ISSN : 2249-7862 (Print) : 2455-877X (Online)
www.aureoleonline.in

Cosine Similarity for Question Matching

Jinsu Ann Mathew[1,*], Joe Jacob[1], Ninan Sajeeth Philip[2]

[1] Department of Physics, Newman College, Thodupuzha.

[2]Artificial Intelligence Research and Intelligent Systems, Thelliyoor

[*] Email: jinsuann91@gmail.com

Abstract: Cosine similarity is a measure that calculates the cosine of the angle between two given n-dimensional vectors in an n-dimensional space. Mathematically, it is defined as the dot product of the vectors divided by their magnitude. When it comes to computing the similarity between two things, the cosine similarity algorithm has proven to be extremely useful. In this paper, we propose the use of cosine similarity to compare the input query to the system database's questions. By turning the words or phrases within the sentence or document into a vectorized form of representation, the degree of similarity between two documents can be quantified. Here, a screenshot of the question is provided along with the expected answer as target. The machine is trained using cosine similarity to predict them correctly. Between each training question and the input question, we compute the cosine similarity. A cosine similarity of 0 would indicate that there are no similarities between the two documents, while a cosine similarity of 1 suggests that the two documents are exactly the same. We calculate the cosine similarity between each query and each document and use the similarity score to determine which document is the closest match for each query. Our research shows that the database can discover the right matching document with a high degree of accuracy, reducing the requirement for human intervention. In contrast to traditional image processing techniques, our study shows how cosine similarity may be used as a very effective machine learning tool.

Keywords: Cosine Similarity, bag of words, similarity score.

INTRODUCTION

In today's world, computational intelligence is being employed more and more. All industries are putting new technologies into practice to enhance their current processes. The concepts of machine learning and data science are being applied in business, finance, manufacturing, healthcare and other industries. The education sector has advanced significantly as a result of Artificial Intelligence (AI) .The use of computer intelligence to enhance the educational system is a subject of extensive research. Routine educational tasks like grading and conducting exams can be automated with AI. A lot of question and answer boards and FAQ pages are available on many websites. Online services have been compiling fairly sizable archives of questions and their answers over the past few years. Finding questions in the archive that are semantically comparable to a user's question is one of the key duties of a question and answer service.

This paper is about the problem of question search. Given a question as a query, we have to return questions that are semantically equivalent or close to the queried

question. This allows for the retrieval of high-quality answers from the archive. Accurate similarity measures between questions are critical in such retrieval systems. Many metrics, including those based on Euclidean distance, Cosine, Jaccard, Dice, and Jensen-Shannon divergence, have been presented in recent years to address various information retrieval and similarity measurement issues. The most popular of the available metrics is cosine, which determines the angle between two vectors. In this study, we use the cosine similarity score between each archived question and the input query to determine which match is most appropriate.

Methodology

The goal of this research is to illustrate how to perform a calculation of similarity between a reference question and an input question. To find similarities across queries, we have used a number of procedures. The first step is the construction of an archive of questions. In order to do this, we provided screenshots of the questions and placed them in a folder. The second step is to extract all the textual information from the image. The preprocessing of this extracted text comes next. The end result of this method is a list of keywords for each document. Using all of these chosen keywords, we then developed a dictionary. A corpus is then constructed and each query is saved as a corresponding bag of words within it. The calculation of similarity comes next. For this, each item in the corpus is converted to corresponding vectors. With questions presented as vectors, we can calculate the similarity of two questions as the correlation between their corresponding vectors, which can be further quantified as the cosine of the angle between the two vectors. Figure 1 depicts the angle in two dimensions, but in reality, there are tens of thousands of dimensions to a document.

