A Questions Answering System on Hadith Knowledge Graph

Kemas Rahmat Saleh Wiharja, Danang Triantoro Murdiansyah, Muhammad Zakiyullah Romdlony, Tiwa Ramdhani & Muhammad Ramadhan Gandidi

School of Computing, Telkom University, Jalan Telekomunikasi No. 1, Terusan Buah Batu-Bojongsoang, Kabupaten Bandung 40257, Indonesia
*E-mail: bagindokemas@telkomuniversity.ac.id

Abstract. Several works have presented the Hadith on different digital platforms, ranging from websites to mobile apps. These works were successful in presenting the text of the Hadith to users, but this does not help them to answer any particular questions about religious matters. Therefore, in this work we propose a question-answering system that was built on a Hadith knowledge graph. To interpret the user questions correctly, we used the Levenshtein distance function, and for storing the Hadith in graph format we used Neo4J as the graph database. Our main findings were: (i) a knowledge graph is suitable for representing the Hadith and also for doing the reasoning task, and (ii) our proposed approach achieved 95% for top-1 accuracy.

Keywords: *Hadith; knowledge graph; semantic; question-answer system; reasoning.*

1 Introduction

After Al-Quran, the Hadith – a collection of traditions containing sayings of the prophet Muhammad with accounts of his daily practice (the Sunnah)[1] – is the most important source for Islamic teachings. With the advent of information technology, many studies have tried to present a digital version of the Hadith on different digital platforms, such as websites (http://sahih-bukhari.com/) or mobile apps (Hadith Collection of Greentech). A digital version of the Hadith helps people in searching for specific Hadiths. Unfortunately, when people search for answers to questions about religious matters, the current digital versions require the user to read and understand the meaning of multiple Hadiths, and then draw their own conclusions. Not everyone is willing to sacrifice their time to read a number of Hadiths and then draw conclusions about their contents. Many people also have insufficient knowledge to understand the text and meaning of the Hadith.

The major problems this research addressed were the following:

1. *Time-consuming process of existing Islamic consultation platforms.* When people ask questions about Islamic matters, consultation with an expert is

Received April 1st, 2022, Revised July 22nd, 2022, Accepted for publication September 1st, 2022.
 ISSN: 2337-5787, DOI: 10.5614/itbj.ict.res.appl.2022.16.2.6

typically the best choice [2]. This can be done synchronously or asynchronously. However, this process is time-consuming. By providing a knowledge graph-based question-answers system, people can obtain instantaneous answers.

2. *Lack of Hadith experts to approach for consultation.* In Indonesia, and generally in other Muslim countries, the number of Hadith experts is smaller compared to the number of experts on Al-Quran [3]. Meanwhile, both Al-Quran and Hadith are considered to be primary sources for Islamic teachings.
3. *The absence of certification for Hadith experts.* For people who memorize Al-Quran there exists a type of certification, known as Hafidz. However, in the Hadith field there is no such certification. This leads to unclarity about the eligibility of Hadith experts.
4. *Knowing the context of the Hadith is necessary.* When people search and read the Hadith without knowing the context, it can lead to wrong answers. That is why all Hadith scholars systematically interpret the book of Hadith according to the context. The use of digital platforms for the Hadith without concern about the context can lead to misinterpretation.
5. *Lack of knowledge of the Arabic language leads to misinterpretation.* Originally, the Hadith was delivered in the Arabic language. However, most of the existing digital platforms for the Hadith are in non-Arabic languages (Ajami). Since the translation process is prone to mistakes, different and contradictory translations are unavoidably found in many Hadith translations. Therefore, the involvement of experts who master the Arabic language is important for the verification and validation of a translation.

