Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

A PROTOTYPE SYSTEM FOR CLUSTERING COVID-19 RESEARCH PAPERS

Abdolreza Abhari
Mahfuja Nilufar

Department of Computer Science
Ryerson University
350 Victoria St
Toronto, ON M5B 2K3, CANADA

ABSTRACT

We build a COVID-19 database where the related papers are added into the system dynamically along with the clustering of similar papers. The clustering has been done by two popular NLP models. The goal is building a database software that compares the whole body of all COVID-19 related papers to find similar ones. We developed a prototype by considering abstracts, titles, and the full body of papers. Simulating different searching scenarios and achieving similarity scores evaluated by the Microsoft academic similarity tool show abstract processing outperforms title processing. The results achieved by the developed prototype also proves the correctness of another hypothesis, which is integrating database search features and NLP methods to compare the whole body of similar papers increases the searching capability even more. However, the time spent on creating the clusters shows the scalability of ideal software that can process COVID-19 papers continuously is a significant challenge.

1 INTRODUCTION

The growth of scientific literature of COVID-19 makes it hard to keep track and find similar papers. Finding similar papers for both semantic and lexical similarity is a time-consuming process that requires manual validation. By building a system prototype, which is a database software that uses NLP text mining technique to find similar research papers for COVID-19 related queries, we examine the scalability and performance of different methods. Identical to our ultimate goal, the prototype is an intelligent relational database system capable of showing the query results without writing any SQL query. We already did COVID-19 trend analysis and showed in (Kazemi 2017) that word2vec is 22% more accurate than tf-idf.

2 METHODOLOGY

The goal of our research is to cluster similar types of scientific papers. We simulated our database software by considering the title, abstract, and whole text body data. Our first approach is WoB tf-idf, and measuring cosine similarity to allow k-means clustering. The second approach is the word2vec model for learning word embeddings from raw text data. We used two different cloud computing instances: A high computing instance with 192 GiB memory and 48 CPU for full body data, and an instance with 16 CPU and 64 GiB memory for the title and abstract clustering. We were able to extract the text body of around 31k papers.

3 RESULTS

On our prototype database software, we showed different scenarios by analyzing three levels of difficulties for the queries related to COVID-19 papers. To find the result of an easy level query, we used a scenario to find the related papers that show the effects of the virus on elderlies. By searching with the keyword 'elderly' on the title or abstract fields of our prototype database software, we get four papers that have the same cluster number and potentially have the answer to this query. The second and third scenarios are queries with average and complex difficulty levels by searching for different keywords.

We used a language similarity package from Microsoft academic graph to validate the clustering results by calculating a similarity score for each group that takes two strings as input and returns similarity scores as output. It uses pre-trained word embedding, which is trained on Microsoft academic graph corpus. The time of forming of clusters in each model was between 4 to 6 hours. The numbers of clusters range from 5,330 to 7,097 for word2vec and from 5,275 to 7,101 for tf-idf in clustering of 44,232 papers. The average members of the clusters are similarly between 6.2 to 8.4 for both methods. More results are shown in Table 1.

Table 1: Similarity score and number of clusters for each method.

Model	**Average similarity score**	**final_score=average _similarity_score *cluster members**	**No of the clusters with one member**	**No of the clusters with more than 20 members**
TF-IDF(title)	0.76	3.69	2,025	363
Word2Vec(title)	0.59	3.74	307	25
TF-IDF(Abstract)	0.79	5.94	707	101
Word2vec(Abstract)	0.77	6.14	224	5
TF-IDF(full_body)	0.84	7.2	497	53
Word2vec(full_body)	0.83	7.4	203	8

By multiplying average similarity scores with the no. of members in each cluster, we get a final score that shows word2vec performs better than tf-idf. Comparing title and abstract clustering we achieved higher accuracy for word embedding than tf_idf. Another observation is adding abstract and full body of papers increases the scores. In all cases, word2vec has less number of single-member clusters and large clusters compared to tf_idf. Overall, we found word2vec performs better than the tf-idf method.

4 CONCLUSION

Since building and performance measurement of a scalable and intelligent system to find similar scientific COVID-19 papers is difficult, we developed a prototype to test such a system. The developed prototype is a combination of a relational database and NLP processing for processing the title, abstract, and full text body of the papers. Firstly, we found that better similarity results can be achieved by processing more words (i.e., abstract versus the titles). Secondly, by simulation different scenarios, we proved that NLP methods should be combined with database features to answer different queries. Finally, the achieved results showed that to build a real system text processing of the whole papers should be combined with the Node2vec model to get a more accurate and scalable system.

ACKNOWLEDGEMENT

This research is funded by FOS Dean Research Fund DRF-CONNECT, Ryerson University, Canada.

REFERENCES

Kazemi, B., and A. Abhari. 2017. "A Comparative Study on Content-based Paper-to-Paper Recommendation Approaches". In *Proceedings of SpringSim - 20th Communications & Networking Symposium,* April 23rd-27th, Virginia Beach, Virginia.