2010). Low-Cost Evaluation in Information Retrieval. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10* (p. 903). https://doi.org/10.1145/1835449.1835675

Carterette, B., Pavlu, V., Kanoulas, E., Aslam, J. A., & Allan, J. (2008). Evaluation over thousands of queries. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08* (pp. 651–658). Singapore: ACM. https://doi.org/10.1145/1390334.1390445

Carterette, B., & Soboroff, I. (2010). The Effect of Assessor Errors on IR System Evaluation. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval-SIGIR'10* (pp. 539–546). Geneva: ACM. https://doi.org/10.1145/1835449.1835540

Chandar, P., Webber, W., & Carterette, B. (2013). Document Features Predicting Assessor Disagreement. In *Proceedings of the 36th International ACM SIGIR*

Conference on Research and Development in Information Retrieval (pp. 745–748). Dublin: ACM. https://doi.org/10.1145/2484028.2484161

Chua, Y. P. (2013). *Mastering Research Statistics* (1st ed.). McGraw- Hill Education (Malaysia) Sdn. Bhd. https://doi.org/https://doi.org/9789675771699

Clarke, C. L. A., Craswell, N., & Soboroff, I. (2009). Overview of the TREC 2009 Web Track. In *NIST Special Publication 500-278: The Eighteenth Text REtrieval Conference Proceedings (TREC 2009)* (pp. 1–9).

Clarke, C. L. A., Craswell, N., Soboroff, I., & Cormack, G. V. (2010). Overview of the TREC 2010 Web Track. In *NIST Special Publication 500-294: The Nineteenth Text REtrieval Conference Proceedings (TREC 2010)* (pp. 1–9).

Clarke, C. L. A., Craswell, N., Soboroff, I., & Voorhees, E. M. (2011). Overview of the TREC 2011 Web Track. In *NIST Special Publication 500-296: The Twentieth Text REtrieval Conference Proceedings (TREC 2011)* (pp. 1–9). TREC.

Clarke, C. L. A., Craswell, N., & Voorhees, E. M. (2012). Overview of the TREC 2012 Web Track. In *NIST Special Publication 500-298: The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)* (pp. 1–8).

Clough, P., & Sanderson, M. (2013). Evaluating the performance of information retrieval systems using test collections. *Information Research*, *18*(2), paper582.

Cohen's kappa. (2017). Retrieved August 20, 2017, from https://en.wikipedia.org/wiki/Cohen%27s_kappa

Collins-Thompson, K. (2011). Enriching Information Retrieval with Reading Level Prediction. In *SIGIR 2011 Workshop on Enriching Information Retrieval (ENIR 2011)* (pp. 1–3).

Collins-Thompson, K., Bennett, P., Diaz, F., Clarke, C. L., & Voorhees, E. M. (2013). TREC 2013 Web Track Overview. In *NIST Special Publication 500-302: The Twenty-Second Text REtrieval Conference Proceedings (TREC 2013)* (pp. 1–15).

Collins-Thompson, K., Macdonald, C., Bennett, P., Diaz, F., & Voorhees, E. M. (2014). TREC 2014 Web Track Overview. In *NIST Special Publication 500-308: The Twenty-Third Text REtrieval Conference Proceedings (TREC 2014)* (pp. 1–15).

Cooper, H., & Hedges, L. V. (1994). *Handbook of Research Synthesis*. Russell Sage Foundation.

Cormack, G. V., & Lynam, T. R. (2007). 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* (p. 753). Amsterdam: ACM. https://doi.org/10.1145/1277741.1277892

Craswell, N., & Hawking, D. (2003). Overview of the TREC-2002 Web Track. In *NIST Special Publication 500-251: The Eleventh Text REtrieval Conference (TREC 2002)* (pp. 1–10).

Craswell, N., & Hawking, D. (2004). Overview of the TREC-2004 Web Track. In *NIST Special Publication 500-261: The Thirteenth Text REtrieval Conference Proceedings (TREC 2004)* (pp. 1–9).

