Content Based Scientific Article Recommendation System Using Deep Learning Technique

Akhil M. Nair, Oshin Benny, and Jossy George

Abstract The emergence of the era of big data has increased the ease with which scientific users can access academic articles with better efficiency and accuracy from a pool of papers available. With the exponential increase in the number of research papers that are getting published every year, it has made scholars face the problem of information overload where they find it difficult to conduct comprehensive literature surveys. An article recommendation system helps in overcoming this issue by providing users with personalized recommendations based on their interests and choices. The common approaches used for recommendation are Content-Based Filtering (CBF) and Collaborative Filtering (CF). Even though there is much advancement in the field of article recommendation systems, a content-based approach using a deep learning technology is still in its inception. In this work, a C-SAR model using Gated Recurrent Unit (GRU) and association rule mining Apriori algorithm to provide a recommendation of articles based on the similarity in the content were proposed. The combination of a deep learning technique along with a classical algorithm in data mining is expected to provide better results than the state-of-art model in suggesting similar papers.

Keywords Gated recurrent unit · Apriori algorithm · Content-Based recommendation

1 Introduction

Recommender systems are the systems that employ algorithms to help in suggesting relevant items to the users based on their needs and interests such as movies to watch, products to purchase, medications for health issues and so on. In the last few decades, as companies like YouTube and Netflix have risen to power, other companies also experienced the need for recommendation systems based on the user's needs. This

A. M. Nair (✉) · O. Benny · J. George
Department of Computer Science, CHRIST (Deemed to Be University), Lavasa, Pune, India
e-mail: akhil.nair@christuniversity.in

V. Suma et al. (eds.), *Inventive Systems and Control*,
Lecture Notes in Networks and Systems 204,
https://doi.org/10.1007/978-981-16-1395-1_70

will provide the customers with a better experience and satisfaction which later contributes to the expansion and profitability of the companies.

Due to the rapid increase in the amount of information that is available in digital form, the issue of information overload is becoming significant these days. This hinders the timely access of relevant information to many users according to their interests. With the explosive growth of information technology, the number of research papers published every year is also increasing drastically. In such scenarios, research scholars find it difficult to search and access relevant papers according to their area of interest [1]. The exponential growth in the number of research papers and articles published every year has made it challenging for scholars to conduct a comprehensive literature review. Many academic papers are getting published through conferences and journals. Research scholars tend to spend a considerable amount of time as well as other resources to gather the relevant information. They might also use Google Scholar or Citeseer to search for articles based on keywords but does not guarantee the significance of articles.

A research paper recommendation system tries to overcome this issue of information overload by providing the researchers with a personalized recommendation of articles based on their preference [2]. Research paper recommender system finds relevant papers based on the users' current requirements which can be gathered explicitly or implicitly through ratings, user profile, and text reviews. The two main approaches of a recommendation system are Content-Based Filtering (CBF) and Collaborative Filtering (CF). A content-based approach requires information related to the items and their features whereas Collaborative Filtering needs the user's historical preferences on a set of items. The state-of-art models of article recommendation systems focus on Collaborative Filtering and citation counts but very few models have focused on the content for finding the recommendations. With the expansion of Artificial Intelligence and Machine Learning, it has become easy to build recommendation systems based on the requirements. The study proposes a Content Based Scientific Article Recommendation (C-SAR) model by combining the Gated Recurrent Unit and association rule mining namely the Apriori algorithm to provide an additional layer of filtration for similar documents thereby converting relevant articles to comparatively higher relevant articles.

2 Related Work

Even though the content-based scientific article recommendation using the deep learning technology is still in its infancy, various approaches are present today to deal with article recommendations. The recommendation of papers based on content was formulated as a ranking problem [3] with two phases NNselect for selection of papers and NNrank for their ranking. Through this work, a contribution of the new dataset OpenCorpus with seven million articles was also made which could be useful for researchers. The model could have gained better accuracy if the metadata regarding the papers were used along with other attributes. A comparative study between two well-known content-based methods, Term Frequency Inverse-Document-Frequency (TF-IDF) and word-embedding were made and implemented [4] using the PUBMED dataset. The word embedding model obtained 15% better accuracy than TF-IDF with a similarity score of 0.7168. To reach a more reliable and accurate result, the set of target and recommended papers must be provided which is considered as a limitation of the model. Simon et al. [5] proposed a system that helps the user to quickly locate items of interest from a digital library. The application employed a TF-IDF weighting scheme and cosine similarity along with keyword-based vector space models and item representation. Instead of looking into the ratings provided by other users, this work has focused on the user's interests and needs. This content-based approach for research paper recommendations based on a user's query in a digital library has produced better results along with additional features that do not exist in the digital library. Stemming of attributes was not done in the model to reduce the loss of context of the search.

