Atlantis Highlights in Engineering (AHE), volume 1
International Conference on Science and Technology (ICST 2018)

Replies Identification in Question Answering System using Vector Space Model

Intan Yuniar Purbasari[1], Fetty Tri Anggraeny, Masti Fatchiyah Maharani
Department of Informatics
Universitas Pembangunan Nasional "Veteran" JawaTimur
Surabaya, Indonesia
[1]intanyuniar.if@upnjatim.ac.id

***Abstract*—Automatic question answering system is an information retrieval system designed to return an answer to a specific question. It is a base system for a chatbot application. This research aims to identify responses in a question answering system using Vector Space Model (VSM). The system compared query questions to possible answers stored in the repository, and the top ten highest similarity results are returned. Dataset used is from students' questionnaires about final project essay and field practice, which are specific topics of interest often asked to program coordinator. There are 100 questions and answers, and the result gives 0.413average precision, 0.893 average recall, and 0.563 average F-Measure.**

Keywords—question answering system; vector space model; information retrieval; data mining

I. INTRODUCTION

Humans and information are two things that cannot be separated. Information is used by humans to develop new ideas, explore a particular science, make decisions, interact, communicate, and answer questions. The human's need for information thus raises the demand for information. In this era, the desire for information can be fulfilled very easily and quickly by the presence of various kinds of information technology such as newspapers, TV, radio, and the Internet.

The Internet, which is a crossing gate for information without taking into account geographical location and time, provides a wealth of unlimited details that can be searched using search engines like google. The abundance of information generated by the search engine results in the potential for users to obtain information that is not relevant. Too many search results also require users to spend more time to sort out the right information according to user's intention.

Chatbots also experience the same problem. As an alternative to search engines, recently chatbots can also be categorized as an information technology that helps users to get information. In [1]the Indonesian Contact Center Association explains that Chatbot is a computer program designed to communicate and interact with users. There are two types of chatbot with regards to the technology used: chatbot with a system of rules and chatbot with the implementation of artificial intelligence. Chatbots with artificial intelligence allows the system to be able to communicate with users like humans. Chatbot works by giving responses to questions entered by users. This response is generated by scanning the entered keyword and then looking for the most suitable counterpart or the most similar pattern of words to what has been declared in the system database. To help computers understand, interpret, and manipulate human language so that they can get specific information, an approach called Natural Language Processing (NLP) is used. Currently, chatbots have been widely used to help simplify human work such as command control, customer care, messaging applications, and virtual assistance.

Text Mining, or some sources use the term Text Analysis, is the process of mining data in the form of text with data sources usually from documents, which are unstructured. The aim is to find words that represent the documents and to analyze the connection between them. Text Mining is widely used to solve problems that require information retrieval or commonly known as Information Retrieval (IR). IR is different with database searching in that in IR the returned results are ranked [2]. Vector Space Model (VSM) is a mathematical model used to measure the similarity between a document and a query that can be used on IR to determine that the text is relevant to information[3].

A previous study [4] has developed an answering system applied in an e-learning context in an open university, where most, if not all, interactions are by online. The resources were taken from course learning materials, messages on discussion boards, and other sources from the Internet. The system applied a morphological analysis of the text's message body, generated tag clouds relevant to the context of each message, and performed useful context search to given resources. They rated the result with a Mean Score of the Usefulness (MSU) from experts (from the scale of 0-3, 1.77 for system's finding information and 1.47 for giving a fast answer) and novices (1.51 for system's finding information and 1.75 for providing a quickresponse).

Meanwhile, an auto-answering system suffers from low hit rate of expected answers. One suggestion to improve the mining performance of text mining process is if there were not a high enough amount of corpus quantity available, repeating to input the same corpus a few times is advisable. Also, the co-occurrence ranges among words should be kept to a single document [5]. More recent research applied an ontology-based Bayesian network to locate desired treatment departments in an

interactive question answering system in a hospital [6]. The ontology integrated close temporal associations between words in the input query and achieved 93% correction rate for the first prediction of treatment department.

A study conducted in [7] have developed a question answering system using Vector Space Model to the Ministry of Education and Culture and the Ministry of Tourism and Creative Economy Culture of Indonesia. Another research [8] also made use of VSM applied in unstructured data, where keywords were generated using TF-IDF score and used to index every file and query.

