International Research Journal of Computer Science (IRJCS)
Issue 08, Volume 4 (August 2017)
ISSN: 2393-9842
www.irjcs.com

Answer Search Indonesian Language Hadith Using Vector Space Model in PDF Document

Bagus Priambodo*
Information System University of Mercu Buana
bagus.priambodo@mercubuana.ac.id

Manuscript History
Number: **IRJCS/RS/Vol.04/Issue08/AUCS10080**
DOI: 10.26562/IRJCS.2017.AUCS10080
Received: 15, July 2017
Final Correction: 24, July 2017
Final Accepted: 28, July 2017
Published: August 2017
Citation: **Bagus, P. (2017). Answer Search Indonesian Language Hadith Using Vector Space Model in PDF Document. IRJCS:: International Research Journal of Computer Science, Volume IV, 01-06.**
DOI: 10.26562/IRJCS.2017.AUCS10080
Editor: Dr.A.Arul L.S, Chief Editor, IRJCS, AM Publications, India

Abstract — Digital text documents are spread in various formats, the most widely used formats today include word format, and PDF format. This research will try to make text search application in text document using vector space approach model. The document format used is a PDF document. Text in PDF will be extracted and then made rank using vector space model. The PDF document consists of ten pages and each page contains a hadith. In general the system can search from the PDF document quite well and able to display the list of results in accordance with the relevance rank with the question.

Keywords— answer retrieval, vector spaces model, text mining

I. INTRODUCTION

Text documents are the most widely generated documents at the blink of an eye. Each organization at least produces dozens of text documents every day. Not to mention the text documents generated by the publication or printing company in the form of books or magazines. Also text documents on the internet. Millions of news sites display millions of text documents every day. Digital text documents are spread in various formats; the most widely used formats today include word format, and PDF format. This text document contains a lot of information. There is at least information in each text document. The process of extracting information from text documents is often referred to as text mining. To search for a particular word in a text document is very easy, just clicking find, and then type a certain word, then the computer will find it for you in particular. But what if we want to find an answer to the question we want. Need an information-processing process or so-called Information Retrieval to do that. Information Retrieval is one of the topics of text mining.

A lot of research has been made to search for answers for text documents. Among them are vector space model (Ismail, 2003), (Hedström, 2005) and recursive data mining [5]. The search for answers with the vector space model method is very good and effective on various documents; in addition, there are also shortcomings [5]. Hadith or Sunnah is the second source of law and life guidance for Muslims. This research will try to make answer search application in Hadist using vector space model approach. The user will try to enter the question, then the system will display five hadith relevant to the question (if any). The most relevant results will be displayed at the top.

II. PROBLEM

Hadith or Sunnah is the second author of law and life guidance for Muslims. Using existing applications the user can search the textbook in the hadith document but cannot search the hadith that best matches the question in question. This study aims to create an application that can search for hadiths that match the questions of the user. This research will prove to make answer search application in Hadist using vector space model approach. The user will attempt to enter the keyword, and then the system will display five traditions relevant to the keyword (if any). The most relevant results will be displayed at the top and then next lower ranking hadith are displayed.

III. TEXT MINING AND VECTOR SPACE MODEL

A. Text mining

Text mining is the operation of extracting hidden information from text documents. Similar to data mining, text mining tasks can be viewed as three steps, process: preprocessing, mining patterns and evaluation results the preprocessing step consists of preprocessing the default text and representing each text in the form Vector. Learning algorithm cannot process a raw document directly. Thus, we must perform a document indexing procedure that maps raw text documents into vector formats. Before documenting indexing, text documents often contain words that can result in lower performance in Model learning, such as misspelled words, abbreviated words, and words with and without originating. This process is referred to as preprocessing text. After we have done preprocessing on the electronic document, we need to define a set of features to represent the text document. A feature refers to an attribute that is a particular characteristic of the data. By viewing text documents, we can think of some characteristics that we can use as a feature to represent text documents. List of all words, list of selected keywords, and list of phrases in text etc. By extracting each feature for each document, we create a feature vector space for the text corpus. Next, we extract the informative patterns during the mining process patterns using various texts-learning algorithms depending on the job. The algorithm analyzes the data, and presents the extracted information. Finally, in the evaluation step, depending on the work performed, data visualization, and the document is taken as extracted information.