Craswell, N., Hawking, D., Wilkinson, R., & Wu, M. (2004). Overview of the TREC 2003 Web Track. In *NIST Special Publication 500-255:The Twelfth Text REtrieval Conference (TREC 2003)* (pp. 1–15).

Culpepper, J. S., Mizzaro, S., Sanderson, M., & Scholer, F. (2014). TREC: Topic Engineering Exercise. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1147–1150). Queensland: ACM. https://doi.org/10.1145/2600428.2609531

De Melo, G., & Hose, K. (2013). Searching the web of data. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *LNCS 7814*, 869–873. https://doi.org/10.1007/978-3-642-36973-5_105

Dependent T-Test using SPSS. (2015). Retrieved January 3, 2016, from https://statistics.laerd.com/spss-tutorials/dependent-t-test-using-spss-statistics.php

Dinçer, B. T., Macdonald, C., & Ounis, I. (2014). Hypothesis testing for the risk-sensitive evaluation of retrieval systems. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval - SIGIR '14* (pp. 23–32). Queensland: ACM. https://doi.org/10.1145/2600428.2609625

Fisher, R. A. (1969). Statistical methods for research workers. *Biological Monographs and Manuals*, (V), 356. https://doi.org/https://doi.org/10.1056/NEJMc061160

Fisher, R. A. (1995). *Statistical methods, experimental design, and scientific inference*. (J. H. Bennett, Ed.). Oxford University Press.

Fleiss' kappa. (2017). Retrieved August 20, 2017, from https://en.wikipedia.org/wiki/Fleiss%27_kappa

Golub, K., Soergel, D., Buchanan, G., Tudhope, D., Lykke, M., & Hiom, D. (2016). A Framework for Evaluating Automatic Indexing or Classification in the Context of Retrieval. *Journal Of The Association For Information Science And Technology*, *67*(1), 3–16. https://doi.org/10.1002/asi

Gövert, N., & Kazai, G. (2002). Overview of the INitiative for the Evaluation of XML retrieval (INEX) 2002. In N. Fuhr, N. Gövert, G. Kazai, & M. Lalmas (Eds.), *Proceedings of the First Workshop of the INitiative for the Evaluation of XML Retrieval (INEX)* (pp. 1–17).

Guiver, J., Mizzaro, S., & Robertson, S. (2009). A few good topics: Experiments in Topic Set Reduction for Retrieval Evaluation. *ACM Transactions on Information Systems*, *27*(4), 1–26. https://doi.org/10.1145/1629096.1629099

Gústafsdóttir, G. U. (2017). Readability tests – Siteimprove Help Center. Retrieved August 1, 2017, from https://support.siteimprove.com/hc/en-gb/articles/114094009592-Readability-tests

Gwet, K. L. (2014). *Handbook of inter-rater reliability : the definitive guide to measuring the extent of agreement among raters* (4th ed.). Advanced Analytics, LLC, 2014.

Harman, D. (1992). Overview of the First Text REtrieval Conference (TREC-1). In *NIST Special Publication 500-207: The First Text REtrieval Conference (TREC-1)* (pp. 1–20). Gaithersburg, Maryland: NIST Special Publication 500-207.

Harman, D. (1995). Overview of the Fourth Text REtrieval Conference (TREC-4). In *NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)* (pp. 1–24). Gaithersburg, Maryland: NIST Special Publication 500-236.

Hauff, C., Hiemstra, D., Jong, F., & Azzopardi, L. (2009). Relying on topic subsets for system ranking estimation. In *Proceeding of the 18th ACM conference on Information and knowledge management CIKM 09* (pp. 1859–1862). https://doi.org/10.1145/1645953.1646249

Hauke, J., & Kossowski, T. (2011). Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. *Quaestiones Geographicae*, *30*(2), 87–93. https://doi.org/10.2478/v10117-011-0021-1

Hawking, D. (2001). Overview of the TREC-9 web track. In *NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC-9)* (Vol. 54, pp. 87–102).

