UNIVERSITI TEKNOLOGI MARA

EVALUATION OF QUERY REFORMULATION STRATEGIES FOR DOMAIN-SPECIFIC INFORMATION SEARCHES: A CASE STUDY OF THE DURIAN FRUIT DOMAIN

AZILAWATI BINTI AZIZAN

Thesis submitted in fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information Technology and Quantitative Sciences)

Faculty of Computer and Mathematical Sciences

September 2022

ABSTRACT

Although search engine technologies have made great strides in helping users find information on the Web, search results are only as good as the keywords and phrases that users use in the search query. Hence, search queries need to precisely formulated. However, users often fail to accurately translate their information needs into correct query words or phrases for a search engine to utilise. This becomes harder when users search for domain-specific information as, in most cases, users are unable to identify the keywords that are appropriate for the domain in the search query. As such, the search engine is unable to locate the relevant documents. This causes users to reformulate the query multiple times in the hopes of retrieving a more relevant set of search results. To address this issue, many researchers propose the use of query reformulation, query refinement, query expansion, or query disambiguation to intentionally build better queries and retrieve more relevant results. However, most of strategies employed to tackle this issue; such as the query log, rhetorical structure, thesaurus, WordNet, ontology, and user profiles; require extensive sources, risky and are time consuming. Therefore, more effective and simpler techniques are needed to obtain better search results as well reduce the need of query reformulation (QR). To that end, this study applied a search engine framework which employs standard methodology in Information Retrieval (IR) to evaluate several reformulation strategies and proposes an operative and effective QR strategy to locate domain-specific information. The fruit domain; specifically, durian; was chosen as the case study. An investigation was first conducted to prove that the issues present at the time of the study as well as the selected domain were still pertinent. Several popular commercial search engines were examined to determine their current search performance in locating domain-specific information on the Web. A group of users was then selected to conduct a task-based search to examine how users structured their queries to obtain the search intent. The results indicated that the most popular search engine (Google) only had an average of P@10 score of 0.463 and mean average precision (MAP) score of 0.649 when searching for durian-related information. The results of the task-based search showed that 84.82% of users reformulate their queries, clearly indicating that users do not obtain relevant search results on the first few tries. As such, several QR strategies that may produce better search results were investigated. Nine strategies were examined by using features, such as query keywords, ontology, the characteristic category of the domain, and the domain name. These features were manipulated using techniques, such as ‘generalization’, ‘specification’, and ‘new’. Of the nine strategies examined, three outperformed the baseline. Combining query keywords with ontology significantly surpassed the baseline MAP score by 2.65%. More interestingly, the characteristic category of the domain, which is considerably simpler and easier to use, also outperformed the baseline MAP score by 2.63%. The findings of this study contribute to the field of IR, through the performance of search engines, user behaviour, test collection and reformulation strategies in searching for domain-specific information.

ACKNOWLEDGEMENT

All praise and infinite gratitude always and forever be to the Almighty Allah, the Creator, the Knower of All, the Merciful for showering me a good experience throughout this study and for all that has bestowed on me. It is with His ascendancy this study is completed. May His endless mercies and blessings be upon his Prophet Muhammad SAW.

I would like to take this opportunity to thank a number of people for their help and support during my study. First and foremost my heartiest gratitude and gratefulness goes to my former supervisor Prof. Dr. Zainab Abu Bakar whose patience and trust in the research and wise supervision have benefited me greatly. She who encouraged me the freedom of individual effort that was essential to fulfilling my work and who provided all conceivable aid in the pursuit of this study.

Next I would like to express my sincere gratitude to my present supervisor, Assoc. Prof. Dr. Nurazzah Abdul Rahman for her unwavering trust, support and motivation that she's always shown me despite her busy schedule and commitments. I really appreciate her commitments.

It is also a pleasure to acknowledge Prof. Dr. Shahrul Azman Mohd Noah for his co-supervision, suggestion and encouragement during this study. His professionalism in supervising my research studies has inspired this PhD journey.

I would also like to express my deep appreciation to the Ministry of Higher Education (MOHE) and the University Technology of MARA (UiTM) for funding my study and granting me the study leave that facilitated the attainment of my PhD degree. Many thanks also go to the lecturers and supporting staff for their most generous assistance.

Last but not least, my thankful gratitude to my husband, my children, my mother, my late father, the whole family, and friends wherever they may be for their numerous acts of kindness, support, prayers, and constant encouragement for endurance and patience.

Thank you very much.

TABLE OF CONTENTS

CHAPTER ONE
INTRODUCTION

1.1 Introduction

As the amount of content on the Web grows exponentially, so does the number of new and inexperience users. This only poses new challenges in information retrieval (IR) as it makes search activities more complex. As search engine technologies are evolving positively to mitigate these issues, most people believe that searching for information on the Web is easy as long as you have a Google search engine. While this is true to some extent, a search engine can only return highly relevant search results if the user can accurately and correctly express and transform their information needs into a search query. Otherwise, the user will receive tons of irrelevant results (Cao, Chen, Baltes, Treude, & Chen, 2021). Therefore, the queries submitted by users affect the search experience (Chen et al., 2021).

Unfortunately, existing studies indicate that users struggle to formulate accurate queries (Zeboudj, 2020; Huang & Efthimiadis, 2009). This causes them to reformulate the same query multiple times in order to obtain more relevant search results. These searches become tougher when users attempt to retrieve domain-specific information as it requires the exploration of information pertaining to a specific area of knowledge in the domain. Domain-specific searches also require more facts, more complex task-oriented information as well as a wider range of search strategies (G. H. Yang, Sloan, & Wang, 2016). This also involves intensive browsing and idea discovery during the search. Therefore, there is a need to address QR issues, especially in the context of domain-specific searches to help users as well as search providers.

To that end, this study examined and provided empirical evidence on QR strategies that may provide better search results when locating domain-specific information on the Web. Several investigations were conducted to determine the current issues faced during searches as well examine existing QR strategies. This study focused on QR for domain-specific searches on the Web. The fruit domain was chosen as the scope while the durian fruit was chosen as the domain-specific knowledge to be retrieved.