B. Information retrieval

The information retrieval (IR) system is utilized to retrieve information relevant to the user's need of information set automatically [1]. One of the common applications of IR systems is the search engine or search engine found on the internet network. As a system, the IR system has various parts that build the system as a whole.

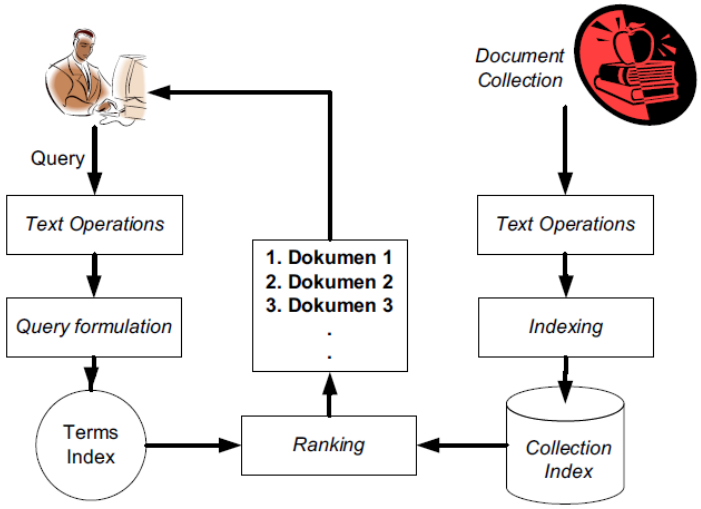

Fig. 1 Illustration retrieval [1]

The first flow starts from the document collection and the second plot starts from the user query. The first flow of processing of the collection of documents into the index database does not depend on the second groove. While the second groove depends on the existence of the index database generated in the first groove.

The parts of the IR system in Figure 2 include:

1) Text Operations (operations on text) which include selection of words in the query or the document (term selection) in transformation document or query into term index (index of words).
2) Query formulation (formulation of the query) that gives weight to index of query words
3) Ranking, searching for documents relevant to the query and sorting the document according to its compliance with the query.
4) Indexing (indexing), builds an index database of document collections. Done first before document search is done.

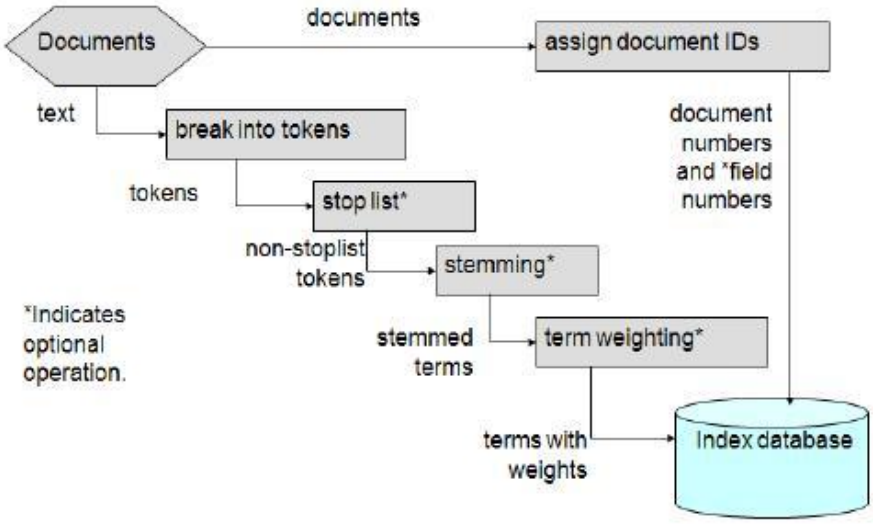

Fig. 2 The process of determining the index [2]