Observing these problems, we realized that we needed to create a question-answer system (QA system) to help users save time and effort in getting answers to questions about the Hadith. According to Lan, *et al.* [4], there are two main approaches to QA systems, namely semantic parsing based methods (SP-based methods), and information retrieval based methods (IR-based methods). SP-based methods answer user questions by constructing the logical form of the question and then executing the logical form against a knowledge base to get the answer. IR-based methods directly retrieve and rank answers from a knowledge base by extracting a question-specific graph from the knowledge base. In both approaches, the role of the knowledge base is very important. Therefore, we decided to use a knowledge base as the repository for our QA system. As for the type of knowledge base, we chose a knowledge graph, because this is widely used by many large companies, such as Google Knowledge Graph (Singhal [5]) and Airbnb (Chang [6]), Amazon (Dong [7]), Facebook (Noy, *et al.* [8]), and LinkedIn (He, *et al.* [9]).

At the time of writing this paper, we had not found any previous paper that proposed a QA system on a Hadith knowledge graph. Therefore, the following are the contributions of this study:

1. It is the first work that developed a QA system using a Hadith knowledge graph.
2. Our Hadith knowledge graph is not a static knowledge graph (KG), since we accommodate feedback from a Hadith expert to be inserted into the KG.

The rest of this paper is organized as follows. In Chapter 1 we introduce the main motivation of our work. In Chapter 2 we review the existing literature on the topic of this study. In Chapter 3 we explain our proposed approach in detail. In Chapter 4 we present our approach's performance and its evaluation. Finally, in Chapter 5 we conclude our paper and mention several subjects for future works.

2 Related Work

2.1 Existing Question Answer System

This section presents an overview of existing question-answering systems related to the Hadith.

Naemah & Saad [10] developed a question-answering system for the Hadith domain by using the support vector machine method. They attempted to enhance the accuracy of their question-answering system for the Hadith by using several methods, i.e., N-gram, WordNet, CS (cosine similarity, LCS (longest common subsequence), SVM (support vector machine), and NER (name entity recognition). N-gram, WordNet, CS, and LCS were used to update and enrich the extracted concepts of user queries based on the formal representation of Hadith answers or documents. SVM and NER were used to classify documents based on relevant subjects and question types in order to reduce the searching scope of the answer documents. The evaluation showed that the best average accuracy achieved by the system was 80%. SVM successfully enhanced the system's accuracy up to 10% compared to using other methods without classification process.

Assegaf, *et al.* [11] developed a question-answering system for the Hadith that can answer factual Islamic questions. Their question-answering system is able to receive input posed in natural language, such as 'What are the best deeds we can do to please Allah?' The question-answering system uses keywords-based search to find matching verses. The system was evaluated by using top-1 and top-5 accuracy. The system achieved 30% for top-1 accuracy, and 52% for top-5 accuracy.

Abdi, *et al*. [12] developed a question-answering system for the Hadith by using linguistic knowledge. According to the authors, all question-answering systems have a common problem, i.e., it is difficult to capture the meaning when comparing a sentence and a user query. The proposed method successfully tackled this problem using several steps. Also, in order to reduce redundant Hadith texts, the proposed method used the greedy algorithm to impose appropriate penalties on sentences. The experimental results showed that the proposed method was able to improve the performance compared with the previous methods on Hadith datasets.

In this work, we built a question-answer system by using a different method than the aforementioned approach. From [13], we know that the knowledge graph-based question-answer system (KGQA) has many benefits compared with other types of question-answer systems. One benefit is that other types of question-answer use intent detection, which requires a lot of data, whereas KGQA relies on a parsing process that breaks down the question into its elementary components. Another benefit of KGQA is its scalability. Since the data format only consists of nodes and edges, KGQA is naturally easy to scale up or down.

It is important to maximize the accuracy of the QA system. One possible way to increase the accuracy of such a system related to Islamic knowledge is to involve a Hadith expert. To relate a query with the comments from an expert, a solution is needed that is naturally designed for traversing relationships, for example, a knowledge graph.

Thus, we built a question-answer system with a Hadith knowledge graph as its database, and we improved its accuracy by considering expert comments. The expert comments were included in the Hadith knowledge graph as nodes.