A comparison study [9] analyzed the use of VSM, Latent Semantic Indexing, and Formal Concept Analysis to perform Information Retrieval task using standard Medline and real-life healthcare datasets. The result showed that all three are similar regarding the ability to return relevant documents, in which the top 10 documents retrieved by all methods were related to the questions.

In [10] the author said that the use of Vector Space Models in the retrieval system was able to provide more evident rating values in information retrieval, partial matching of keywords, and also produce reference results that matched the needs.

The topic domain of Q&As in this research is the final project essay (undergraduate thesis) and field practice. Students tend to ask similar questions regarding prerequisite courses, required documents, procedures, deadlines, schedules, and other inquiries to the program coordinator. This research will apply VSM to represent questions and answers stored in the repository and also questions asked by users. The questions asked will then be compared to that of available resource and the system will return the most probable answers matching the questions.

II. METHODOLOGY

A. Dataset

The dataset for this experiment was in the form of typical questions compiled from questionnaires result of 42 students and answers from the program coordinator within the specified topic domain. There is 100 pairs of questions and answers in the Indonesian language collected and stored in MS-Word (.docx) format.

B. Design

The prototype design methodology of this information retrieval system is in Fig. 1. It starts with (1) document collection as a dataset; (2) preprocessing stage, which includes text mining preprocessing steps: tokenization, filtering, and stemming; (3) indexing; (4) representing each document in Vector Space Model; (5) query processing, which includes restructuring process; (6) calculating "distance" between the query and the stored documents using cosine similarity; and (7) returning the document/answer with the highest cosine value.

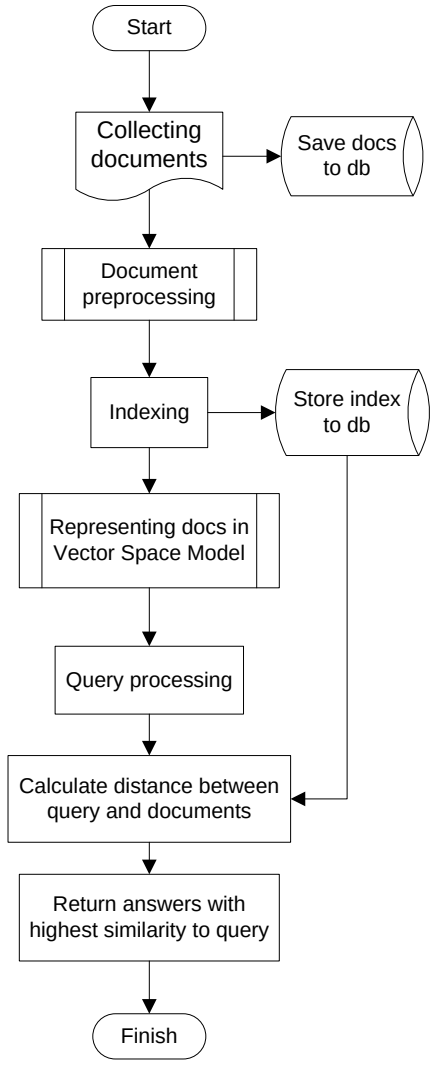

Fig. 1. System Methodology

1. Preprocessing
   a. Tokenization: a process to separate sentences into words by deleting punctuation and changing to all lower cases.
   b. Filtering: The extraction of essentialwords from the tokenization process with stoplist removal. Some stop words examplesare*yaitu*, *adalah*, *bila*, *cukup*, *dan*, *itu*, *iya*, stored in a one-dimensional array.
   c. Stemming: The identification of base words by removing all affixes (prefix, infix, suffix, and confix) according to morphological structures of the Indonesian language using Nazief-Adriani algorithm [11].The results are words referred to as terms.
2. Vector Space Model (VSM) representation

   In VSM, we applied the Term Frequency-Inverse Document Frequency(TF-IDF) method to weigh each document containing terms in the topic domain. This research used binary term frequency, in which if a word exists in a document, a value of 1 will be assigned, else a value of 0 will be given. An example is as follows. Assume

there are three documents, D_1, D_2, and D_3, in the database, and a query, Q, where:

D_1 : *Bagaimana format bukulaporan PKL?* (What is the format of a field practice report?)
D_2 : *Kapan jadwalsidang PKL dilaksanakan?* (When is the schedule of a field practice seminar?)
D_3 : *Bagaimana format berkas seminar proposal?* (What is the format of a final project essay proposal for a seminar?)
Q : *Bagaimanasusunanlaporan PKL yang benar?* (What is the correct arrangement of a field practice report?)