Hawking, D., & Craswell, N. (2002). Overview of the TREC-2001 web track. In *Nist Special Publication Sp* (pp. 61–67).

Hawking, D., Craswell, N., & Thistlewaite, P. (1999). Overview of TREC-7 Very Large Collection Track. In *NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC 7)* (pp. 1–13).

Hawking, D., Voorhees, E. M., Craswell, N., & Bailey, P. (1999). Overview of the TREC-8 Web Track. In *NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC 8)* (pp. 1–18).

Hersh, W., Turpin, A., Price, S., Chan, B., Kraemer, D., Sacherek, L., & Olson, D. (2000). Do batch and user evaluations give the same results? In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR'00* (pp. 17–24). Athens,Greece: ACM.

Hiemstra, D., & Graham, M. (2009). Information Retrieval Models. In A. Goker & J. Davies (Eds.), *Information Retrieval : Searching in the 21st Century* (Vol. 65, pp. 1–26). John Wiley & Sons. https://doi.org/10.1017/CBO9781107415324.004

Hopkins, W. G. (2000). Measures of Reliability in Sports Medicine and Science. *Sports Medicine*, *30*(1), 1–15.

Hosseini, M., Cox, I. J., Milic-Frayling, N., Shokouhi, M., & Yilmaz, E. (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* (pp. 901–910). Portland, Oregon: ACM. https://doi.org/10.1145/2348283.2348403

Independent t-test in SPSS Statistics | Laerd Statistics. (2017). Retrieved January 8, 2017, from https://statistics.laerd.com/spss-tutorials/independent-t-test-using-spss-statistics.php

Inter-rater reliability. (2017). Retrieved November 28, 2016, from http://www.cookbook-r.com/Statistical_analysis/Inter-rater_reliability/

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*(4), 422–446. https://doi.org/10.1145/582415.582418

Jayasinghe, G. K., Webber, W., Sanderson, M., Dharmasena, L. S., & Culpepper, J. S. (2015). Statistical comparisons of non-deterministic IR systems using two dimensional variance. *Information Processing & Management*, *51*(5), 677–694. https://doi.org/10.1016/j.ipm.2015.06.005

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR'05* (pp. 154–161). Salvador: ACM. https://doi.org/10.1145/1076034.1076063

Jones, S., & Rijsbergen, C. Van. (1975). Report on the need for and provision of an "ideal" information retrieval test collection. *Computer Laboratory, University of Cambridge*, *46*(December 1975), 73–81.

Jones, T., Thomas, P., Scholer, F., & Sanderson, M. (2015). Features of Disagreement Between Retrieval Effectiveness Measures. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'15* (pp. 847–850). Santiago: ACM. https://doi.org/10.1145/2766462.2767824

Jones, T., Turpin, A., Mizzaro, S., Scholer, F., & Sanderson, M. (2014). Size and Source Matter: Understanding Inconsistencies in Test Collection-Based Evaluation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM'14* (pp. 1843–1846). ACM. https://doi.org/10.1145/2661829.2661945

Kando, N. (2004). Evaluation of Information Access Technologies at NTCIR Workshop. In C. Peters, J. Gonzalo, M. Braschler, & M. Kluck (Eds.), *Comparative Evaluation of Multilingual Information Access Systems. CLEF 2003. Lecture Notes in Computer Science, vol 3237*. Springer, Berlin, Heidelberg.

Kando, N., Kuriyama, K., Nozue, T., Eguchi, K., Kato, H., & Hidaka, S. (1999). Overview of IR Tasks at the First NTCIR Workshop. In *Proceedings of the first NTCIR workshop on research in Japanese text retrieval and term recognition* (pp. 11–44). National Center for Science Information Systems.