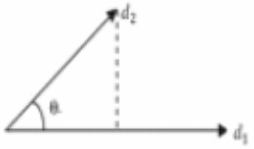


Figure 1: Angle between documents

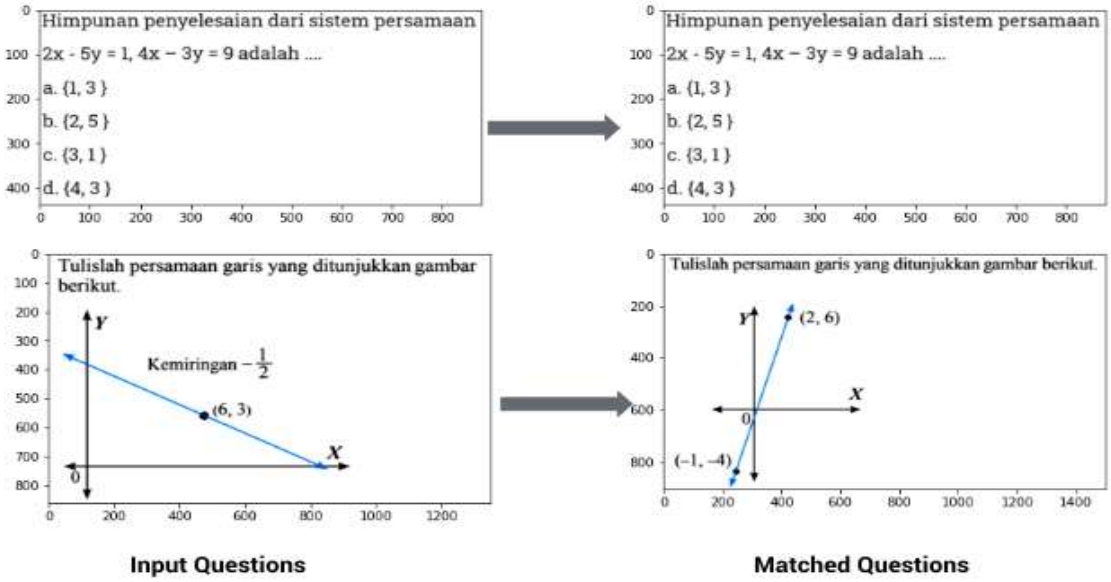
Based on vector similarity, similarity between two vectors can be defined as

$$\text{Cosine Similarity}\left(\vec{d}_1, \vec{d}_2\right) = \frac{\vec{d}_1 . \vec{d}_2}{\vec{d}_1 \times \vec{d}_2}$$

where $\overline{d1}$ and $\overline{d2}$ are n-dimensional vectors and n is the no:of keywords in the dictionary. Therefore , each dimension represents a term with its weight in the document, which is non-negative. As a result, the cosine similarity is non-negative and bounded between [0,1]. So if the angle between the vectors is 0 degrees, then the cosine similarity is 1 or we can say that both are similar.
So, when an input question is given, we convert it into a vector after performing all of the preprocessing steps mentioned above, and then find the similarity score with each

question in the archive. As the best match, the one with the highest similarity score is chosen.

CONCLUSION



This paper demonstrates how to use cosine similarity to find similar questions in an archive. In the figure above, we can see two input questions and corresponding matched questions chosen based on similarity score. The experimental results show that by measuring cosine similarities, we can automatically find similar question pairs. A large number of similar question pairs can be gathered by applying this technique to many different question collections. The proposed similarity measures can be used to cluster question-answer pairs, and the clusters can then be used to improve the performance of question and answer retrieval systems.

REFERENCES

1. Jeon, Jiwoon & Croft, W. & Lee, Joon. (2005). Finding semantically similar questions based on their answers. 617-618. 10.1145/1076034.1076156.

2. Huang, Anna. (2008). Similarity measures for text document clustering. Proceedings of the 6th New Zealand Computer Science Research Student Conference.

3. Li, Baoli. (2013). Distance Weighted Cosine Similarity Measure for Text Classification. 10.1007/978-3-642-41278-3_74.

4. Duan, Huizhong & Cao, Yunbo & Lin, Chin-Yew & Yu, Yong. (2008). Searching Questions by Identifying Question Topic and Question Focus.. 156-164.
5. Jeon, Jiwoon et al.(2005) “Finding similar questions in large question and answer archives.” *International Conference on Information and Knowledge Management*.

6. Xia, Peipei & Zhang, Li & Li, Fanzhang. (2015). Learning Similarity with Cosine Similarity Ensemble. Information Sciences. 307. 10.1016/j.ins.2015.02.024.

7. Rahutomo, Faisal & Kitasuka, Teruaki & Aritsugi, Masayoshi. (2012). Semantic Cosine Similarity.

8. Bahel, V., & Thomas, A. (2021). Text similarity analysis for evaluation of descriptive answers. *ArXiv, abs/2105.02935*.

9. Gunawan, Dani & Sembiring, C & Budiman, Mohammad. (2018). The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. Journal of Physics: Conference Series. 978. 012120. 10.1088/1742-6596/978/1/012120.