Each question document has an answer pair, so if a query matches with a document, the system will return the answer from the matching document.

Table 1 lists the TF-IDF values for all documents and the query.

TABLE I. TF-IDF VALUES FOR DOCUMENTS AND QUERY

Token	TF-IDF			
	Q	***D1***	***D2***	***D3***
Bagaimana (how)	0.176	0.176	0	0.176
Kapan (when)	0	0	0.477	0
Format	0	0.176	0	0.176
Lapor (report)	0.477	0.477	0	0
PKL (abbrv. Field practice)	0.176	0.176	0.176	0
Jadwal (schedule)	0	0	0.477	0
Sidang (seminar)	0	0	0.477	0
Laksana (happen)	0	0	0.477	0
Seminar	0	0	0	0.477
Proposal	0	0	0	0.477
Berkas (file)	0	0	0	0.477
Susun (arrange)	0	0	0	0

3. Distance calculation

The distance between the query and each document is computed using Cosine Similarity function (1):

$$\text{Cosine}\theta_{\text{Di}} = \frac{Q \,.\, D}{|Q|^{*}|D|} = \frac{\sum_{i=1}^{n} W_{qj} \,.\, W_{ij}}{\sqrt{\sum_{i=1}^{n} W_{qj}^{2} \,.\, W_{ij}^{2}}} \quad (1)$$

Where:

Q = Query
D = Document
|Q| = Query Vector
|D| = Document Vector
W_{qj} = Weight of term in query vector *j*
W_{ij} = Weight of term in document vector *j*
Using values from Table I, the cosine similarity values are:

$$\text{Cosine}_{\text{Q.D1}} = \frac{(0.176*0.176)+(0*176)+(0.477*0.477)+(0.176*0.176)}{0.535*0.563}$$
$$= 0.953$$

$$\text{Cosine}_{\text{Q.D2}} = \frac{(0.176*0)+(0*477)+(0.176*0)+(0.176*0.176)+(0*0.477)+(0*0.477)+(0*0.477)}{0.535*0.968}$$
$$= 0.058$$

$$\text{Cosine}_{\text{Q.D3}} = \frac{(0.176*0.176)+(0*176)+(0.477*0)+(0.176*0)+(0*0.477)+(0*0.477)+(0*0.477)}{0.535*0.860}$$
$$= 0.65$$

Thus, the most relevant document is Document 1 and the answer pair from Document 1 is returned, which is: *Laporan PKL terdiri dari Sampul, Lembar Pengesahan, Surat Keterangan Selesai PKL, Abstrak, Kata Pengantar, Ucapan Terima Kasih, Daftar Isi, Daftar Gambar, Daftar Tabel, BAB I Pendahuluan, BAB II Gambaran umum tempat PKL, BAB III Pelaksanaan dan Pembahasan, BAB IV Penutup, Daftar Pustaka, dan Lampiran* (Field report consists of a cover, validation sheet, letter of completion of field practice activity, abstract, preface, gratitude notes, table of contents, table of figures, table of tables, Chapter I Introduction, Chapter II General Description of Field Practice Company, Chapter III Implementation and Discussion, Chapter IV Closing Statements, References, and Appendices).

III. RESULTS AND DISCUSSION

We experimented on 3 (three) testing scenarios. Scenario I had questionshaving the same meaning as the questions in the database but in different writing.Scenario II consists of items having the samecopy as the questions in the database. In all two scenarios, both stored questions and answers were stemmed. Also, each has five questions asked. In scenario III, only the questions were stemmed and stored in database, and it has all ten questions asked from scenario I and II.Scenario I and IIare based on data stored on one database, while scenario IIIis based on data in another database. The total number of questions were 20 questions.