A model that considers the paper's topic and ideas that helps the non-profiled users to obtain a set of relevant papers was introduced [6] which required a single input paper and the main themes were derived as subqueries. The model acquired an accuracy of 80% based on the NDCG metric and proved to be one of the efficient ways for recommendations without a user profile. The inclusion of indexing features may help the model to achieve better performance. A centrality-based approach was proposed by Abdul Samad et al. [7] that analyses the textual and topological similarity measures in a paper recommendation system. A comparison is made based on the performance of the Cosine and Jaccard measure. When performed on the dataset, it was found that topological-based similarity through Cosine achieved 85.2% accuracy, and using Jaccard obtained 61.9%. On the other hand, textual-based similarity obtained 68.9% citation links on abstract and 37.4% citation links on the title. Both similarity measures analyzed only the symmetric relationship of papers. Sometimes, it is necessary to consider asymmetric relationships as well to provide better recommendations. For a DeepWalk based method, a Recall of 0.2285 and NDCG of 0.3602 was obtained which used deep learning based bibliographic content specific recommendation [8]. The matrix used in the model is based only on the paper vector, but not on the citation information. The deep walk based method is expected to outperform other existing models only if the paper contents and citations are used. A deep learning-based study [9] used a Recurrent Neural Network for modelling a recommendation system. The explicit and implicit feedback collected from users was used along with a semantic representation of the paper's title and abstract. The feedback was checked for matching with the representation. The model used Long Short-Term Memory (LSTM) for the semantic representation of articles. The paper recommendation is purely based only on the user actions and their feedback which is not always appreciated. An Advanced Personalized Research Paper Recommendation System (APRPRS) [10] based on User-Profile which applies keyword expansion through semantic analysis was implemented and achieved an accuracy of 85% and user satisfaction level of 89%. A possibility of limitless expansion in the case of keywords makes the model less efficient since it does not have any limits of numbers. Another LSTM based approach Ask Me Any Rating (AMAR) secured an

F1@10 score of 0.66 when experimented on Movielens and DBbook dataset [11]. This article also introduced AMAR Extended Architecture where along with user profile and item representation, item genres were also considered for recommendation. Lack of proper hyperparameter optimization and regularisation makes the model performs less inadequately.

A novel neural probabilistic approach was implemented by Wenyi Huang, et al., [12] which jointly learns semantic representations of citation context and cited papers. The probability of citing a document is calculated using a multi-layer neural network that improved the quality of citation recommendation. The neural probabilistic model together with word representation and document representation learning is employed for citation recommendation and the word2vec model is used for this purpose of learning both word and document representation. The representation of the words and documents are to be learned simultaneously from the citation context and the document pairs that are cited. Zhi Li and Xiaozhu Zou in [13] discussed a comprehensive summary research paper recommendation system by discussing the state-of-art academic paper recommendation methodologies, its advantages, and disadvantages with evaluation metrics and the available datasets. This work is helpful to refer to the basic methods of research paper recommendation system and its performance evaluation metrics. By providing a detailed description of each terminology associated with a research paper recommendation system, this work is well appreciated among the scholars and is very insightful.

A novel concept-based research paper recommendation system that represents research articles in terms of their topics or semantics [14] has managed to acquire accuracy of 74.09% based on the Normalized Discounted Cumulative Gain (NDCG) evaluation metric. Distributed representation of words could be combined to obtain a unique vector for candidate documents which might result in better recommendations. The average accuracy of 88% was achieved for a content-aware citation recommendation system [15] by computing citation relations using the global citation method and cross-reference. The model includes three algorithms for Own Citation relation extraction, Cross-reference calculations, and similarity checking among the papers. A limitation of this model is that it does not consider the year of a paper published or other relationships such as co-authorship. A Convolutional Neural Network (CNN) based recommendation system to predict the latent factors from text information achieved an RMSE value of 3.3481 [16]. The major novelty of the proposed recommendation algorithm is that the text information is used directly to make the content-based recommendation without tagging. The recommendation process involves using the text information regarding the input learning resource, maybe the content itself, or a brief introduction of the content. For the CNN model, the input and its output must be established in the beginning itself which may not be possible at least in a few scenarios.

