Enhanced Topic Modelling using Dictionary For Questions and Answers Problem

Maryamah Maryamah, Agus Zainal Arifin, Riyanarto Sarno, Rizka Wakhidatus Sholikah
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
maryamahfaisol02@gmail.com, agusza@if.its.ac.id, riyanarto@if.its.ac.id, rizka16@mhs.if.its.ac.id

***Abstract*— Making Questions and Answers (QA) with large data and a broad context of problems can cause the desired document to sometimes be irrelevant. QA in terms of religious-social issues have a broad context, so they need to be firstly introduced to the topic. However, the questions raised in the Questions and Answers problem have short text criteria that must be correctly identified by the topic according to the relevant answers. In this paper, we proposed topic modeling for questions and answers with improved term weighting in special words in religious-social problems. The process consisted of preprocessing, making improved dictionary and modeling topic based on dictionary. The result obtained was in the form of topics from input short text which assisted in taking relevant topics, so that correct answers to the questions were obtained.**

Keywords—Topic Modelling, Term Weighting, Dictionary, Religious-social problem

I. INTRODUCTION

Questions and Answers (QA) is a popular communication system nowadays. It is because QA provides convenience for users who would like to obtain information by directly giving answers according to the questions desired by the user. QA is divided into two, namely manual and automatic. The manual QA is done by humans to answer questions; the precision of the answer is assured but there is always possibility regarding wrong answers when too many questions were asked. Repetitive questions lead to inefficiency of manual QA.

Automatic QA can answer questions automatically and efficiently by the system compared to manual QA. However, the system must understand the questions and choose the correct answers. These answers can be easily understood if the context is not too broad; if the context is too broad, it causes system confusion in understanding questions and choosing the correct answer. A system that responds to automatic QA is called Artificial Intelligence (AI); AI is programmed to to answer QA correctly. With the advancement of AI, QA has been rapidly developed. There are several previous studies which discuss how to program automatic QA so that it can provide relevant answers, one of which Visual Question Answering (VQA), given the image and a natural language question about the image to provide an accurate natural language answer [1], recurrent neural model which generates natural language questions from documents, conditioned on answers [2].

Understanding the questions given by users on the QA can be done by detecting the topic of the question and matching the topic of the answer. The application of modeling topics can detect hidden information [3]. The process of detecting topic can be done by using several topic modeling methods, namely Latent Dirichlet Allocation (LDA) [4], Latent Semantic Analysis (LSA) [5], Probabilistic Latent Semantic Analysis (PLSA) [6], Latent Semantic Indexing (LSI) [7], dan Probabilistic Latent Semantic Indexing (PLSI) [8].

However, merely applying the topic modeling method is insufficient to understand the questions in the religion context. It requires technical terms or related terms in a particular field [9]. The technical term used in this study is based on Islamic jurisprudence (fiqh). Islamic jurisprudence is a science that regulates Sharia law obtained based on Islamic law. In addition, non-specific and general questions are prone to producing irrelevant documents. It usually contains special words which generates the problem. In order to understand those particular word, it is necessary to do additional weight if the question contains the word. These particular words depend on the problem to be solved. These words will be built into a dictionary that can be used for certain problems.

In this paper, we proposed build an additional dictionary based on fiqh knowledge on topics modeling method for questions and answers problems. The topic modeling used is LDA and improved based on the words corresponding to the dictionary. The words in the questions contained in the dictionary are given more weight, so that the result is not based on the number of words mentioned but specific words that correspond to the problem. The modeling topics comparation for this paper were LSI and LDA. The result from both methods were compared based on the number of important topics in each document. The topic modeling was tested based on the expert in the field of Islamic jurisprudence.

II. RELATED WORK

A. Question and Answer

Question and Answer (QA) is part of Information Retrieval (IR) and Natural Language Processing (NLP). The QA system can provide answers to questions that are automatically submitted by users. The basic difference between QA and IR is in the input and output produced. In QA, input obtained from users in the form of questions or sentences in natural language is not in the form of keyword. The output is in the form of answers to the questions entered, not in the form of documents or data.

QA can be divided into 2 types based on the problem domain, namely closed-domain and open-domain [10]. Closed-domain QA is a system dedicated to certain topics or domains, for example health, sports, technology, e-libraries, etc. Open-domain QA does not only depend on a topic, but

978-1-7281-2133-8/19/$31.00 ©2019 IEEE

[12] P. Pathak, M. Das Gupta, N. Nayak, and H. Kohli, "AQuPR : Attention based Query Passage Retrieval," pp. 1495–1498, 2018.

[13] K. Bi, Q. Ai, and W. B. Croft, "Revisiting Iterative Relevance Feedback for Document and Passage Retrieval," *arXiv Prepr. arXiv1812.05731*, pp. 1–8, 2018.

[14] K. Bi, Q. Ai, and W. B. Croft, "Iterative Relevance Feedback for Answer Passage Retrieval with Passage-level Semantic Match," *Eur. Conf. Inf. Retrieval. Springer*, pp. 558–572, 2019.

[15] L. K. Sharma and N. Mittal, "Answer Extraction in Question Answering using Structure Features and Dependency Principles," *arXiv Prepr. arXiv1810.03918*, 2018.

[16] S. Khan and K. T. Kubra, "Improving Answer Extraction For Bangali Q / A System Using Anaphora-Cataphora Resolution," *2018 Int. Conf. Innov. Eng. Technol.*, pp. 1–6, 2018.

[17] Y. Wang, A. M. Asce, J. E. Taylor, and M. Asce, "DUET : Data-Driven Approach Based on Latent Dirichlet Allocation Topic Modeling," *J. Comput. Civ. Eng.*, vol. 33, no. 3, 2019.

[18] R. Rahim, N. Kurniasih, M. D. Irawan, and Y. H. Siregar, "Latent Semantic Indexing for Indonesian Text Similarity," *Int. J. Eng. Technol*, vol. 7, no. 2.3, pp. 73–77, 2018.

[19] A. M. Al-zoghby and K. Shaalan, "Ontological Optimization for Latent Semantic Indexing of Arabic Corpus," *Procedia Comput. Sci.*, vol. 142, pp. 206–213, 2018.