Table II lists questions, answer candidates, and relevancy status between questions and selected answers from all four scenarios (in each scenario, 2questionswere taken as examples). The system generated the top 10 answer candidates based on the similarity values, and the bold-style answer is selected to be the one with the highest similarity.

TABLE II. QUESTION, ANSWER CANDIDATES, AND THEIR RELEVANCE

No.	Scenario	Query	Candidate Answer Documents and similarity values		System output	Relevance
1.	I	*PKL dilaksanakan maksimal berapa lama?* (How long the PKL/field practice takes place?)	**D3=0.894** D17= 0.774 D41= 0.774 D49= 0.774 D30= 0.632	D40=0.632 D10= 0.632 D6=0.632 D1=0.632 D7=0.632	**[D3]** *Lama pengerjaan PKL adalah 1-3 bulan di instansi* (The length of PKL is 1-3 months in the company)	YES
2.	I	*Apakah akan ada seminar proposal pada bulan berikutnya?* (Will there be a proposal seminar the following month?)	**D13= 1** D58 = 1 D54 = 0.816 D61 = 0.816 D3 = 0.577	D49 = 0.577 D90 = 0.577 D44 = 0.577 D15 = 0.577 D83 = 0.577	**[D13]** *Seminar proposal dilaksanakan setiap 2 bulan 1 kali* (Proposal seminar is held once every 2 months)	YES
3.	II	*Berapa lama PKL dilaksanakan?* (How long does the PKL/field practice take place?)	**D3=1** D49= 0.866 D30= 0.707 D40= 0.707 D17= 0.707	D10=0.707 D7= 0.707 D1=0.707 D6=0.707 D26=0.707	**[D3]** *Lama pengerjaan PKL adalah 1-3 bulan di instansi* (The length of PKL is 1-3 months in the company)	YES
4.	II	*Kapan jadwal seminar proposal?* (When is the next proposal seminar?)	**D13= 1** D10=0.707 D14 =0.707 D91 =0.707 D54 =0.707	D58 =0.707 D61 =0.707 D64 =0.707 D97 =0.707 D90 =0.500	**[D13]** *Seminar proposal dilaksanakan setiap 2 bulan 1 kali* (Proposal seminar is held once every 2 months)	YES
5.	III	*Apa syarat mengajukan skripsi?* (What are the requirements to submit a thesis?)	**D16=1** D89= 0.866 D59=0.866 D85= 0.707 D52=0.866	D44=0.707 D70=0.707 D2=0.707 D15=0.707 D34=0.707	**[D16]** *Membawa draft skripsi ke calon dosen pembimbing dan melengkapi form berkas skripsi* (Bring the thesis draft to a potential supervisor(s) and complete thesis submission forms)	YES
6.	III	*Berapa lama batas maksimal pengerjaan skripsi?* (How long is the maximum limit of a thesis work?)	**D17=1** D48 = 0.912 D97= 0.707 D57= 0.707 D96= 0.577	D30= 0.577 D39= 0.577 D3=0.577 D45= 0.577 D95=0.577	**[D17]** *Satu tahun* (one year)	YES

From Table 2, in Scenario I, both samples gave relevant answers to the questions asked. From all fivequestions of scenario I, the system was able to provide all applicable answers.In Scenario II, since the queries' wording are the same as that of stored in the database, the returned answers were all relevant. In scenario III, the system gave all relevant answers (as long as the topic is within the domain) although only the key questions' stem stored in the database.

System's evaluation is measured using precision, recall, and F-measure values for all 20 tests data which are calculated using formula (2), (3), and (4):

$$\text{Precision} = \frac{\text{Number of relevant docs retrieved}}{\text{Number of docs retrieved}} \quad (2)$$

$$\text{Recall} = \frac{\text{Number of relevant docs retrieved}}{\text{Number of relevant docs in collection}} \quad (3)$$

$$\text{F. Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

To make ashort-enough list to calculate the values for Precision, Recall, and F-Measure, we limit the number of retrieved documents of each query to five results. Thus, to assess the relevance of the retrieved documents, we will use pooling approach, where relevanceis evaluated from the top *k* documents returned from a document collection [12] in binary value: *relevant* or *non-relevant*. In this experiment, the value of *k* is 5.