3 Overview of Architecture

3.1 Gated Recurrent Unit (GRU)

GRU is a gating mechanism in the Recurrent Neural Network. This architecture is proposed to deal with long sequences of data. GRUs are known to be the advanced version of standard Recurrent Neural Networks which uses an update gate and a rest gate in a normal RNN to minimize the vanishing gradient problem. These gates are two vectors that would decide which information needs to be passed to the output.

The gating mechanism creates a memory control of values processed over time. This consists of two gates that control the flow of data through states. They are the update gate and reset gate. Two of them can be considered as vector entries that perform a convex combination. The combination decides on which hidden state information should be updated or reset the hidden state when required. By this, the network learns to filter out irrelevant temporary observations. Figure 1 shows the architecture of GRU.

Gate *rt* controls updates on the internal memory, which is not propagated to the next state. Gate *zt* controls how much of internal memory should be considered in the next state. Equations 1 and 2 represent operations realized by gates *rt* and *zt* to result, and equation 3 shows how the next hidden state is computed in a GRU unit.

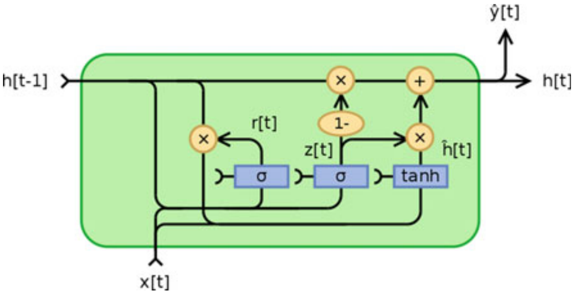
$$rt = \sigma(W_r h_{t-1} + U_r x_t + br) \tag{1}$$

$$zt = \sigma(W_z h_{t-1} + U_z x_t + bz) \tag{2}$$

$$ht = z_t \otimes h_{t-1} \oplus (1 - z) \otimes \tan h(W_h x_t + U_h(r_t \otimes h_{t-1}) + bh)) \tag{3}$$

The update gate *zt* determines the amount of past information that is to be passed into the next state. The reset gate *rt* decides on how much of the previous information has to be neglected, thus resetting the gate values. Both the gates share the same formula but differ in the weights and usage of gates.

Fig. 1 Gated recurrent unit architecture



3.2 Apriori Algorithm

Association rule learning being rule-based machine learning is used to discover interesting relationships and patterns between variables in large databases. Apriori is an algorithm for frequent itemset mining and association rule learning. It identifies the frequent individual items and later extends them to larger and larger sets as long as they appear often in the database. The three most commonly used ways to measure association are support, confidence, and lift. Support defines how popular an item set is, by measuring the proportion of transactions in which an item set appears. Confidence gives the measure of how likely an item is purchased if another item is purchased. Lift is similar to confidence but it measures how likely an item (Y) is purchased when another item (X) is purchased while controlling the popularity of Y.

The key property of the algorithm states that all the subsets of a frequent itemset must be frequent and that if an itemset is infrequent, all its supersets will be infrequent.

4 Methodology

In this work, a model called Content-Scientific Article Recommendation (C-SAR) were proposed in which both GRU and Apriori algorithm are used for finding similar sets of articles. The combination of high-level deep learning, GRU, and a data mining Apriori algorithm will filter the most similar set of documents. Table 1 describes the steps involved in building the C-SAR model.

