Contextual Journal Recommendation and Query Search Engine Using Word Embedding

1st Nayanita Saha
Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur
M21CS010

2nd Swapnil S. Mane
Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur
P21CS014

***Abstract*—A large number of research articles are being published in conferences and journals. Most researchers identify their related work by searching for keywords or reading only renowned conferences and journal articles. Furthermore, researchers frequently fail to publish articles in appropriate journals due to a lack of contextual understanding about publications. In order to search the query in an article, it is also necessary to comprehend the semantics of the text in unstructured data. Natural language processing aids in the extraction of usable facts from unstructured textual material. One approach to get around this problem is to use word embedding. This article proposes a Context-based Journal Recommendation system based on word embedding and a Context-based Search Engine. The dataset is self-contained, and the distance is calculated using cosine similarity. The built system provides contextual functionality that the existing entity-based journal recommender and search engine do not. The proposed methodology will aid aspiring researchers in determining which journal to submit their work to for publication.**

***Index Terms*—Text Context, Search Engine, Natural Language Processing, Word Embedding, Journal Recommendation, Unstructured Data**

I. Introduction

In the era of social media where data is increasing continuously and for this we need to have a search engine that can filter the information according to the content that the user gives to a system. Here the recommendation system comes in picture which provides an approach to facilitate the user's desire. Some recommendation systems recommend research papers while some recommend articles, books, and useful chapters on behalf of their interesting research area or topic. It is very challenging for the recommendation system to provide a perfect match as per the requirement of people. In this paper we have implemented the Journal Recommendation System (JRS) that helps the author in publishing their articles or papers in a suitable journal. A large portion of input data is in the form of unstructured text data. This creates a need for the system to analyze the useful insights from this unstructured data. Natural language processing (NLP) plays a vital role in this scenario. The searching interface helps the user to get the most relevant data concerning the search query. Search query is the most used way for retrieving relevant data from the data repositories. As the size of data is very large, it becomes a challenge for a searching system to get only the relevant document as it also depends on the context of keywords used in the system. In our proposed system we have used the concept of search query by using NLP and machine learning, to determine the context of the keyword. There are various ways to determine the semantic of the text, like word embedding, here we will assign a vector value to the word according to its context, and words with similar contexts will be placed close to each other. So after getting the context of the text or keyword the system will search for the most relevant documents from the data repository. The proposed system would mainly work on determining the context of the keyword in the query by using different word embedding techniques. The system will also make the query more machine-readable so as to get relevant data from the documents. The further reports explored are as follows, Related work of the report in section II. The proposed method is in section III. Experiment and result analysis in section IV. Finally, the conclusion is in section V.

II. Related Work

In NLP, word embedding can be used for a variety of purposes. The embedding of the word vectors aids in the identification of a list of words used in comparable contexts to a particular term. Many researchers have used various word embedding approaches to create diverse systems. Word embeddings include Word2vec, Glove, Bert, Elmo, and others. Roman et al. [1] had developed a methodology for determining the numerical representation of text properties using contextualized word embedding. They also looked at how well a number of machine-learning methods performed when it came to a numerical text representation. Jeffrey Pennington et al. [2] developed the Glove model. The model is trained on a big corpus with millions of tokens, such as Wikipedia. The glove is used to assign vectors based on the word's global context. On similarity challenges and named entity identification, it also beats related models. As indicated by its performance of 75% on a recent word analogy challenge, the model generates a vector space with a meaningful substructure. Using Machine Learning and two-word embedding approaches, Khatri et al. [3] created a sarcasm detection system. This system makes use of Bert and glove embedding. The response vector is determined using the glove, while the context vector is obtained using BERT. With logistic regression as a classifier, the combination of glove and BERT embedding yields a maximum F-measure score of 0.690.