2.2 Knowledge Graph

In Indonesian, the term ‘knowledge base’ can be interpreted as a database that contains knowledge. This definition implies that a knowledge base is a passive repository. Adding knowledge to this repository can be done by taking new knowledge from outside the knowledge base and then inserting the new knowledge into the existing knowledge or by creating a reasoning system that is able to add new knowledge based on two or more existing knowledges. For example, the first knowledge is: ‘North Sumatra is part of Indonesia’, while the second knowledge is: ‘Medan is the capital of North Sumatra’. Then the reasoning system can conclude a third knowledge, namely ‘Medan is part of Indonesia’. This combination of knowledge base and reasoning system, as proposed by Ehrlinger & Wöß in [14], is called a knowledge graph.

The word 'graph' in knowledge graph must not be interpreted as a graphic, because here it means the relationship or relationships between an element of knowledge (commonly called an entity) and other elements of knowledge. Take for example the fact that Medan is the capital of North Sumatra. There are two elements of knowledge, or entities, namely the Medan entity and the North Sumatra entity. These two entities are connected by a relationship, namely the 'capital of' relationship. The most natural formats for describing entities and relationships are nodes and arcs. Both are known in the mathematical world as graph formats. Figure 1 depicts the knowledge about Medan as the capital of North Sumatra in graph form.

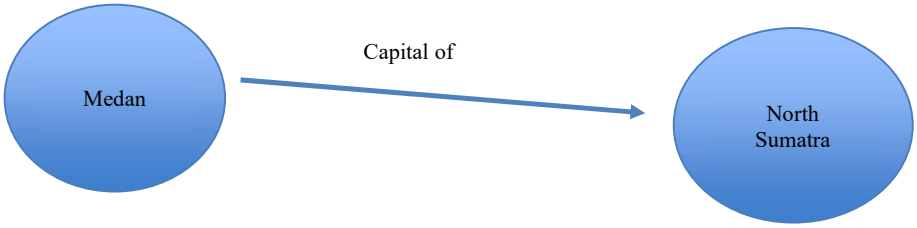

Figure 1 Example of a simple knowledge graph.

3 Our Approach

In this section, we describe our proposed approach. First, we introduce the common flow of our QA system. Then, we discuss in detail the components of our proposed approach.

3.1 The Common Flow

Given a question that is given in a natural language text, our QA system will convert the text into a cypher query syntax using a set of patterns that was defined beforehand. After the conversion process is done, the cypher query will search the knowledge graph according to the path that is written in the cypher query. Once the data has been found, the QA system will return the result to the user.

The following is an example of input into our QA system, the cypher query, and the output. The input: 'Show all Hadiths that mention eating in the *matn*.' The rule: Show all Hadith that mention $[Topic] in the *matn*. The cypher query:

```
MATCH (b:Book)<-[:FROM_BOOK]-(m:Matn)-[:NARRATED_BY]->(s:Sanad) WHERE toLower(m.matn) =~ '(?i).*\\b{eating}\\b.*' RETURN b.book AS Book, m.number AS Number, m.matn AS Matn, s.name AS Sanad
```

Explanation of the cyper query: Ne04J starts the traversing process from nodes of the type `Matn`. From these nodes, Neo4J picks the `Matn` nodes that have two patterns: (1) those that have a relationship `FROM_BOOK` with nodes of the type `Book`, and (2) those that have a relationship `NARRATED_BY` with nodes of the type `Sanad`**.** For all `Matn` nodes that fulfill these two patterns, Neo4J selects only those that have a string containing '`eating`' as a label inside the node content.

Finally, the QA system reports how many Hadiths mention eating in the *matn* (55 Hadiths). For each Hadith, we display the following information: the number of the book, the number of the Hadith in the book, the string value for the *matn*, and the string value for the *sanad*.