The methodology of the C-SAR model is described in Fig. 2 in two phases. In the first phase, Gated Recurrent Unit (GRU) technique is used to obtain the similarity of documents and the adjacency matrix. In the second phase, association rule mining based Apriori algorithm would be applied to filter out the most relevant set of documents among the similar documents. Since the model is content-based, the 'title' feature is extracted from the AAN dataset followed by data cleaning. The absence of null or insignificant values indicated that the data is cleaned. For a better result, removal of stop words followed by stemming and lemmatization was done to the 'title' feature. Padding is also done to the text data after one-hot encoding to

Table 1 The steps of the proposed C-SAR model

Step 1	The feature (title) is extracted from the AAN dataset
Step 2	The text data is converted to vectors using Glove pre-trained embedding
Step 3	Similarity probability is calculated for the documents using GRU
Step 4	The probabilities are replaced by 1's and 0's based on a fixed threshold
Step 5	An adjacency matrix is created with the new values(1's and 0's)
Step 6	The new matrix with similarity score is passed to the Apriori Algorithm
Step 7	The set of frequently occurring documents are obtained

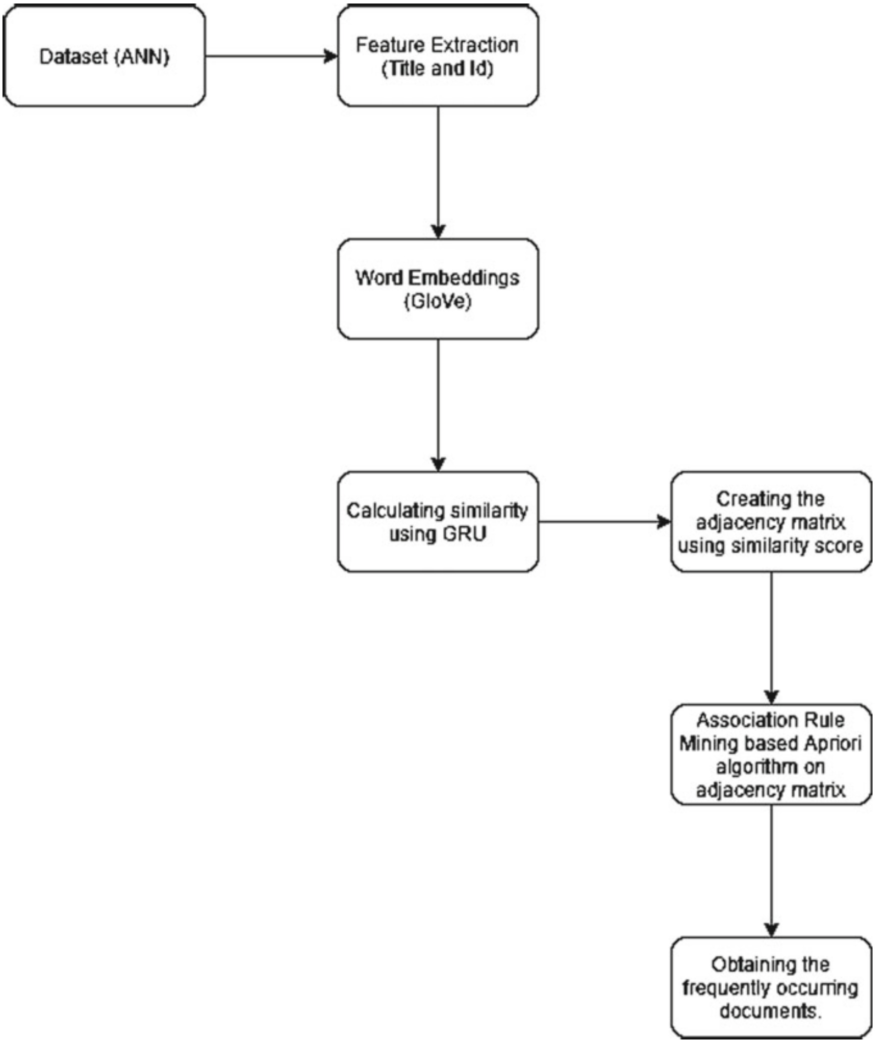

Dataset (ANN)
Feature Extraction (Title and Id)
Word Embeddings (GloVe)
Calculating similarity using GRU
Creating the adjacency matrix using similarity score
Association Rule Mining based Apriori algorithm on adjacency matrix
Obtaining the frequently occurring documents.

Fig. 2 The proposed C-SAR model

make sure that the sequences of data are of the same length to avoid ambiguity. It is performed to make all sequences in a batch to be of standard length.