Table III provides all questions result regarding the retrieved documents to the returned answers of five documents each. There was a total of 100 returned documents for all three scenarios. The table also lists all relevant documents for each query so that the recall values can be calculated.

In scenario I, a total of 25 documents were retrieved having ten relevant documents and 15 non-relevant documents. In scenario II, 25 documents were retrieved with ten relevant documents and 15 non-relevant documents. While in scenario III, 50documentswere returned with 22documents are relevant, and 28 documents are not relevant. One important thing to note is that, for all queries in all three scenarios, the first ranked returned documents were all relevant to each query.

Table IV lists the evaluation of precision, recall, and F-measure for all 20 tests data query calculated using formula (2), (3), and (4).

Table IV shows that, for all three scenarios, the average precision value is 0.413, the average recall value is 0.893, and the F-Measure value is 0.563. From this result, we can say that the precision of the system is still below expectation. However, in Questions and Answers application domain, we all agree that the system will only be required to give the highest matching answer, not a set of possible ranked answers. In this case, the result of this experiment has already satisfied that requirement,

since the first ranked returned documents were all relevant to each query.

TABLE III. SUMMARY OF RETRIEVED DOCUMENTS FOR ALL QUERIES

Question	Scenario	Retrieved Docs	Assessment of Relevancy	Total Relevant Docs
Q1	I	D6, D54, D29, D34, D97	*Relevant: 1 Non-relevant: 4*	2
Q2	I	D3, D17, D41, D49, D30	*Relevant: 2 Non-relevant: 3*	2
Q3	I	D13, D58, D54, D61, D3	*Relevant: 1 Non-relevant: 4*	1
Q4	I	D16, D90, D15, D60, D96	*Relevant: 2 Non-relevant: 3*	2
Q5	I	D1, D7, D8, D11, D4	*Relevant: 4 Non-relevant: 1*	5
Q6	II	D6, D54, D8, D94, D26	*Relevant:2 Non-relevant:3*	2
Q7	II	D3, D49, D30, D40, D17	*Relevant: 2 Non-relevant: 3*	2
Q8	II	D13, D10, D14, D91, D54	*Relevant: 2 Non-relevant: 3*	2
Q9	II	D17, D49, D3, D58, D41	*Relevant: 2 Non-relevant: 3*	2
Q10	II	D91, D99, D10, D14, D92	*Relevant: 2 Non-relevant: 3*	2
Q11	III	D6, D26, D96, D8, D5	*Relevant:2 Non-relevant:3*	2
Q12	III	D3, D17, D48, D39, D30	*Relevant: 1 Non-relevant: 4*	1
Q13	III	D13, D53, D43, D64, D63	*Relevant: 3 Non-relevant: 2*	3
Q14	III	D16, D89, D59, D85, D52	*Relevant: 2 Non-relevant: 3*	4
Q15	III	D8, D1, D11, D7, D6	*Relevant: 3 Non-relevant: 2*	3
Q16	III	D6, D26, D8, D53, D18	*Relevant:3 Non-relevant:2*	3
Q17	III	D3, D39, D30, D17, D10	*Relevant: 2 Non-relevant: 3*	2
Q18	III	D13, D53, D63, D10, D90	*Relevant: 2 Non-relevant: 3*	3
Q19	III	D17, D48, D97, D57, D96	*Relevant: 2 Non-relevant: 3*	2
Q20	III	D18, D93, D8, D6, D97	*Relevant: 2 Non-relevant: 3*	3

TABLE IV. SYSTEM'S EVALUATION

Scenario	Precision	Recall	F-Measures
I	10/25 = 0.4	10/12 = 0.833	(2*0.4*0.833)/ (0.4+0.833) = 0.540
II	10/25 = 0.4	10/10 = 1	(2*0.4*1)/(0.4+1) = 0.571
III	22/50 = 0.44	22/26 = 0.846	(2*0.44*0.846)/ (0.44+0.846) = 0.579
Average	**0.413**	**0.893**	**0.563**

Still, from Table IV, we can conclude that the recall hit rate is satisfying (above 85%) which means that almost all relevant documents were returned in the top five results. Scenario II achieved a perfect recall value which returned 100% relevant documents. The slightly low precision value was balanced with the high recall value, which made the harmonic average F-Measure value of the system is 0.563.