3.2 Components of our QA System

We provide the complete architecture of our approach in Figure 2. From this figure, it can be seen that our approach has the following three modules:

1. Text parsing module. We use regular Python expressions to extract the words from the sentence of the user query. For example: the user query is: **Is eating cake halal?** From this query, we are interested in two things: (1) the topic word and the pattern of the sentence. The first module will extract **eating cake** as the topic label and then the module will report that the pattern of the query is: '`Is $[Topic] [Halal/Haram]?`' The text parsing module receives patterns as input. We defined the following six patterns: Show all Hadith that mention `$[Topic]` in the *matn*; Show all Hadiths narrated by the `$[Sanad]` that talk about `$[Topic]`; Show all Hadiths that talk about `$[Topic]` and `$[Topic]`; Is `$[Topic]` [Halal/Haram]?; What is the sanad that mentions `$[Topic]` in the Hadith?, Is there any Hadith that talks about `$[Topic]?` We note that user queries are prone to errors, therefore we use the Levenshtein distance module included in Neo4J to measure the similarity between the user query and the patterns that we have defined beforehand.
2. Translator module. This module translates a pattern + topic to a cypher query. In this module, we construct a cypher query by applying text concatenation towards the topic and pattern produced by the text parsing module. Then, we send this cypher query to Neo4J, which will execute the query and traverse the knowledge graph using its graph traversing mechanism. We collect the answers from Neo4J using the output module.
3. Output module. This module is responsible for displaying the result of the query that is reported by Neo4J.

As for the expert comment function, every time an expert gives a comment, the ModifyKG function will be triggered (see Figure 3). This function creates a node in Neo4J containing the content of the comment and a relationship/edge that connects the comment to a certain Hadith in our knowledge graph. The new node and edge will enrich the existing Neo4J knowledge graph.

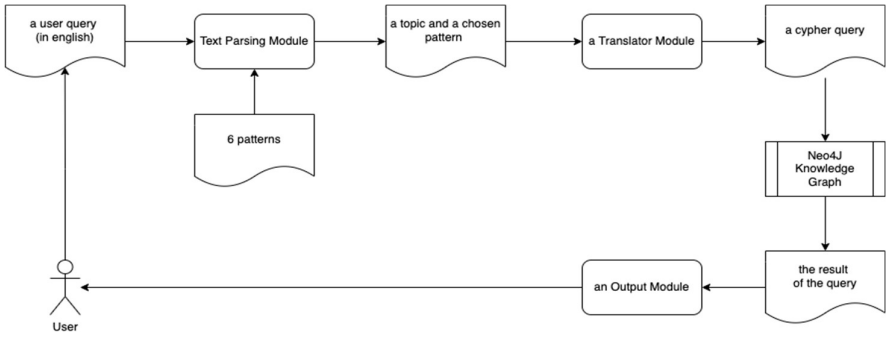

Figure 2 Architecture of our approach.

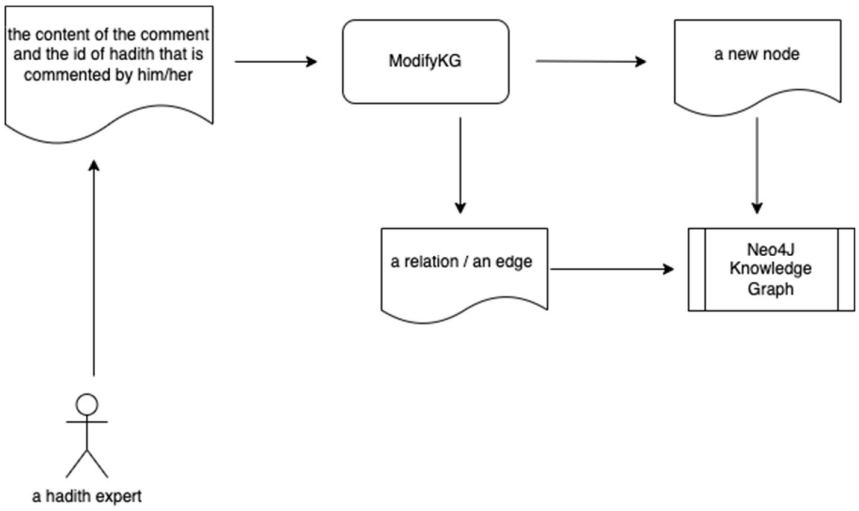

Figure 3 Illustration of the expert comment function.