Karlgren, J. (2000). The Basics of Information Retrieval: Statistics and Linguistics. *Studies in Mathematical Statistics Theory and Applications*, *18*(1997), 129–131. https://doi.org/10.4135/9781849208741

Kulinskaya, E., Morgenthaler, S., & Staudte, R. G. (2014). Significance Testing: An

Overview. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science* (pp. 1318–1321). Springer Berlin Heidelberg.

Landers, R. (2015). *Computing (ICC) as Intraclass Estimates Correlations of Interrater Reliability in SPSS*. *The Winnower 2:e143518.81744*. https://doi.org/10.15200/winn.143518.81744

Lewandowski, D. (2008). The retrieval effectiveness of web search engines: considering results descriptions. *Journal of Documentation*, *64*(6), 915–937. https://doi.org/10.1108/00220410810912451

Mandl, T. (2008). Recent Developments in the Evaluation of Information Retrieval Systems: Moving Toward Diversity and Practical Applications. *Information Retrieval*.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval*. *Cambridge: Cambridge University Press*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1109/LPT.2009.2020494

Mitra, M. (2008). Overview of FIRE 2008. *Working Notes from FIRE 2008 (FIRE'08)*.

Mizzaro, S., & Robertson, S. E. (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* (pp. 479–486). ACM.

Moffat, A., Scholer, F., & Thomas, P. (2012). Models and metrics:IR Evaluation as a User Process. In *Proceedings of the Seventeenth Australasian Document Computing Symposium on - ADCS '12* (pp. 47–54). https://doi.org/10.1145/2407085.2407092

Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, *27*(1), 1–27. https://doi.org/10.1145/1416950.1416952

NTCIR project overview. (2016). Retrieved September 27, 2016, from http://research.nii.ac.jp/ntcir/outline/prop-en.html

One-way ANOVA - An introduction to when you should run this test and the test hypothesis | Laerd Statistics. (2017). Retrieved August 16, 2017, from https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php

Pavlu, V., Rajput, S., Golbus, P. B., & Aslam, J. A. (2012). IR system evaluation using nugget-based test collections. In *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12* (pp. 393–402). New York, New York: ACM Press. https://doi.org/10.1145/2124295.2124343

Psychometrics. (2017). Retrieved August 20, 2017, from https://en.wikipedia.org/wiki/Psychometrics

Rasmussen, E. (2003). Evaluation in Information Retrieval. *The MIR/MDL Evaluation Project White Paper Collection Edition 3*, 45–49.

Ravana, S. D., & Moffat, A. (2009). Score Aggregation Techniques in Retrieval Experimentation. In *Conferences in Research and Practice in Information Technology Series* (Vol. 92, pp. 59–67).

Robertson, S. (2008). A new interpretation of average precision. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR'08* (pp. 689–690). New York, New York: ACM.

Robertson, S. E., & Kanoulas, E. (2012). On per-topic variance in IR evaluation. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR'12* (pp. 891–900). Portland, Oregon: ACM. https://doi.org/10.1145/2348283.2348402

Robertson, S., Kanoulas, E., & Yilmaz, E. (2010). Extending average precision to graded relevance judgments. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval-SIGIR'10* (pp. 603–610). New York, New York: ACM Press. https://doi.org/10.1145/1835449.1835550

Rubin, A., & Babbie, E. R. (2009). *Essential research methods for social work* (2nd ed.). Brooks/Cole, Cengage Learning.

Rust, J., & Golombok, S. (2015). *Modern psychometrics* (2nd ed.). Routledge.

Ruthven, I. (2014). Relevance behaviour in TREC. *Journal of Documentation*, *70*(6), 1098–1117. https://doi.org/10.1108/JD-02-2014-0031

Sakai, T. (2007). On the reliability of information retrieval metrics based on graded relevance. *Information Processing and Management*, *43*(2), 531–548. https://doi.org/10.1016/j.ipm.2006.07.020

Sakai, T. (2014). Statistical reform in information retrieval? In *ACM SIGIR Forum* (Vol. 48, pp. 3–12). New York, New York: ACM. https://doi.org/10.1145/2641383.2641385