Najafabadi et al. [4] developed a tag recommendation model based on word embedding. Metadata for target items like photos, movies, and Web pages are extracted using tag recommendation models. Most previous efforts use statistical attributes like co-occurrence patterns or phrase frequency to forecast possible tags for a target item to improve tag quality in tag recommendation services. The author has developed a novel tag suggestion system that uses word embedding to analyze the relationship between words in a text and the target object. They focused on feature learning methods and grammatical links between words in a text or sentence. They employed the Skip-gram model to optimize feature values and learn the representation vector of words for tag suggestion, which exhibits advantages of up to 10% in precision over earlier research methods using real data from the Movie lens dataset. Silva-Fuentes et al. [5] used word2vec to determine the meaning of information in an information retrieval system. The terms in the search query are used for searching and any additional terms added to the query. Therefore word embedding is used to expand the search query. Jain et al. [6] have implemented a Journal Recommendation System Using Content-Based Filtering in which they have taken user input as title, abstract keywords and have designed the model by the use of Single Value Decomposition and LSA. The similarity between the input and journal details has been measured by finding the Euclidean distance of vectors, where vectors consist of normalized journal description word frequencies.

III. Proposed System

In unstructured data, the proposed system implements a journal recommendation system and a contextual queries search engine. This section delves into the proposed system's architecture flow. We have gathered various journal publications for this implementation. With the help of pre-defined tools, the data gathering system retrieved the text from all articles. The extracted text is unstructured format needs to do some preprocessing steps. Special symbols, tags, stop words, row text, and other items were removed during preprocessing. Then, using stemming and lemmatization, reduce the term's redundancy. Data preparation creates structured text data that aids in implementation. The proposed system will train the word embedding model to utilize this structured dataset. The vector of each word is obtained through word embedding, which is determined by the co-occurrence matrix. The system used a GolVe word embedding model for this experiment, which is explained in the experiment section. The system has taken an input proposed research article that has yet to be published. Preprocessing is performed on the input article, and then the word embedding model is used to obtain the vector. The distance between two vectors is measured using cosine similarity. The cosine similarity between all of the articles in the dataset and the author's proposed article is calculated. The proposed system would recommend the author rank-wise journals for the respective proposed paper utilizing the cosine similarity proposed system.

This research also proposes a context-based query search engine in addition to the journal recommendation. A journal article with an unstructured format is used for searching input the query. For this task, the system has gone through the sentences and applied a pre-defined sentence tokenization model for sentence extraction. Each sentence of a journal article and an input query vector are obtained using the word embedding method. The cosine similarity with the journal article sentences determines the search query's outcome. The most context-related sentence from journal articles is returned as the search query result. This finger 1 is the proposed model technique for recommending journals and employing word embedding for a context-based query search.

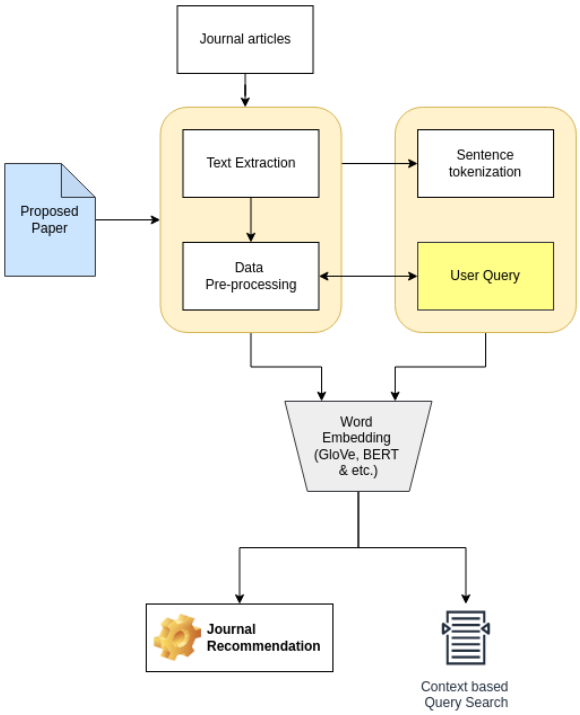

Fig. 1. Workflow of the proposed method

IV. Experiment Details and Result Discussion

A. Data Preparation

For this experiment, we have gathered the research articles of some journals manually. For this implementation, four journals are considered with five articles. The journals are from IEEE and springers publishers which are related to the computer science and engineering domain. The first process is to convert the file into a regular text file. There is a need to extract only the text from the text file, and the images should be neglected as there is no need for an image, and our system is not extracting text from image files. We have used the tika tool for extracting the text data from the portable document format. So the corpus is formed from the documents, but it contains a lot of unwanted text and symbols that are of no use, and it is necessary to clean our corpus, and only the required text should be there. So the next step is data pre-processing.