4 Evaluation

To evaluate our approach, we focused on the following two aspects: coverage and correctness (accuracy). Coverage and correctness (accuracy) are common parameters to assess the quality of knowledge graphs [15]. Related to the coverage aspect, we are interested in measuring how complete our knowledge

graph is compared to the original Hadith dataset. Related to the correctness aspect, we involved a Hadith expert who is fluent in the Arabic language and has memorized many Hadiths to help us in measuring two parameters: top-1 accuracy and top-5 accuracy.

4.1 Platform for Knowledge Graph

Due to the completeness of the features and the breadth of support from third parties, we chose Neo4J [16] as the platform for storing and managing the Hadith in graph format.

4.2 The Data Source for Our Knowledge Graph

To populate our knowledge graph, we choose the Hadith collection from http://sahih-bukhari.com/. This website consists of nine volumes, with several books in each volume. Table 1 shows the statistics of the Hadiths in http://sahih-bukhari.com.

Table 1 The statistics of the Hadith in http://sahih-bukhari.com.

Volume	Number of Books	Number of Hadiths in the volume
1	12	832
2	14	823
3	24	895
4	6	40
5	3	749
6	2	582

The free license of Neo4J that we used, only allows 1,024 characters for each node. Therefore, we were only able to insert 6,412 Hadiths into our knowledge graph.

4.3 Coverage

For the coverage aspect, we checked whether our knowledge graph contained 6,412 Hadiths from http://sahih-bukhari.com. In this experiment, we randomly selected twenty Hadiths from http://sahih-bukhari.com and then we compared the content and the narrator of these Hadiths with the Hadiths we had in our knowledge graph. Table 2 shows the result of this comparison.

Based on the result presented in Table 2, we could conclude that apart from Hadiths with more 1024 characters, our Hadith knowledge graph contained all Hadiths from the original source (http://sahih-bukhari.com/).

Table 2 The result for coverage test.

Node ID in KG	Position of Hadith in Original Source	Comparison Result
876494	the 494th Hadith in the 8th volume, and 76th book	same
765294	the 294th Hadith in the 7th volume, and 65th book	same
224511	the 511th Hadith in the 2th volume, and 24th book	same
771665	the 665 Hadith in the 7th volume, and 71th book	same
334402	the 402th Hadith in the 3th volume, and 34th book	same
333243	the 243th Hadith in the 3th volume, and 33th book	same
880762	the 762th Hadith in the 8th volume, and 80th book	same
14238	the 238th Hadith in the 1th volume, and 4th book	same
661534	the 534th Hadith in the 6th volume, and 61th book	same
762171	the 171th Hadith in the 7th volume, and 62th book	same
332226	the 226th Hadith in the 3rd volume, and 32nd book	same
87369	the 69th Hadith in the 8th volume, and 73th book	same
18471	the 471th Hadith in the 1th volume, and 8th book	same
45267	the 67th Hadith in the 4th volume, and 52th book	same
221245	the 245th Hadith in the 2th volume, and 21st book	same
18438	the 438th Hadith in the 1st volume, and 8th book	same
660237	the 237th Hadith in the 6th volume, and 60th book	same
452232	the 232th Hadith in the 4th volume, and 52nd book	same
18372	the 372nd Hadith in the 1st volume, and 8th book	same
452199	the 199th Hadith in the 4th volume, and 52nd book	same

4.4 Correctness

In this experiment, we focused on the following two parameters: top-1 accuracy and top-5 accuracy. Top-1 accuracy means that given a question, our QA system will return several answers, and from all the answers, if the most correct answer appears in the first position, one point will be added for top-1 accuracy. Top-5 accuracy means that given a question, our QA system will return several answers, and if one out of all top-5 answers is correct, one point will be added for top-5 accuracy.

