Cairo University
Institute of Statistical Studies & Research
Department of Computer &Information Sciences

BUILDING RELEVANCE JUDGMENTS WITHOUT POOLING

by
AbdelRahim AbdelSabor AbdelHalim Mohammed

Supervised by

Prof. Mahmoud Riad Mahmoud
Institute of Statistical Studies and Research

Dr. Kareem Darwish
Faculty of Computer & Information

A thesis submitted to the Institute of Statistical and Research, Cairo University,
In partial fulfillment of the requirements for the degree of
Master of Science in
Department of Computer and Information Sciences

Cairo University

Institute of Statistical Studies & Research

Department of Computer &Information Sciences

Approval Sheet

BUILDING RELEVANCE JUDGMENTS WITHOUT POOLING

by

AbdelRahim AbdelSabor AbdelHalim Mohammed

A thesis submitted to the Institute of Statistical and Research, Cairo University,
In partial fulfillment of the requirements for the degree of
Master of Science in
Department of Computer and Information Sciences

Approved by the Examining Committee

Prof. Mahmoud Riad Mahmoud

Prof. Mohsen Rashwan

Prof. Mervat Gheith

TABLE OF CONTENT

LIST OF TABELS III

LIST OF FIGURES.................. III

ACKNOWLEDGMENTS...IV

ABSTRACT V

CHAPTER I INTRODUCTION 1

1.1 Introduction 2

1.2 Problem Definition... 3

1.3 Objectives 4

1.4 Contributions and Methodology of Work 4

1.5 Testing and Evaluation 5

1.6 Thesis Outlines 6

CHAPTER II BACKGROUNDS 7

2.1 What is Information Retrieval (IR)? 8

2.1.1 Brief History of IR... 8

2.1.2 IR Objective 10

2.2 Information Retrieval System... 11

2.3 IR System Keys... 12

2.3.1 Documents Indexing.. 12

2.3.2 Query.. 19

2.3.3 Ranked List 19

2.3.4 Relevance 20

2.3.5 Relevance Feedback “Query Expansion” 21

2.4 Retrieving Models.. 22

2.4.1 Vector Space Model 23

2.4.2 KL-Divergence Model (KL-Div)... 25

2.4.3 OKAPI BM25 Model 27

2.5 IR Evaluation 28

2.5.1 Mean Average Precision (MAP).. 29

2.5.2 R-Precision.. 31

2.5.3 Bpref.. 32

2.6 Information Retrieval & Arabic 33

2.7 Documents Clustering 35

CHAPTER III BUILDING RELEVANCE JUDGMENTS 38

3.1 Introduction 39

3.1 Exhaustive Search 39

3.2 Pooling Method 40

3.3 Interactive Searching and Judging (ISJ) 41

3.4 Interactive Relevance Feedback and Judging 42

3.5 Sanderson and Hedio Method (2004) 43

3.5 Incremental Test Collections 44

3.6 Minimal Test Collections 45

THAPTER IV RELEVANCE FEEDBACK CLUSTERING 46

4.1 Introduction 47

4.2 The RFC Method: 47

4.3 Experiments: 50

4.2.1 Document Collection and Topics 50

4.2.2 IR system 52

4.2.4 Evaluation 53

CHAPTER V RESULTS AND DISCUSSION 54

CONCLUSION & FUTURE WORK 60

REFERENCES 63

LIST OF TABLES

Table 5.1: Indexing using words as index terms 55

Table 5.2: Indexing using character 4-gram as index terms 56

Table 5.3: Indexing using character 3-gram as index terms 57

LIST OF FIGURES

Figure 2.1 Basic IR System. 11

Figure 2.2 Basic idea of the inverted index ……………………………………….…..13

Figure 2.3 Stop-words list example ……………….………………………………...16

Figure 2.4 Tokenization Example……………………..……………….……………16

Figure 2.5 Building an inverted index ……………………...…………..…………18

Figure 2.6 Google search ranked-list ……………………………...…….………...21

Figure 2.7 Cosine similarity ……………………………………………...……….24

Figure 2.8 Precision & Recall Measures ………………………………….…….…...29

Figure 3.1 Exhaustive Search ……………………………………………….……....33

Figure 3.2 Creating relevance judgments with pooling ……………….….…………34

Fugure 3.3 Sanderson & Hedio Method ……………………………….…….….…...43

Figure 4.1 The search component of RFC ……………………….…….………….48

Figure 4.2 Flow chart or the RFC Method …...………………………….….….......48

Figure 4.3 Clustering relevant documents example ………………….................42

Figure 4.4 TREC 2002 CLIR Document ………………………………….………...51

Figure 4.5 TREC 2002 CLIR Topic ……………………………………….....…51

ACKNOWLEDGMENTS

