UNIVERSITI TEKNOLOGI MARA

GRAPH-BASED AMBIGUITY HANDLING IN THE CHAIN OF NARRATORS FOR HADITH INFORMATION RETRIEVAL

NURSYAHIDAH BINTI ALIAS

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

College of Computing, Informatics and Mathematics

December 2023

ABSTRACT

Information Retrieval (IR) is essential in retrieving information automatically and quickly. Therefore, a wide range domain area applies, and testing the IR models with a standard dataset is provided. Besides, much attention is given to hadith data sets in retrieving information automatically. Hadith is divided into a sanad and matan. These parts are the essential factors in hadith. Sanad refers to a chain of narrators. A degree of authenticity in hadith can be determined by the chain of narrators which are: sahih (sound), hasan (good), and da'if (weak). Hadith IR focuses primarily on the matan part, which retrieves hadith based on a topic that is more appropriate with the existing traditional techniques in IR, which are terms and frequency. However, current research in the hadith IR working with the sanad part is in indexing individual narrator names in the chain of narrators. The indexing could be used in retrieving hadith documents. However, the document cannot be retrieved if a user needs to retrieve the hadith according to the authenticity determined by the chain of narrators, not the individual narrator. Therefore, this thesis concerns hadith document retrieval based on a chain of narrators. Three main studies were conducted, each with a different and specific aim. The first was a study conducted by a literature review of traditional indexing and retrieval techniques in IR and hadith IR, the chain of narrators, an ambiguous element in the chain of narrators, and existing NLP tools for Malay language and hadith IR. The first study's findings led us to draw a conceptual model of hadith IR that handles ambiguous elements in the chain of narrators based on a graph. The second study aims to design a graph-based indexer and rule-based hadith IR. The findings in the second study resulted in a test collection, indexer, and retrieval design based on 1000 hadith from the Shahih Bukhari Book validated by hadith experts. Two indexers designed are the name entity representation and the pair name entity representation. Then, both indexers were implemented in the rule-based retrieval. The first rule-based retrieval implements the inverted index using the name entity representation named RRN. The second rule-based retrieval implements the inverted index using the pair name entity representation named RRPN. The third study uses fifty queries to evaluate the proposed rule-based hadith IR: RRN and RRPN. The evaluation was conducted by testing the fifty queries in the traditional IR model, the Boolean Model, RRN, and RRPN to get a retrieved result. Then, each retrieved result calculates the recall, precision, and F-measure. Next, the statistically significant difference among the Boolean Model, RRN, and RRPN is tested using the sign test. Two pairs of observations are investigated for their significant difference: i)The effectiveness of the Boolean Model against RRN, ii)The effectiveness of the Boolean Model against RRPN, and iii)The effectiveness of the RRN against RRPN. The sign test proves that RRN and RRPN retrieval are able to generate significantly better results than the Boolean Model at the 5% level. The sign test also proves no significant difference in effectiveness between RRPN and RRN at the 5% level. But, the sign test also proves that the RRPN can generate significantly better results than the RRN at the 10% level. The results show that there is an improvement in the effectiveness of RRN and RRPN compared to the Boolean Model. Accordingly, it will benefit future researchers and designers toward better development of hadith retrieval based on a chain of narrators. The future system would help hadith scholars reduce time in the Takhrij al-Ḥadīth process, which refers to the investigation of hadith into the original book with the complete chain of narrators in retrieving hadith documents with similar chains.

ACKNOWLEDGEMENT

Firstly, Alhamdulillah, praise Allah (S.W.T) for allowing me to embark on my Ph.D. and complete this long and challenging journey.

First, my special thanks go to my supervisor, Professor Madya Dr. Nurazzah Abd Rahman, and co-supervisor, Dr. Normaly Kamal Ismail, who was willing to share their experience and invaluable insights and guidance in completing this journey. Thank you for never giving up on me. I would also like to express my special thanks to all the experts for their cooperation throughout the study.

Finally, this thesis is dedicated to the loving memory of my mother, , and my very dear late father, Alias bin Chik, for his vision and determination to educate me. This piece of victory is dedicated to both of you. Thank you very much to my husband, Safrizal bin Kabir, my sibling and their family, and my in-law family, who always supported me in the good and bad times. To my kid, thank you for giving me your time to complete this thesis, and Insya Allah, we will spend more time together after this. Alhamdulillah.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Background of the Study

Information retrieval (IR) is essential in retrieving documents automatically in many domains. It is because the IR classical models are fundamentally based on texts and use the text of the documents to rank them. The text is words represented as a term, and it is independent. The traditional retrieval strategy is based on set theory, algebraic and probabilistic, known as Boolean, Vector, and Probabilistic Models. The indexing of those models treats the text as a string with no particular structure. However, information on the structure might be important to the user for particular searches (Baeza-Yates & Ribeiro-Neto, 2011).

IR has a standard dataset, which is Text Retrieval Conference (TREC). The standard dataset comes with standard relevant judgment and evaluation. Hadith IR is a research area whose dataset is not the standard dataset. Hadith IR has been divided into two areas in the hadith part: the sanad and matan parts. It is because, in terms of Hadith science, determining the authenticity of a hadith can be accomplished by examining its sanad and matn (Mghari et al., 2022). Sanad is an essential aspect of the hadith because it indicates the chain of the narrators who transmit the hadith (Luthfi et al., 2022). Recognizing the chain of narrator has a crucial role in authorizing the category of particular hadith.

All acceptable hadiths fall into three general categories which are sahih (sound), hasan (good), and da'if (weak). Saḥīḥ (sound) refers to those with a reliable and uninterrupted sanad and a matan (text) that does not contradict the orthodox belief. Hasan (good) refers to those with an incomplete sanad or narrators of questionable authority. Daʿīf (weak) are those whose matan or narrators are subject to serious criticism (Adam Zeidan, 2023). Those previously mentioned processes were conducted manually by hadith scholars by examining a chain of narrators in hadith documents. Hadith scholar identifying narrators in the chain of narrators and searching through the many volumes of biographical dictionaries were laborious because of the ambiguity of the names of the narrators in the sanad of the hadith. Manually finding biographical information on every narrator found in even a set of 10-20 hadith could take many days