To decide the correctness of the answers from our QA system, we asked a Hadith expert to help us. For the top-1 accuracy and top-5 accuracy experiments, we asked the Hadith expert to create twenty questions. These questions were used for measuring both parameters. From Table 3, we can see that for both top-1 and top-5 accuracy, the accuracy of our QA system increased when we used the Levenshtein distance function. This function helps in finding answers from user queries that have a similar meaning but different words than the content of the Hadiths in our Hadith KG.

Table 3 Result of the coverage test.

Question	with Levenshtein Distance			without Levenshtein Distance		
	Top 1	Top 5	Top 5 Prob	Top 1	Top 5	Top 5 Prob
Is there any Hadith that mentions Hijrah?	1	3	0,6	0	0	0
Is it haram to drink alcohol?	1	5	1	0	0	0
What is the punishment for illegal sexual intercourse?	1	1	0,2	0	0	0
Is there any Hadith that mentions the Throne?	1	4	0,8	1	4	0,8
Is there any Hadith that mentions the Qadr Night?	0	2	0,4	0	0	0
Is there any Hadith that mentions the Friday prayer?	1	5	1	0	0	0
Is it halal/haram for a woman to travel for more than three days without a Mahram?	1	2	0,4	0	0	0
Show all Hadith that mention war in the *matn.*	1	2	0,4	1	2	0,4
Show all Hadiths about food and drink	1	5	1	1	5	1
Which sanads mention the Wasq?	1	1	0,2	1	1	0,2
Show all Hadiths narrated by Ibn Umar mention sin.	1	3	0,6	1	3	0,6
Show all Hadiths that mention angels in the *matn.*	1	5	1	1	5	1
Show all Hadiths that mention fighting in the *matn.*	1	3	0,6	1	3	0,6
Is there any Hadith that mentions the Adhan?	1	5	1	1	5	1
Is harming halal or haram?	1	5	1	1	5	1
How many times are the words of the Adhan repeated?	1	3	0.6	0	0	0
Which sanad mentions the Badr war?	0	1	0,2	0	0	0
Show all Hadiths narrated by Abu Huraira that mention fasting.	1	5	1	1	5	1
Which sanads mention the Adhan?	1	4	0,8	0	0	0
Is there any Hadith that talks about Heraclius?	1	1	0,2	1	1	0
Sum	18	65	13	11	39	7,6

Take as an example, the second user query: 'Is it haram to drink alcohol?' There was no word 'haram' related to drinking in our entire Hadith KG. Therefore, we could not find any answer when we searched this word syntactically. However, when we used the Levenshtein distance function, we were able to find that the word 'haram' has the same meaning as 'prohibited'. The word 'prohibited' is related to alcoholic drinks in a number of Hadiths.

By applying Levenshtein distance in our approach, we were able to increase the top-1 accuracy from 55% to 90% and we were also able to increase the top-5 accuracy from 38% to 65% (see Tables 4 and 5).

Table 4 Scores for top-1 accuracy.

Top 1 % with Levenshtein Distance	Top 1% without Levenshtein Distance
0.9	0.55

Table 5 Scores for top-5 accuracy.

Top 5% with Levenshtein Distance	Top 5% without Levenshtein Distance
0.65	0,38

To further improve the performance of our QA system in terms of the top-1 accuracy and top-5 accuracy metrics, we added an expert comment function. With this function, an expert can give their comments about a certain Hadith. For example, the expert choses a node in our Hadith KG that mentions the Night of Qadr (let us assume the id of this node is 660228). The expert entered the following comment: 'This Hadith tells us the virtues in the Qadr Night or the Night of Qadr.' Our system then inserts this comment as a new node in the Hadith KG that is connected to node 660228.

Besides inserting a new node into the Hadith KG, by enabling the expert comment function, our system will start the traversing process from nodes that have expert comments. If our system finds a node, then it will collect all the Hadith nodes that are connected to this node. The expert comment function increased the accuracy of our approach. We tested this expert comment function for all queries in Table 2, and we now achieved 95% top-1 accuracy.