GloVe, coined from Global Vectors, is an unsupervised learning algorithm that is used to get vector representation of words. This is done by mapping words into a meaningful space where similar words occur together. Such pre-trained embeddings can capture the semantic and syntactic meaning as they are trained on large datasets. They also boost the performance of the model. The main idea behind Glove is to derive relationships between words from Global Statistics and is found to be better than another embedding like Word2Vec. GRU units take up the vector representation and measure the probability of similarities between the input sequences. Two sequences were passed into the GRU model at a time and the similarity score was calculated. A threshold of 0.55 was fixed to filter the most and least similar documents. The results are replaced by 1's and 0's based on the probability scores. The presence of 1 indicates that the documents are similar to each other whereas 0 indicates the least similarity. An adjacency matrix with the results is achieved.

After obtaining the adjacency matrix, it is passed on to the association rule mining Apriori algorithm. This produces an output of the set of frequently occurring documents from the AAN dataset. The detailed architecture of the proposed method for the recognition of handwritten MODI script is discussed in the next subsection.

4.1 Architectural Details of GRU in C-SAR

The proposed architecture of GRU used in the C-SAR model is described in Fig. 3. The 'title' of the two sets of papers are converted into vectors using the Glove

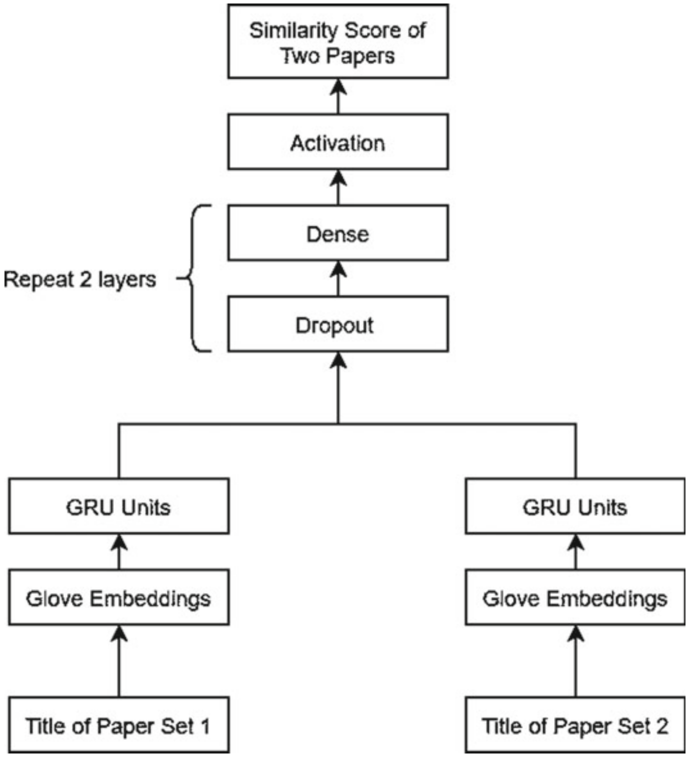


Fig. 3 The architecture of GRU in C-SAR model

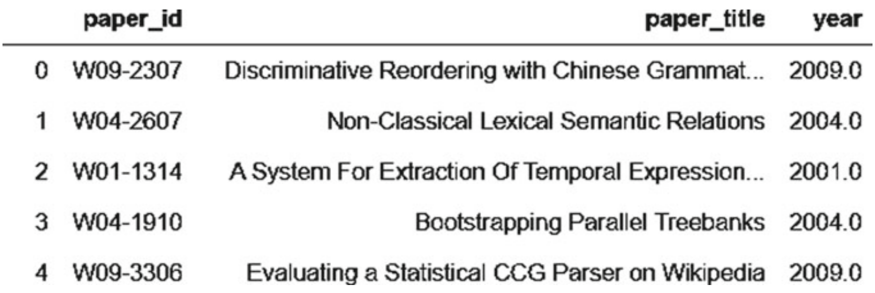
embedding after proper cleaning. The vector format of the data is then separately passed on to two GRU units. Later, the output of the GRU units is concatenated and passed on to dropout and dense layers (2 layers are repeating). A sigmoid activation function calculates the probability of similarity between the two input sequences.