C. Vector Space Model

It is widely used for the purpose of extracting information. It uses the underlying spatial proximity metaphor to determine the semantic proximity. Vector space model proposes a framework in which partial matching between documents or documents and queries are possible. The weights of terms in vectors are used to calculate the degree of similarity between documents. The correlation between vectors can be measured by measuring the cosine of angles between two vectors. During the indexing process, the contents of the document are transferred to the representation vector. Documents and queries in vector space models are represented as feature vectors that represent terms that occur within themselves, namely in the Collection. Each document is represented by a n-dimensional vector, where n is the number of index terms in the collection. The value of each feature indicates the presence or absence of a term given in the given document. Therefore, it is possible to define the relevance of a document to a query (or document to a document) by specifying the number of terms that they have in common.

$$\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{i1} & \cdots & w_{n1} \\ w_{21} & w_{22} & \cdots & w_{i2} & \cdots & w_{n2} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ w_{1j} & w_{2j} & \cdots & w_{ij} & \cdots & w_{nj} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ w_{1m} & w_{2m} & \cdots & w_{im} & \cdots & w_{nm} \end{bmatrix}$$

Fig. 3 Weight value in the document matrix

The elements in the vector may also have a weight attached to it The database can be represented as a document term matrix with the document as Different n line terms t1, t2..., tn, each column represents the assignment of specific terms for documents m d1, d2..., d m collections. For each document keyword combination there can be weight, where the weight may be zero or one of the binary vector or real for the weighted vector, wi j. In the Model vector query space is solved by translating queries into vectors with the same class as the document, comparing them and the similarity rating.

IV. METHODOLOGY

In conducting this research, we conduct various activities which are step in doing research, that is:

A. Data preparation.

Preparing a PDF text document containing ten hadiths, each page contains one hadith, so all of it consists of ten pages.-

B. Propose Algorithm

Extract text from PDF document, after text from pdf extracted, and then stemming will be done in accordance with Indonesian language. Furthermore, giving the index and weight. The user will enter the question and the result of the search will be displayed to the user. All sequence can be seen in figure 4 below.

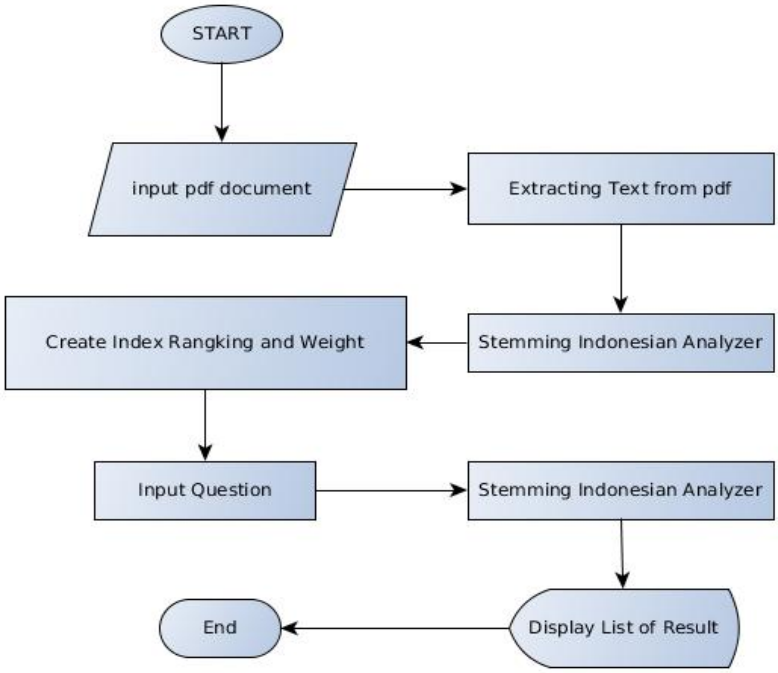

Fig. 4 Flow chat of Application

V. EXPERIMENT AND DISCUSSION

A. Prepare question

Pada eksperimen ini kami menyiapkan tiga buah pertanyaan yaitu : "puasa" (fasting), "niat" (intention), and "kapan puasa ?" (When fasting date?).