In this step, we are removing unwanted symbols, punctuation marks, multiple spaces replaced with single space, converting whole text into lowercase, removing all the special characters, stopwords elimination, lemmatization, etc. After the data pre-processing step, the corpus contains only meaningful words. Now the clean corpus is fed into two modules simultaneously to paragraphs to sentences converter and Glove model for extracting predefined vectors. The paragraph to sentences converter makes use of a sentence tokenizer from NLTK. The NLTK is a python library and can be used for many purposes, mainly for NLP. Sent-tokenize [7] smartly split the paragraph into sentences with the help of punctuation marks. The sent-tokenize function uses an instance of Sent Tokeniz from the nltk.tokenize module, which is already being trained and thus very well knows to mark the end and beginning of sentence at what characters and punctuation.

B. Word embedding

After converting every paragraph into sentences, we need the vector for the whole sentence, which can be done by using various word embedding techniques like word2vec, Glove, Fast text, Bert, etc. The Glove is used in this system as it assigns the vector depending upon the word's global context, unlike word2vec which is dependent on the local context. The Glove is one of the techniques used to find the word vector [8]. Word vector is the representation of words in vector space such that words with similar contexts cluster together while different words repel. The Glove model productively influences statistical data by using only non-zero values of word-word co-occurrence matrix for training. Word-word cooccurrence is a sparse matrix. So only the nonzero elements are considered. The Glove is used in this system as it assigns the vector depending upon the word's global context, unlike word2vec which is dependent on the local context. The dimension of each vector can be of various dimensions, say 50, 100, 200, and 300. So Glove provides a pre-trained model trained on big datasets like Wikipedia, Twitter, etc., which can be downloaded directly from the web. It contains millions of words along with their vectors in various dimensions. But there can be some words from the corpus that are not in the pre-trained model, so the vector assigns them zero.

With the help of glove embedding the sentence is embedded into a sentence vector and stored in a database for future use. The dimension of vectors is also important so all the possibilities are checked while testing. With the help of word embedding the main objective of determining the context of the search query and submitted article is achieved. The search result gives relevant documents depending upon the context of the submitted article and search query.

C. Journal Recommendation

For journal recommendations, the proposed system uses word embedding for paragraphs. The vectors are determined for each article of the journals and the input article by the author is yet to be published. The dimensions of the glove embedding for this experiment are 300D. From this, we will get the effective context-related journal article using cosine similarity.

TABLE I
SAMPLE RESULT OF JOURNAL RECOMMENDATION

Rank	Research Paper	Journal	Cosine Similarity
1	On the structural equivalence of co-residents and the measurement of village social structure	Social Networks An International Journal of Structural Analysis	0.9589
2	Artificial neural networks applied for predicting and explaining the education level of Twitter users	Social Network Analysis and Mining	0.9557
3	The influence of social status and network structure on consensus building in collaboration networks	Social Network Analysis and Mining	0.9538
4	Impact of survey design on the estimation of exponential-family random graph models from egocentrically-sampled data	Social Networks An International Journal of Structural Analysis	0.9453
5	Feature Extraction and Selection for Emotion Recognition from Electrodermal Activity	IEEE Transactions on Affective Computing	0.9376

The following table captures the most related research papers for the recommendation done on Article of Kundu and pal is "Fuzzy-Rough Community in Social networks". The cosine similarity is calculated for each research paper and depending upon the cosine similarity value, ranking is done for the same.
The data displayed on the tabular format are having four columns described below:

- Rank: Rank denotes the most accurate research paper based on the Cosine similarity
- Research paper: The name of the research paper that matches the user query
- Journal: It displays the Journal name to which the research paper belongs to
- Cosine Similarity: Cosine similarity is a measure of similarity which calculates the similarity between the research paper with the user query

The figure 2 shows the graphical representation of the above result. The colorful dots in the graph represent the Journal name for the respective research article. For plotting, we have used (Principle Component Analysis) PCA to convert the 300-dimensional vector to the two-dimensional. The X-axis and Y-axis are the principal component 1 and principal component 2. In the plot, the violet point represents the submitted article that is yet to publish "Kundu proposed paper". The nearby points are the most contextually relevant which is under in the dotted circle.