Sakai, T., & Kando, N. (2008). On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval*, *11*(5), 447–470. https://doi.org/10.1007/s10791-008-9059-7

Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends® in Information Retrieval*, *4*(4), 247–375. https://doi.org/10.1561/1500000009

Sanderson, M., & Joho, H. (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 SE-SIGIR'04* (pp. 33–40). Sheffield, South Yorkshire: ACM. https://doi.org/doi: 10.1145/1008992.1009001

Sanderson, M., & Zobel, J. (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* (pp. 162–169). Salvador: ACM. https://doi.org/10.1145/1076034.1076064

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 138–146). Seattle, Washington: ACM.

Scholer, F., Moffat, A., & Thomas, P. (2013). Choices in Information Retrieval Evaluation. In *Proceedings of the 18th Australasian Document Computing Symposium on - ADCS'13* (pp. 74–81). Brisbane, Queensland: ACM.

Scholer, F., Turpin, A., & Sanderson, M. (2011). Quantifying Test Collection Quality Based on the Consistency of Relevance Judgements. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval-SIGIR'11* (pp. 1063–1072). Beijing: ACM. https://doi.org/10.1145/2009916.2010057

Sheskin, D. J. (2011). Parametric Versus Nonparametric Tests. In *International Encyclopedia of Statistical Science* (pp. 1051–1052). Springer Berlin Heidelberg. https://doi.org/https://doi.org/10.3386/w0913

Shi, H., Tan, Y., Zhu, X., & Wu, S. (2013). Measuring stability and discrimination power of metrics in information retrieval evaluation. In Yin H. et al. (Ed.), *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 8206 LNCS, pp. 8–15). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41278-3_2

Shokouhi, M., Craswell, N., & Robertson, S. (2009). Are Evaluation Metrics Identical With Binary Judgements? *Learning*, *10*.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Shuttleworth, M. (2009). Internal Consistency Reliability. Retrieved August 20, 2017, from https://explorable.com/internal-consistency-reliability

Smith, E. A., & Senter, R. J. (1967). Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (6570th)*, 1–14.

Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management-CIKM'07* (pp. 623–632). Lisbon: ACM. https://doi.org/10.1145/1321440.1321528

Smucker, M. D., Allan, J., & Carterette, B. (2009). Agreement Among Statistical Significance Tests for Information Retrieval Evaluation at Varying Sample Sizes. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval-SIGIR'09* (Vol. 2, pp. 630–631). Boston, MA: ACM. https://doi.org/10.1145/1571941.1572050

Smucker, M. D., & Clarke, C. L. A. (2012). Time-based calibration of effectiveness measures. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR'12* (p. 95). Portland,

Oregon: ACM. https://doi.org/10.1145/2348283.2348300

Soboroff, I., Nicholas, C., & Cahan, P. (2001). Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval-SIGIR'01* (pp. 66–73). New Orleans, Louisiana: ACM. https://doi.org/10.1145/383952.383961

Statistics | Definition of Statistics by Merriam-Webster. (2017). Retrieved July 24, 2017, from https://www.merriam-webster.com/dictionary/statistics

Stefan, B., Clarke, C. L. A., Yeung, P. C. K., & Soboroff, I. (2007). Reliable Information Retrieval Evaluation with Incomplete and Biased Judgements. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR'07* (pp. 63–70). Amsterdam: ACM. https://doi.org/10.1145/1277741.1277755

The CLEF Initiative. (2016). Retrieved September 27, 2016, from http://www.clef-initiative.eu/

TREC 2004 Robust Track Guidelines. (2005). Retrieved July 23, 2017, from http://trec.nist.gov/data/robust/04.guidelines.html

Trochim, W., Donnelly, J. P., & Kanika, A. (2015). *Research Methods: The Essential Knowledge Base* (2nd ed.). Cengage Learning.

Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, *24*(3), 478–514. https://doi.org/10.1007/s10618-011-0238-6

Turpin, A. H., & Hersh, W. (2001). Why batch and user evaluations do not give the same results. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval-SIGIR'01* (pp. 225–231). New Orleans, Louisiana: ACM. https://doi.org/10.1145/383952.383992

Two-way ANOVA in SPSS Statistics - Step-by-step procedure including testing of assumptions | Laerd Statistics. (2017). Retrieved August 16, 2017, from https://statistics.laerd.com/spss-tutorials/two-way-anova-using-spss-statistics.php

Types of reliability. (2017). Retrieved August 20, 2017, from http://changingminds.org/explanations/research/design/types_reliability.htm#par

Urbano, J., Marrero, M., & Martín, D. (2013a). 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* (Vol. 308, pp. 925–928). Dublin: ACM. https://doi.org/10.1145/1321440.1321528

Urbano, J., Marrero, M., & Martín, D. (2013b). On the Measurement of Test Collection Reliability. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR'13* (pp. 393–402). Dublin: ACM. https://doi.org/10.1145/2484028.2484038

Varathan, K. D., Sembok, T. M. T., Kadir, R. A., & Omar, N. (2014). Semantic Indexing For Question Answering System. *Malaysian Journal of Computer Science*, *27*(4), 261–274.

Verma, M., Yilmaz, E., & Craswell, N. (2016). On Obtaining Effort Based Judgements for Information Retrieval. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining - WSDM '16* (pp. 277–286). San Francisco, California: ACM. https://doi.org/10.1145/2835776.2835840

Villa, R., & Halvey, M. (2013). Is relevance hard work?: Evaluating the effort of making relevant assessments. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval-SIGIR'13* (pp. 765–768). Dublin: ACM.

Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, *36*(5), 697–716. https://doi.org/10.1016/S0306-4573(00)00010-8

Voorhees, E. M. (2001). Evaluation by Highly Relevant Documents. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval-SIGIR'01* (pp. 74–82). New Orleans, Louisiana: ACM.

Voorhees, E. M. (2002). Overview of TREC 2002. In E. . Voorhees & L. P. Buckland (Eds.), *NIST Special Publication 500-251: The Eleventh Text REtrieval Conference (TREC 2002)* (pp. 1–13). Gaithersburg, Maryland.

Voorhees, E. M. (2002). The Philosophy of Information Retrieval Evaluation. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), *Evaluation of Cross-Language Information Retrieval Systems. CLEF 2001. Lecture Notes in Computer Science, vol 2406* (pp. 355–370). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45691-0_34

Voorhees, E. M. (2004). Overview of the TREC 2004 Robust Retrieval Track. In E. M. Voorhees & L. P. Buckland (Eds.), *NIST Special Publication: SP 500-261 The Thirteenth Text Retrieval Conference (TREC 2004)*. Gaithersburg, Maryland.

Voorhees, E. M. (2005). Overview of the TREC 2005 Robust Retrieval Track. In E. M. Voorhees & L. P. Buckland (Eds.), *NIST Special Publication 500-266: The Fourteenth Text REtrieval Conference Proceedings (TREC 2005)*. Gaithersburg, Maryland. https://doi.org/10.1145/1147197.1147205

Voorhees, E. M. (2007). Overview of TREC 2007. In E. M. Voorhees & L. P. Buckland (Eds.), *NIST Special Publication 500-274: The Sixteenth Text REtrieval Conference Proceedings (TREC 2007)* (pp. 1–16). Gaithersburg, Maryland.

Voorhees, E. M., & Buckley, C. (2002). The effect of topic set size on retrieval experimental error. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval-SIGIR'02* (pp. 316–323). Tampere: ACM. https://doi.org/10.1145/1076034.1076166

Voorhees, E. M., & Harman, D. (1999). Overview of the Eigth Text Retrieval Conference (TREC-8). In E. M. Voorhees & D. K. Harman (Eds.), *NIST Special Publication*

500-246: The Eighth Text REtrieval Conference (TREC 8) (pp. 1–24). Gaithersburg, Maryland.