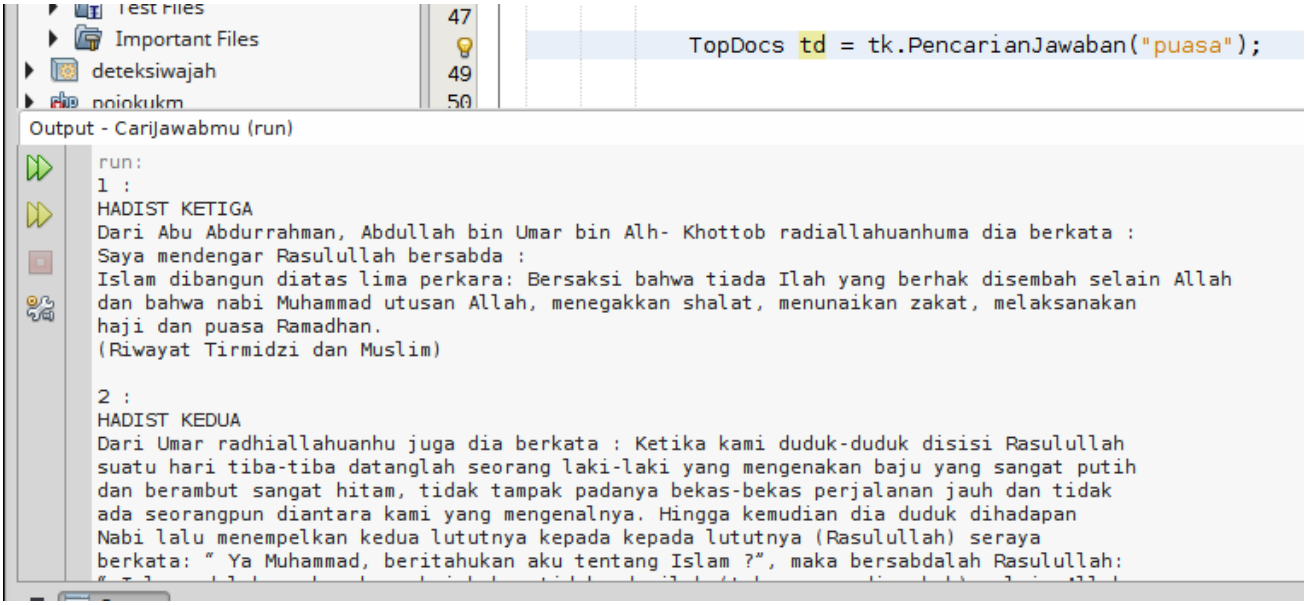

Fig. 5 Result of fasting question

B. Result

Result of experiment is shown in figure 5, figure 6, and figure 7 below.
The first test with the keyword of fasting featured two hadiths: the third hadith in the first and the second hadith in the second.

The third test with the intention question, shows only one hadith that is the first hadith

