KESDR : A Knowledge-Enriched Semantic Scheme for Scientific Document Recommendation

Vaibhava Lakshmi R[1], Gerard Deepak[2] and Santhanavijayan A[3]

[1] Department of Computer Technology, Madras Institute of Technology, Chennai, India
[2] Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India
[3] Department of Computer Science and Engineering, National Institute of Technology, Tiruchirapalli, India

vaibhavi18092002@gmail.com
gerard.deepak.christuni@gmail.com
vijayana@nitt.edu

Abstract. The Recommendation systems are serving users in the task of picking out items that cater their demands and wishes. In educational research, recommendation systems can issue similar scientific papers to researchers, by facilitating the search process. It has become a fascinating area of study because of the exponentially increasing number of scientific articles. In this paper, a KnowledgeEnriched Semantic Scheme for Scientific Document Recommendation is put forth. The proposed KESDR approach is evaluated using several performance measures by comparison with other models like HDR, BDR, BKA and K-Means+SVM. For this project, the documents are procured from Pharmaceutical domain. The proposed KESDR model gives better performance compared to other models, with a precision of 94.49 % which surpasses the other models in comparison, due to the knowledge aggregation it uses from several heterogeneous knowledge sources such as LOD cloud, DBpedia and NELL.

Keywords: Recommendation systems, K-Means, SVM.

1 Introduction

A recommendation system comes under the category of Information filtering systems that apprises about the preference or rating, an item might get from a user. A plethora of things can be recommended by the system like books, articles, advertisements, jobs, news and movies. Recommendation systems have multifarious applications in the fields of Economics, entertainment, education and scientific research. Recommendation systems are widely used by multinational companies like Youtube, Tinder, Netflix and Amazon. Rapid increment in the number of scholarly articles on internet poses a huge challenge of finding relevant articles. At times, even a highly germane paper is missed by researchers who are neophytes owing to the lack of experience in finding the most

pertinent articles [1]. Scientific document recommendation systems, in specific, recommend scientific articles to research professionals according to their interests. They focus on helping researchers to alleviate information overload and search for pertinent articles by calculating and ranking publication records and by suggesting top N papers to the user elicited from their research ardors [2]. The recommendation algorithms being used are incessantly updated and their accuracy is also ameliorating with time.
Our work incorporates Semantic Intelligence. Semantic Intelligence refers to getting rid of the disparity between human and computer comprehension by making a machine to learn to think in terms of objects like a human does. Hence, Semantic Intelligence technologies play a significant role in developing artificially intelligent Knowledge-based Systems. Semantic Intelligence helps to handle unstructured information and leverage Semantic technologies [21]. Semantic recommender systems use Semantic Intelligence which entirely based on a knowledge base generally defined as a concept diagram or an Ontology. Semantics-oriented recommendation involves augmentation of auxiliary and background crowd-contributed, community-verified knowledge with statistical principles and methods, which could be combined with traditional AI or ML or DL models based on the boundedness, fluidity and the nature of the problem [22].

1.1 Motivation

The rapid spurt in quantity of published content and the effect of its plenitude is known as information explosion. As the quantity of obtainable data increases, managing the information becomes more and more onerous, that can culminate in information overload. Therefore, recommendation systems help to prohibit information overload to a huge extent in a multitude of domains. Also, Machine Learning or Deep learning models alone can't learn from all the entities in the web as, web has got 120 Zettabytes of indexed information. Hence, these models must be assisted by semantically compliant reduction techniques. Scientific documents recommendation is very tedious. Although, data is available readily, there is paucity of domain centric knowledge. Therefore, semantically compliant inbounding and out bounding methodologies are needed for retrieval of scientific documents. The baseline models are not compliant with the cohesive structure of the Web 3.0 where the information density is extensively high. As a result, the proposed Semantically Driven Knowledge Centric model works based on Machine Intelligence which is amalgamation of Machine Learning and Semantic inference and helps overcome the challenges faced using the baseline models.

