UNIVERSITI TEKNOLOGI MARA

HIERARCHICAL FUZZY LOGIC RANKING FUNCTION FOR MALAY TRANSLATED HADITH INFORMATION RETRIEVAL SYSTEM

SHAIFUL BAKHTIAR BIN RODZMAN

Thesis submitted in fulfilment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science)

Faculty of Computer and Mathematical Sciences

September 2022

ABSTRACT

The ranking function is a predictive algorithm that is used to establish a simple ordering of documents according to their relevance. It is a critical step in Information Retrieval (IR) because the quality of the results in IR System is fundamentally dependent on it. Several IR Models are proposed over the years in an attempt to yield the best results in IR. One of them is the BM25 Model. It improves the ranking process by using the term-frequency factor as a multiplier in the ranking function. There is a growing consensus that BM25 yields the best results compared to the Vector Space Model specifically for general collections. Hence, it is widely used as a baseline model for new ranking method evaluation in substitution to the vector model. Just like other IR models, many documents retrieved by BM25 are irrelevant to the user. The majority of existing IR systems often base their retrieval judgment solely on query representation and document collections. Other aspects like retrievability and quality indicators of the documents, accuracy of the top results, and human judgment are largely ignored, especially in Malay Translated Hadith IR System that has a common problem like Fabrication of Hadith. The literature review discovered no research has applied the complete integration of retrievability improvement of the documents, the quality indicator in Hierarchical Fuzzy Logic System, and document ranking in producing and improving results of the relevant Malay Translated IR system. Hence, this study has three objectives. First, to develop an Ontology Concept of Malay Translated Hadith Document and utilize its information in the calculation of novel Ontology BM25 Score in order to improve the retrievability of Ranking Function. Second, to develop a Fuzzy Logic Controller of Mamdani-type Fuzzy Inference System based on the BM25 Model in the Malay IR that includes positive and negative quality ranking indicators and improves the results by using positive documents and demoting negative documents. The quality indicators of 1) Ontology BM25 Score, 2) Fabrication Rate of Hadith, 3) Shia Rate of Hadith, 4) Positive Rate of Hadith, and 5) Expert Judgment with Z-Numbers are also introduced. Third, to evaluate the results by comparing documents from Hadith experts with the result from BM25 Model original score and Vector Space Model. The data set are collected from Malay Translated Sahih Bukhari Hadith, Malay Translated Fabricated Hadith, and Shia Malay Translated Hadith domains. The researcher improved the original BM25 ranking function by increasing the documents' retrievability and introducing five new ranking indicators to the ranking function score. The search results confirm the improvement in Recall (Retrievability) and Precision@10 and Mean Average Precision/MAP (ranking function score). Out of 30 queries, the proposed research yielded better retrievability result in 10 queries compared to BM25 original score with just only one better result. In average the proposed research achieved the value of Recall, 0.7998 compared to BM25 original score, 0.7813 from the perfect score 1. In terms of ranking, out of 30 queries, the proposed research yielded better result on 27 queries on all evaluation metrics such as Precision at Rank 10, %no measures and MAP compared to the BM25 original score which only yielded better result in two queries. In average the proposed research achieved the value of Precision at Rank 10, 0.9413 compared to BM25 original score, 0.5044 from the perfect score 1. In conclusion, the new five quality indicators with the Hierarchical Fuzzy Logic Controller have the potential to improve the results of Malay Translated Hadith IR.

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

CHAPTER ONE
INTRODUCTION

1.1   Research Background

Research in Malay Information Retrieval started around twenty-one years ago in 1995. It marked a new area in Computer Science. The research keeps growing because of the increasing number of Malay language documents and the variety of available documents with the theme of information characteristics and its features that have the potential to be utilized in the document retrieval process as well as the document ranking of Information Retrieval. According to Nurazzah et al. (2010), this scenario proves the need for reliable and effective Malay Text Document Retrieval methods since the tools provide assistance for the user with the aim to locate documents that satisfy their specific needs (Agbele, 2014).

As a summation, Information retrieval (IR) focuses on the searching of documents for information that satisfies a user's need, along with the document representations. In Traditional IR, document representations are made by a document indexing process (extracting meaningful term) that provides a logical view of the documents. In the searching process, the user query will be processed in the same manner to produce the request representation and the representation of the document will be matched according to specific matching conditions (rules). Then the results of the searching process will be presented to the user in a form of a ranked list that contains the most relevant documents (Baeza-Yates & Ribeiro-Neto, 2011).

Baeza-Yates and Ribeiro-Neto (2011) defined the ranking process as the most important aspect because it can show the effectiveness and accuracy of Information Retrieval. Two complex processes are involved in the ranking process: (a) the conception of a logical framework for representing documents and queries and (b) the definition of the ranking function that computes a rank for each document with regard to a given query. The logical framework directly affects to the computation of document ranks, which are then used to sort the document ranks in response to a given query. This means the definition of ranking function is related to the logical framework that represents the document, such as matching query and document process stated in Othman (1993) by the task of indexing, index language, document clustering