Overall, the developed system has performed well on returning the relevant answer given a query. The limited topic domain constraint is assumed to contribute to the quite high F-Measure value. The system is still having difficulties in recognizing and differentiating some words. For example, in the question, *Berapa lama batas maksimal pengerjaan skripsi* (How long is the maximum time limit for the thesis work?). The system is unable to differentiate *skripsi* (thesis) and *PKL* (practical fieldwork). Thus, in the top five returned documents, it not only gave answers regarding the time limit of thesis work but also regarding the time limit of practical fieldwork, which is regarded as a non-relevant answer.

IV. CONCLUSION

Based on the experiment that has been completed, it can be concluded that the use of Vector Space Model for document representation can be applied in a Question and Answer system. The average precision score is 0.413, average recall score is 0.893, and the F-Measure score is 0.563. The number of queries and documents should be increased and varied more if we want to have a more reliable result, within a broader topic domain. Also, using another approach such as the Latent Semantic Analysis [13] might be considered to improve the recognition of more related words to the query. Some questions need to be repeated with different writings, so that system truly tests the reliability in recognizing the question. A newer approach on Vector Space Model, known as Generalized Vector Space Model, might also be used instead to include the computation of correlation between terms [14].

ACKNOWLEDGMENT

The writers would like to thank the Faculty of Computer Science Universitas Pembangunan Nasional "Veteran" Jawa Timur for its support for this work to be published.

REFERENCES

[1] MZ, "What is Chatbot?," 2017. [Online]. Available: http://icca.co.id/apa-itu-chatbot/. [Accessed: 20-Apr-2017].
[2] B. J. Jansen and S. Y. Rieh, "The Seventeen Theoretical Constructs of Information Searching and Information Retrieval," J. Am. Soc. Inf. Sci. Technol., vol. 61, no. 8, pp. 1517–1534, Aug. 2010.
[3] G. Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," Commun. ACM, vol. 18, no. 11, pp. 613–620, Nov. 1975.
[4] J. Moré, S. Climent, and M. Coll-Florit, "An Answering System for Questions Asked by Students in an e-Learning Context," Int. J. Educ. Technol. High. Educ., vol. 9, no. 2, pp. 229–239, Jul. 2012.
[5] K. Aman and F. Ishino, "Application of Text Mining into Auto Answering System and Improvement of Mining Performance," in Project E-Society: Building Bricks, 2006, pp. 166–175.
[6] J.-F. Yeh, Y.-J. Huang, and K.-P. Huang, "Ontology-based Bayesian network for clinical specialty supporting in interactive question answering systems," Eng. Comput., vol. 34, no. 7, pp. 2435–2447, 2017.
[7] L. Jovita, A. Hartawan, and D. Suhartono, "Using Vector Space Model in Question Answering System," Procedia Comput. Sci., vol. 59, pp. 305–311, 2015.
[8] R. Jayashree and N. Niveditha, "Natural Language Processing Based Question Answering Using Vector Space Model," in Proceedings of Sixth International Conference on Soft Computing for Problem Solving, 2017, pp. 368–375.
[9] C. A. Kumar, M. Radvansky, and J. Annapurna, "Analysis of a Vector

Space Model, Latent Semantic Indexing and Formal Concept Analysis for Information Retrieval," Cybern. Inf. Technol., vol. 12, no. 1, 2012.
[10] I. Irmawati, "Information Retrieval in Documents using Vector Space Model," J. Ilm. FIFO, vol. 9, no. 1, pp. 74–80, 2017.
[11] M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. E. Williams, "Stemming Indonesian: A Confix-stripping Approach," ACM Trans. Asian Lang. Inf. Process., vol. 6, no. 4, pp. 1–33, Dec. 2007.
[12] C. D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval. Cambridge, UK: Cambridge University Press, 2008.
[13] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," J. Am. Soc. Inf. Sci., vol. 41, no. 6, pp. 391–407, 1990.
[14] S. K. M. Wong, W. Ziarko, and P. C. N. Wong, "Generalized Vector Spaces Model in Information Retrieval," in Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 18–25, 1985.