In the beginning, I would like to express all praise to Allah for His grace and mercy. Then, I would like to express sincere appreciation to my parents AbdelSabor El-Madany and Hadya Raee and my brothers Mohammed, Hassan, my sisters Aya, Mona, Wafaa, and my fiancee Radwa. I also would like to thank my best friends Hossam Sayed, Ahmed Omran, Mamdouh Saber, and Mahmoud Rashad.

Heartfelt gratitude and special thanks are due to Dr. Kareem Darwish who dedicated countless hours teaching and mentoring me. I also extend my gratitude to my advisor Prof. Mahmoud Riad for his help and valuable advice and to all the members of my dissertation committee, Prof. Mohsen Rashwan and Prof. Mervat Gheith.

ABSTRACT

A novel method is represented in this thesis for building query relevance judgments without system pooling using subtopic clustering in conjunction with relevance feedback. The new method is referred to as Relevance Feedback Clustering (RFC).

RFC builds on a previously developed method [*Sanderson & Hedio, 2004*] that uses relevance feedback to replace manual interactive query reformulation.

RFC method is shown to be robust in building relevance judgments even in the absence of proper text processing tools, as demonstrated for Arabic with minimal processing, and to be consistently better than the one suggested by Sanderson and Hedio. RFC was applied to the TREC 2002 CLIR test collection, which contains Arabic newswire articles from 'Agence Française de Presse' (AFP).

The thesis also reports the conditions under which the produced relevance judgments and official TREC relevance judgments rank different systems in ways that highly correlate.

The experimental results show that using the new method with OKAPI BM25 weighting and incrementally increasing expansion terms produces good relevance judgments that rank different systems in a way that highly correlates with the ranking produced by the official TREC relevance judgments.

In addition, using bpref seems to slightly improve the comparison between systems over MAP and is more tolerant of changes in the relevance judgments. The success of the method suggests that feedback using subtopic clusters

individually more successfully probes the subtopics to find more relevant documents.

On the other hand, using all documents from all subtopics together in feedback may dilutes the effectiveness of feedback (or perhaps confuses the IR system that tries to find loosely related, as opposed to closely related, documents at once).

Lastly, using clustering in interactive retrieval applications where users merely need to mark the first occurring document of a cluster in the ranked list as relevant to indicate the system that other documents in the same cluster are potentially relevant can potentially improve feedback in such an applications.

CHAPTER I

INTRODUCTION

1.1 Introduction

The vast amounts of information make it an attractive resource for answering a variety of queries that users may have. Information Retrieval is one of the most common approaches that allow users to find information on the huge resources where a user can specify a string of keywords and expect to retrieve relevant documents, possibly ranked by their relevance to the query.

Relevance is the matching of a document with an information need expressed through a query. Relevant documents determined by human assessors for document-query pairs are called Relevance Judgments. Relevance judgments are important parts of constructing information retrieval collections and perhaps the most costly to construct [*Sanderson & Hedio, 2004*].

The information retrieval collection "Test Collection" consists of a set of documents, a set of topics (statements of information need), and a set of relevance judgments that list which documents are relevant to which topics. The cost and effort associated with building relevance judgments significantly outweighs the cost of collecting document and constructing topics.