D. Context-based query search

The context-based query search engine is based on a sentence word embedding vector. The word embedding vector is determined for the input query and sentences of articles. The cosine similarity is used for identifying the context-relevant sentences of the query. The result of a context-based search engine is the rank-wise context-relevant sentences with

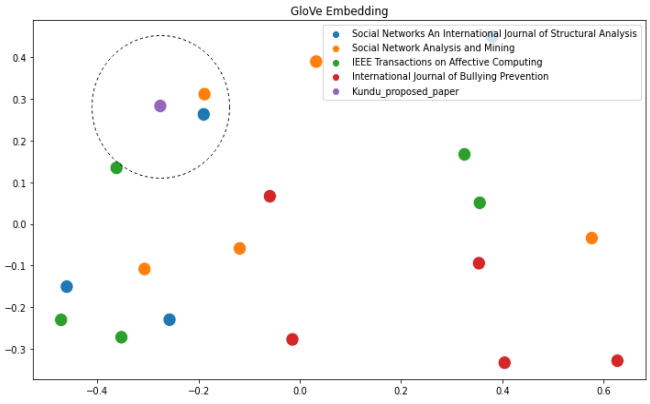

Fig. 2. Contextual position of each article with respective journals

TABLE II
SAMPLE RESULT OF CONTEXTUAL QUERY SEARCH ENGINE

Query	**modularity is in the community detection?**
1	Modularity Approaches in Community Detection Various algorithms of community detection are analyzed based on modularity.
Article	An Analysis of Overlapping Community Detection Algorithms in Social Networks
Similarity	0.87716
2	Overview of Modularity Adaptation based Community Detection.
Article	An Analysis of Overlapping Community Detection Algorithms in Social Networks
Similarity	0.7682

respective journal articles. So from this proposed model users will get the context-related sentences with their research article also. Along with relevant sentences, their contextual similarity score is obtained by cosine similarity.

Table 2 depicts the result of the search query given by the user. The following table captures the result of two search queries in terms of article name and Similarity. For every input query, our search model returns the most similar article name based on the similarity value. A high similarity value indicates the high accuracy of the article with the input.

V. CONCLUSION AND FUTURE SCOPE

The proposed system is entirely reliant on the context of the data and research articles provided. The method proposed combines two applications: journal recommendation and query search. For extracting contextual information, the suggested approach uses a GloVe pre-trained word embedding model. The query is efficiently searched utilizing contextual data from unstructured data in the portable document format. The recommendation's outcome will assist authors in identifying acceptable journals, speeding up their publication process, and enhancing the author's or user's experience. In the future, we'll look into word embedding approaches like Elmo, Bert, and others, but there are a lot of additional factors to consider. Techniques for query expansion are also a desirable addition to any search engine.

REFERENCES

[1] Muhammad Roman, Abdul Shahid, Muhammad Irfan Uddin, Qiaozhi Hua, Shazia Maqsood. 2021. Exploiting Contextual Word Embedding of Authorship and Title of Articles for Discovering Citation Intent Classification

[2] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.

[3] Khatri, Akshay. "Sarcasm detection in tweets with BERT and GloVe embeddings." arXiv preprint arXiv:2006.11512 (2020). arXiv:2006.11512

[4] Maryam Khanian Najafabadi, Madhavan a/l Balan Nair, Azlinah Mohamed.2021 Tag recommendation Model using feature learning via word embedding

[5] Silva-Fuentes, Miguel A., Hugo D. Calderon-Vilca, Edwin F. Calderon-Vilca, and Flor C. Cardenas-Marin˜o. "Semantic Search System using Word Embed- dings for query expansion." In 2019 IEEE PES Innovative Smart Grid Tech- nologies Conference-Latin America (ISGT Latin America), pp. 1-6. IEEE, 2019. https://doi.org/10.1109/ISGT-LA.2019.8894992

[6] Sonal Jain, Harshita Khangarot, Shivank Singh. 2019 Journal Recommendation System Using Content-Based Filtering

[7] https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/ .Last accessed NOVEMBER 1, 2018

[8] pengyan510,https://github.com/pengyan510/nlp-paper-implementation.Last ac-cessed May 5, 2021