The dropout layer is used for regularisation where the inputs and the connections are excluded from getting activated. This helps in reducing overfitting and improves the model's performance. The dense layer is used to get the activation of the dot product of the input and kernel in a neural network. Dense layers with units of 128 and 64 were used in the model along with the dropout. The activation function used in the model is sigmoid which resulted in a value between 0 and 1. It is used for the models especially when the output need is a probability. They are commonly used for binary classification. Adam optimization along with binary cross-entropy loss is used in the C-SAR model to produce better results.

5 Experimental Study

The experiment is conducted on the Intel i5 processor with 8 GB RAM, using python programming. Description of the data set and the details of the experiment are given in this section. The model was implemented using the ACL Anthology Network (AAN) 2014 released dataset which is 387 MB file. The file consists of folders author_affiliations, citation_summaries, paper_text, and release. It consists of around 23,766 papers along with 1, 24,857 paper citations. Only the feature 'paper_title' along with 'paper_id' were extracted from the dataset. Figure 4 shows the AAN dataset with id and title required for the study which is extracted from the dataset. The Glove pre-trained embedding file was downloaded and used to get the encodings of the text data. It contained 400000 trained words and their representations.

The GRU function from Keras was used in implementing the model. It resulted in the probability of how one document in the AAN dataset is similar to another document. The adjacency matrix obtained from the GRU model is then passed on to the Apriori function from mlxtend which resulted in the set of most frequently



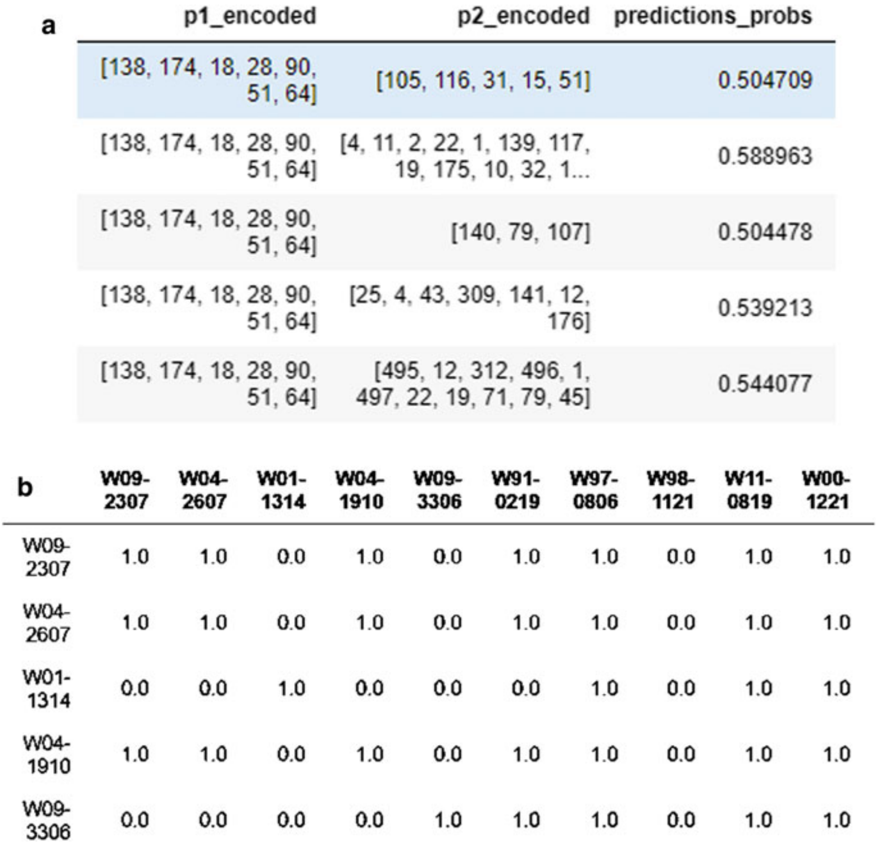
	paper_id	paper_title	year
0	W09-2307	Discriminative Reordering with Chinese Grammat...	2009.0
1	W04-2607	Non-Classical Lexical Semantic Relations	2004.0
2	W01-1314	A System For Extraction Of Temporal Expression...	2001.0
3	W04-1910	Bootstrapping Parallel Treebanks	2004.0
4	W09-3306	Evaluating a Statistical CCG Parser on Wikipedia	2009.0

Fig. 4 AAN dataset

occurring documents along with the minimum support. The minimum support value is a base value used to filter out the documents based on the support value. The set of documents below the minimum support will be rejected while processing.