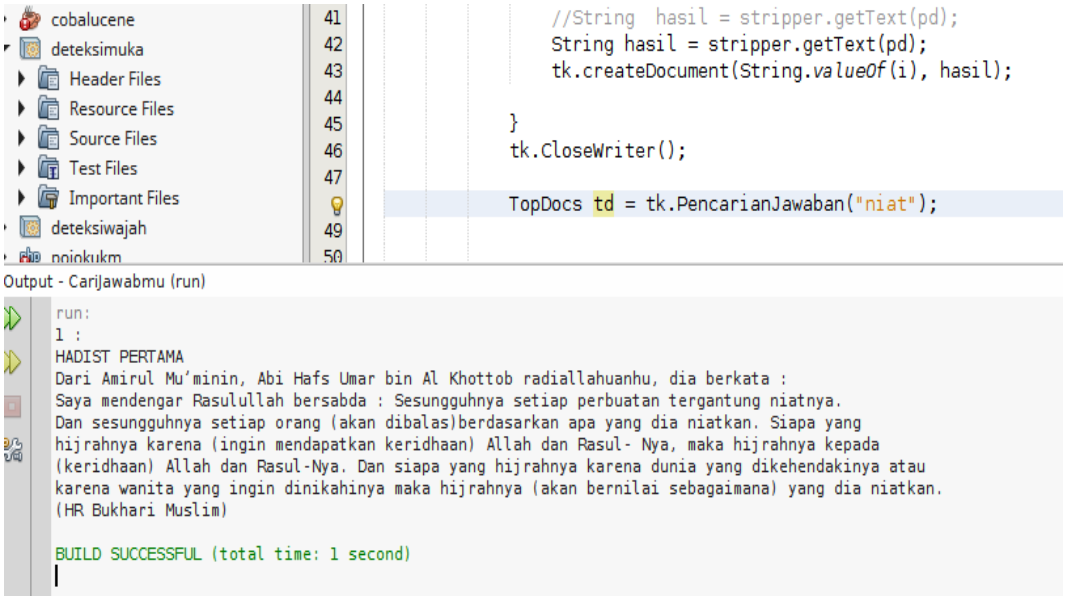

Fig. 6 Result of intention question.

The third test with the question of when fasting presents two hadiths namely the third hadith in the first and the third hadith is second.

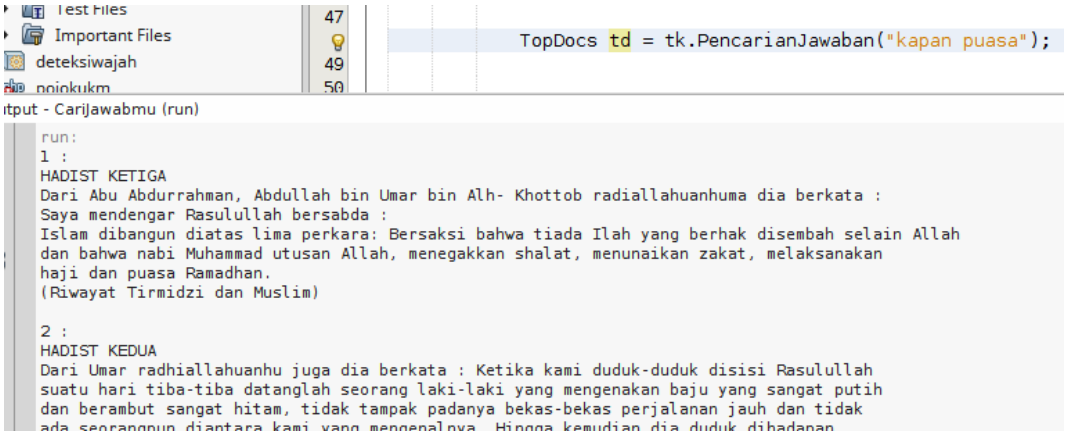

Fig. 7 Result of intention question.

VI. CONCLUSIONS

Experimental results show by weighting using the vector space model approach, the system can provide answers to the hadith most closely related to the question of the user. The most related hadiths will be displayed in the top sequence. However, the system cannot provide answers in the form of sentences. But it only provides answers to the list of hadiths that are most relevant to the question. Our next research will make the system able to give answers in the form of sentences. In addition to the list of hadiths that are sorted with the most relevant of the questions

REFERENCES

1. Bunyamin, Hendra. Algoritma Umum Pencarian Informasi Dalam Sistem Temu Kembali Informasi Berbasis Metode Vektorisasi Kata Dan Dokumen.Jurnal Informatika Ukm, No. 2, Vol. I, Hal 85-87. 2005.
2. Dwi, Ratih Puspita. Analisis Usability Pada Search Engine Dengan Dokumen Teks Terstruktur. IttTelkom, 2009.
3. Hedström, Anna "Question Categorization For A Question Answering System Using A Vector Space Model", Department Of Linguistics And Philology Språkteknologi Programmet, Uppsala Universitet, 2005.
4. Israml, Ismail, Yukawa, Takashi., "Question Answering System Using Concept-Based Vector Space Model", International Journal Of Computer Vision, Kluwer Academic, Netherlands, 2004
5. Honlor, Apirak "Sequential Patterns and Temporal Patterns for Text Mining", Rensselaer Polytechnic Institute Troy, New York. 2011