Voorhees, E. M., & Harman, D. (2000). Overview of the Ninth Text REtrieval Conference (TREC-9). In E. M. Voorhees & D. K. Harman (Eds.), *NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC-9)* (pp. 1–14). Gaithersburg, Maryland.

Voorhees, E. M., & Harman, D. (2001). Overview of TREC 2001. In E. M. Voorhees & D. K. Harman (Eds.), *NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001)* (pp. 1–15). Gaithersburg, Maryland.

Voorhees, E. M., & Harman, D. K. (2005). TREC: Experiment and Evaluation in Information Retrieval. *Digital Libraries and Electronic Publishing*, *1*. https://doi.org/10.1162/coli.2006.32.4.563

Watters, C. (1999). Information retrieval and the virtual document. *Journal of the American Society for Information Science*, *50*(11), 1028–1029. https://doi.org/10.1002/(SICI)1097-4571(1999)50:11<1028::AID-ASI8>3.0.CO;2-0

Webber, W., Moffat, A., & Zobel, J. (2010). The Effect of Pooling and Evaluation Depth on Metric Stability. In *The 3rd International Workshop on Evaluating Information Access (EVIA 2010)* (pp. 7–15). Tokyo.

Webber, W., Moffat, A., Zobel, J., & Sakai, T. (2008). Precision-at-ten Considered Redundant. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval-SIGIR'08* (pp. 695–696). Singapore: ACM. https://doi.org/10.1145/1390334.1390456

Webber, W., Toth, B., & Desamito, M. (2012). Effect of written instructions on assessor agreement. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval-SIGIR'12* (pp. 1053–1054). Portland, Oregon: ACM. https://doi.org/10.1145/2348283.2348465

Yilmaz, E., & Aslam, J. A. (2006). Estimating Average Precision with Incomplete and Imperfect Judgments. In *Proceedings of the 15th ACM international conference on Information and knowledge management-CIKM'06* (Vol. 16, pp. 102–111). Arlington, Virginia: ACM. https://doi.org/10.1007/s10115-007-0101-7

Yilmaz, E., Verma, M., Craswell, N., Radlinski, F., & Bailey, P. (2014). Relevance and Effort: An Analysis of Document Utility. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14* (pp. 91–100). Shanghai: ACM. https://doi.org/10.1145/2661829.2661953

Zobel, J. (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* (pp. 307–314). Melbourne: ACM. https://doi.org/10.1145/290941.291014

LIST OF PUBLICATIONS AND PAPERS PRESENTED

Journal papers:

1. Rajagopal, P., Ravana, S.D., Srivastava, H. and Taheri, M.S (2017). Document level assessment of document retrieval systems in a pairwise system evaluation. *Information Research, 22*(2), paper 752. Retrieved from http://InformationR.net/ir/22-2/paper752.html (Archived by WebCite® at http://www.webcitation.org/6r2QsbQ2T) – **Published**

2. Rajagopal, P and Ravana, S.D. (2017). Measuring the Reliability of Ranking in Information Retrieval Systems Evaluation, *Malaysian Journal of Computer Science* – **Accepted**

3. Ravana, S.D., Rajagopal, P., Balakrishnan, V. (2015). Ranking retrieval systems using pseudo relevance judgments, *Aslib Journal of Information Management*, *67*(6), pp.700-714, https://doi.org/10.1108/AJIM-03-2015-0046 – **Published**

4. Rajagopal, P., Ravana, S.D. & Ismail, M.A. (2014). Relevance judgments exclusive of human assessors in large scale information retrieval evaluation experimentation. *Malaysian Journal of Computer Science*, *27*(2), pp.80-94. – **Published**

Conference proceedings:

5. Rajagopal, P., & Ravana, S. D. (2016). Document level assessment for pairwise system evaluation. In *2016 3rd International Conference on Information Retrieval and Knowledge Management, CAMP 2016 - Conference Proceedings* (pp. 77–81). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/INFRKM.2016.7806339 – **Published**