Initial attempts for building "complete" relevance judgments focused on exhaustively searching for "all" relevant documents in a set of document for each topic. The amount of manual labor required for exhaustive search placed a limitation on the possible sizes of document sets, and most of the resulting test collections were small. The main advantage of this method is that "all" relevant documents are found, but this process is labor intensive and impractical for large collections [*Sanderson & Hedio, 2004*].

To avoid exhaustive search, a new method called "pooling" was employed. Pooling combines the top K (typically K = 100) search results from varying retrieval systems and removes duplicates from a pool. The documents in this pool are the only documents that would be examined for relevance [*Sparck-Jones et al., 1975*]. It was assumed that nearly all relevant documents would be found in the pool. The randomized documents in a pool would be manually assessed for relevance, thereby forming the relevance judgments set.

Pooling is widely used in different evaluation frameworks such as the Text REtreival Conference (TREC). Organizing groups that contribute to such pools requires a level of organization beyond what most researchers are able to provide. Thus, recent work has focused on constructing IR test collections with limited system pooling [*Cormack et al., 1998*] or without system pooling [*Soboroff & Robertson, 2003*], [*Sanderson & Hedio, 2004*], [*Carterette et al, 2005; 2006*]. The techniques involve the manual or automatic reformulation of queries in an iterative search process.

1.2 Problem Definition

Building relevance judgments is the most expensive and laborious part of building retrieval test collections. Common approaches for creating standard tests such as pooling, which is the most common, are generally very expensive. Some proposed methods build relevance judgments without pooling. Existing methods for building relevance judgments without pooling have not been tested for less studied languages such as Arabic and hence their reliability is not certain.

If existing methods do not show sufficient reliability for less studied languages, they would need to be modified or further developed to attain the desired level of reliability.

1.3 Objectives

The main objective of this thesis is to devise a robust repeatable method for building IR test sets cheaply and rapidly, even for languages in which little is known about how to effectively rank and retrieve such as Arabic. Hence, limited linguistic resources or prior knowledge of the language are required.

The proposed method is tested on the TREC CLIR 2002 (Arabic Collection) to ascertain the reliability of the proposed method.

1.4 Contributions and Methodology of Work

This thesis proposes a novel technique for building relevance judgments without system pooling that improves on existing techniques through the use of subtopic clustering. The technique will be referred to as the Relevance Feedback Clustering (RFC). The intuition behind this approach is that most topics contain one or more subtopics and documents typically address subtopics as opposed to whole topics. For example, a topic about "the civil war in Iraq" can be divided into subtopics involving "the role of Sunni insurgency in the war," "the role of Iran and Syria," "the effect of American occupation," and so forth. Each subtopic would have specific words that are frequent and/or unique to it. Therefore, if documents addressing specific subtopics can be used in reformulation of the initial query, the resulting query can better probe the subtopic to find more relevant documents [*AbdelSabor et al., 2007].*

1.5 Testing and Evaluation

The proposed technique is tested using the TREC 2002 Cross-Language IR (CLIR) Arabic test collection using two setups. In the first setup, minimal Arabic processing is employed to stress test the technique. This setup is used to simulate a language where researchers have limited linguistic resources or knowledge. In the second setup, effective processing is employed to ascertain the full potential of the technique. Arabic was chosen because it poses unique challenges in orthography and morphology that complicate IR in general and the potential creation of relevance judgments. Such challenges led to problematic relevance judgments for the TREC 2001 CLIR collection [*Gey & Oard, 2001*] [*Oard & Gey, 2002*]. The problems stemmed from single contributions to the pool containing large numbers of unique documents not found by any other group. This suggests that the relevant documents found were grossly incomplete [*Buckley & Voorhees, 2004*].

1.6 Thesis Outlines

The rest of the thesis will be organized as follows:

Chapter 2: presents a general background on information retrieval system,

Chapter 3: surveys existing literature on building relevance judgment's methods,

Chapter 4: presents relevance feedback clustering method, experimental setup, and evaluation,

Chapter 5: discusses the experimental results, and

Chapter 6: concludes the thesis and presents future work.

CHAPTER II
BACKGROUNDS