5 Conclusion

In this paper, we proposed a question-answering system on a Hadith knowledge graph. Pioneering a QA system on a knowledge graph, our approach showed a promising result in terms of coverage and correctness. This proves that a knowledge graph is suitable for representing the Hadith and also for doing the reasoning task. Our approach also proves that the addition of expert comments to the Hadith knowledge graph can improve the top-1 accuracy parameter.

For future works, we are planning to make a Hadith knowledge graph for other languages, such as Arabic or Indonesian. We are also interested in using machine learning techniques for improving the performance of our system.

Acknowledgements

This work was supported by Telkom University. We would like to thank the people behind http://sahih-bukhari.com/ for providing the Hadith texts.

References

[1] Oxford Advanced Learner's Dictionary, "Hadith." *Hadith Noun - Definition, Pictures, Pronunciation and Usage Notes,* https://www.oxfordlearnersdictionaries.com/definition/english/hadith. (06 August 2022).

[2] Arrahman, F., Lhaksmana, K.M. & Murdiansyah, D.T., *Topic Classification of Islamic Consultation Question and Answer Using Supervised Learning*, 2021 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS), DOI: 10.1109/ICADEIS52521.2021.9701944. pp. 1-6, 2021.

[3] Munadi, M. & bin Salman, A.M., 2018, October. *The Development of Hadith Expert Comparison of Science Curriculum of Hadith Ma'had Aly and PTKIN*, in International Conference on Islamic Education (ICIE 2018), Atlantis Press, pp. 58-65., 2018.

[4] Lan, Y., He, G., Jiang, J., Jiang, J., Zhao, W.X. & Wen, J.R., *A survey on complex knowledge base question answering: Methods, challenges and solutions*, Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21) Survey Track, pp.4483-4491, 2021.

[5] Singhal, A., *Introducing the Knowledge Graph: Things, not Strings*, Google Blog, from https://blog.google/products/search/introducing-knowledge-graph-things-not/ (16 May 2012).

[6] Chang, S., *Scaling Knowledge Access and Retrieval at Airbnb*, The Airbnb Tech Blog, from https://medium.com/airbnb-engineering/scaling-knowledge-access-and-retrieval-at-airbnb-665b6ba21e95 (4 September 2018).

[7] Dong, X.L., *Challenges and Innovations in Building a Product Knowledge Graph*, in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2869-2869), 2018.

[8] Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson A. & Taylor, J., *Industry-scale Knowledge Graphs: Lessons and Challenges*, ACM Queue **17**(2), pp. 48-75, 2019.

[9] He, Q., Yang, J. & Shi, B., *Constructing Knowledge Graph for Social Networks in a Deep and Holistic Way*, in Companion Proceedings of the Web Conference 2020, pp. 307-308, 2020.

[10] Naemah, N. & Saad, S., *Question-answering System Supporting Vector Machine Method for Hadith Domain*, Journal of Theoretical & Applied Information Technology, **95**(7), pp. 1510-1524, 2017.

[11] Assagaf, J., *Question Answering Machine for Islamic Factual Questions Based on English Translation of Hadith*, eProceedings of Engineering, Undergraduate Thesis, Telkom University, Bandung, Indonesia, 2019.

[12] Abdi, A., Hasan, S., Arshi, M., Shamsuddin, S.M. & Idris, N., *A Question-answering System in Hadith Using Linguistic Knowledge*, Computer Speech & Language, **60**, 101023, 2020.
[13] Blumauer, A., *Question Answering based on Knowledge Graphs*, LinkedIn Blog, 2021. (21 July 2022).
[14] Ehrlinger, L. & Wöß, W., *Towards a Definition of Knowledge Graphs,* SEMANTiCS (Posters, Demos, SuCCESS), **48,** pp. 1-4), 2016.
[15] Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G.D., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S. & Ngomo, A.C.N., *Knowledge Graphs*, Synthesis Lectures on Data, Semantics, and Knowledge, **12**(2), pp. 1-257, 2021.
[16] Miller, J.J., *Graph Database Applications and Concepts with Neo4j*, Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA, **2324**(36), pp. 141-147, 2013.