6 Results and Discussion

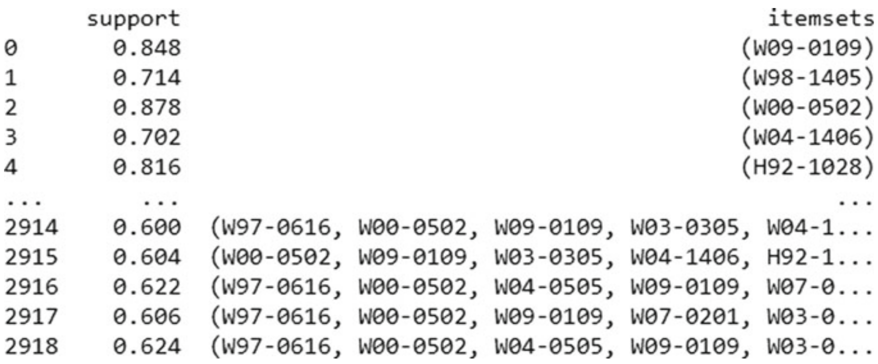
The proposed C-SAR model is expected to outperform the existing state-of-art methods. It is because of the combination of a deep learning technique, Gated Recurrent Unit, and association rule mining Apriori algorithm. The objective of the proposed method is to find a set of most similar documents from the AAN dataset. In the first phase of the C-SAR model, the employment of GRU provided the probability score of the similarity of the documents. Figure 5a shows the sample similarity scores obtained using GRU.



p1_encoded	p2_encoded	predictions_probs
[138, 174, 18, 28, 90, 51, 64]	[105, 116, 31, 15, 51]	0.504709
[138, 174, 18, 28, 90, 51, 64]	[4, 11, 2, 22, 1, 139, 117, 19, 175, 10, 32, 1...	0.588963
[138, 174, 18, 28, 90, 51, 64]	[140, 79, 107]	0.504478
[138, 174, 18, 28, 90, 51, 64]	[25, 4, 43, 309, 141, 12, 176]	0.539213
[138, 174, 18, 28, 90, 51, 64]	[495, 12, 312, 496, 1, 497, 22, 19, 71, 79, 45]	0.544077

	W09-2307	W04-2607	W01-1314	W04-1910	W09-3306	W91-0219	W97-0806	W98-1121	W11-0819	W00-1221
W09-2307	1.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0
W04-2607	1.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0
W01-1314	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0
W04-1910	1.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0	1.0
W09-3306	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	1.0

Fig. 5 **a** Similarity score based on GRU **b** sample adjacency matrix using GRU



```
        support                                                  itemsets
0         0.848                                                (W09-0109)
1         0.714                                                (W98-1405)
2         0.878                                                (W00-0502)
3         0.702                                                (W04-1406)
4         0.816                                                (H92-1028)
...         ...                                                       ...
2914      0.600   (W97-0616, W00-0502, W09-0109, W03-0305, W04-1...
2915      0.604   (W00-0502, W09-0109, W03-0305, W04-1406, H92-1...
2916      0.622   (W97-0616, W00-0502, W04-0505, W09-0109, W07-0...
2917      0.606   (W97-0616, W00-0502, W09-0109, W07-0201, W03-0...
2918      0.624   (W97-0616, W00-0502, W04-0505, W09-0109, W03-0...
```

Fig. 6 Set of frequently occurring documents in AAN dataset

Figure 5b represents the adjacency matrix obtained after replacing the probabilities resulted from the GRU with 1's and 0's based on the threshold value. The adjacency matrix is generated to provide cleaner data to the ARM-based model. This adjacency matrix is used in the Apriori algorithm to find out the most frequently occurring paper_ids and paper_titles to create a narrower range but relevant papers out of the dataset.

The adjacency matrix shown in Fig. 6 has been fed to the Apriori algorithm. The maximum support value received for the paper is 0.848 which depicts 84% of the time paper_id w09-0109 has occurred in the dataset for the given iterations. The minimum support value provided for the algorithm is 0.35 which has been set as a filter to remove all the data items which have a frequency less than 35% and attained the results as shown in Fig. 6.

The huge size of the dataset caused scalability issues and threw memory errors. The increase in the number of data can lead to better accuracy of the model. At the same time, the usage of memory and CPU also gets increased with a number of iterations over a huge dataset.

7 Conclusion

With the increase in the number of scientific publications and the number of research papers, the recommendation system for articles is gaining much significance. It is important for any scholar to get a set of relevant papers related to their field of study. In this work, a Content-based Scientific Article Recommendation (C-SAR) model was proposed based on a deep learning technique. Especially, this model focuses on checking for the papers based on the similarity in their title. A Gated Recurrent Unit method was employed for finding the similarity of documents and association rule mining Apriori algorithm to filter the most frequently occurring set of documents from a similar set. The model is expected to outperform existing models that use simple K-Means Clustering and user representations.

One of the limitations of this model is the memory and time constraints associated with the implementation. The performance can be increased by using cloud services and other super configuration machines. In the future, the efficiency of the model can be improved by considering an optimal threshold value for getting the similarity matrix and a better minimum support count based on the total number of transactions.

References

1. X. Bai, M. Wang, I. Lee, Z. Yang, X. Kong, F. Xia, Scientific Paper Recommendation: A Survey. IEEE Access **7**, 9324–9339 (2019)
2. M. Asim and S. Khusro, “Content Based Call for Papers Recommendation to Researchers”, 12th International Conference on Open Source Systems and Technologies Lahore, Pakistan, 2018, pp. 42–47
3. C. Bhagavatula, S. Feldman, R. Power and W. Ammar,” Content-Based Citation Recommendation”, Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, Vol.1, Jun 2018
4. B. Kazemi and A. Abhari,” A Comparative Study on Content-Based Paper-To-Paper Recommendation Approaches In Scientific Literature”, SpringSim-CNS, Apr 2017, pp. 23–26
5. S. Philip, P.B. Shola and A.O. John, “Application of Content-Based Approach in Research Paper Recommendation System for a Digital Library” International Journal of Advanced Computer Science and Applications, 2014
6. D. Hanyurwimfura, L. Bo, V. Havyarimana, D. Njagi and F. Kagorora,” An Effective Academic Research Papers Recommendation for Non-profiled Users”, International Journal of Hybrid Information Technology, Vol. 8, 2015, pp. 255–272
7. A. Samad, M. A. Islam, M. A. Iqbal and M. Aleem,” Centrality-Based Paper Citation Recommender System”, EAI Endorsed Transactions on Industrial Networks and Intelligent Systems, Jun 2019
8. L. Guo, X. Cai, H. Qin, Y. Guo, F. Li and G. Tian,” Citation Recommendation with a Content-Sensitive DeepWalk based Approach”, International Conference on Data Mining Workshops, Beijing, China, 2019, pp. 538–543
9. H. A. M. Hassan,” Personalized Research Paper Recommendation using Deep Learning”, Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, Jul 2017, pp. 327–330
10. K. Hong, H. Jeon and C. Jeon,” Advanced Personalized Research Paper Recommendation System Based on Expanded User Profile through Semantic Analysis”, International Journal of Digital Content Technology and its Applications, 2013, pp. 67–76
11. A. Suglia, C. Greco, C. Musto, M. Gemmis, P. Lops and G. Semeraro,”A Deep Architecture for Content-based Recommendations Exploiting Recurrent Neural Networks”, Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, Jul 2017, pp. 202–211
12. W. Huang, Z. Wu, C. Liang, P. Mitra, and C.L. Giles,” A Neural Probabilistic Model for Context Based Citation Recommendation”, *AAAI*, 2015
13. Z. Li and X. Zou,” A Review on Personalized Academic Paper Recommendation”, Computer and Information Science, 2019
14. R. Sharma, D. Gopalani and Y. Meena, “Concept-Based Approach for Research Paper Recommendation” PReMI, 2017

15. M. A. Arif, “Content aware citation recommendation system,” International Conference on Emerging Technological Trends, Kollam, 2016, pp. 1–6
16. J. Shu, X. Shen, H. Liu, B. Yi and Z. Zhang,” A content-based recommendation algorithm for learning resources”, Multimedia Systems, 2017