1.2 Contribution

The proposed model has the following implementations: 1) The TF-IDF model is used to derive the most informative terms. 3) Semantic Similarity is calculated from the most informative terms and the pre-processed query terms so as to construct the set of query relevant informative terms. Knowledge aggregation is followed. 4) The dataset is classified by using the features from the enriched query terms into several categories using the Decision tree classifier. 5) Based on the pre-dominant classes identified from the categories, the Semantic Similarity is computed initially using the Normalized Google

Distance with a threshold of 0.75. All the matching classes are taken into consideration and every instance inside every class is further substituted into Semantic Similarity estimation with the enhanced query terms using the Simpson's Diversity Index and the Cosine Similarity. 6) Based on the increasing order of Cosine similarity, ranking is done and yielded to the user.

1.3 Organization

The paper is organized as follows: Section II presents the literature survey and the existing methodologies. In section III, the architecture of the proposed framework is elucidated. Section IV describes the implementation details along with the results obtained. Finally, in section V, the conclusion is presented along with future works.

2 Related works

Zeynep Akkalyoncu Yilmaz et. al., [1] have put forth Birch, a system that uses BERT for document retrieval by integration with the open-source Anserini information retrieval toolkit to manifest end-to-end search for huge document collections. N. Sakib et. al., [2] have proffered a novel hybrid approach that combines a Content Based Filtering (CBF) recommender module and a Collaborative Filtering (CF) recommender module independently. To enhance efficiency of the recommendation, these two approaches separately include public contextual metadata and paper-citation relationship information. L.Guo et. al., [3] have introduced a scrupulous way for co-authorship modeling that contains the topics of their published papers and the co-author network organization. A graph-based recommendation model consisting of three layers that combines meticulous co-authorship as well as author-paper, paper-keyword relations and paper-citation is created..

T.Dai et. al., [4] have put forward a novel low-rank and sparse matrix factorization-based paper recommendation (LSMFPRec) system for different authors. This approach can effectively assuage the sparsity and cold start problems that occur in traditional matrix factorization based collaborative filtering approaches.X. Bai et. al., [5] have expounded the benefits and significance of the paper recommendation systems. Also, several recommendation algorithms are inspected for comparison. X. Kong et. al., [6] have come forth with a recommendation system for scientific papers, called VOPRec, using vector representation learning of paper in citation networks. It is shown that VOPRec transcends state-of-the-art paper recommendation baselines measured using F1, precision,NDGC and recall.

F.Xia et. al., [7] have presented a novel recommendation approach which includes minutiae on common author relations amongst articles. This approach produces more pertinent recommendations for relevant researchers when compared to a Baseline method.Yeon-Chang Lee et al., [8] have come forth with a hybrid approach suitable for paper recommendation combining the content-based and the graph-based approaches.

Also, several methods with four datasets of DBpia in the context of paper recommendation using content-based or graph-based recommendation are discussed.W. Zhao et. al., [9] have proposed a literature recommendation method based on knowledge-gap, so as to support researchers to cater literature support. A graph-based technique to traverse appropriate knowledge paths, which can help a researcher to learn the required knowledge according to the cognition pattern is designed so as to bridge the knowledge gap. Sharma et. al., [10] have provided a brief overview of popular algorithms and previous systems developed to solve the problem of information explosion. Wu et. al., [11] have proposed an attack model corresponding to recommendations. In the model, the current recommended status and a specified item are scrutinized to estimate the upshots of several attack decisions (addition or deletion of facts), thereby giving rise to the optimal attack combination.
Saar Kuzi et. al., [12] have elucidated the merits of combining deep neural network models and lexical models for the retrieval stage of scientific documents. A leveraging of both semantic and lexical retrieval models is put forward.Rospocher et. al., [13] have assessed a document retrieval methodology employing Knowledge Extraction techniques and Linked Open Data. This approach was implemented in the KE4IR system. In [14-20], several semantically inclined approaches in support of the proposed framework have been depicted.

3 Proposed architecture

techniques and Linked Open Data. This approach was implemented in the KE4IR system. In [14-20], several semantically inclined approaches in support of the proposed framework have been depicted.

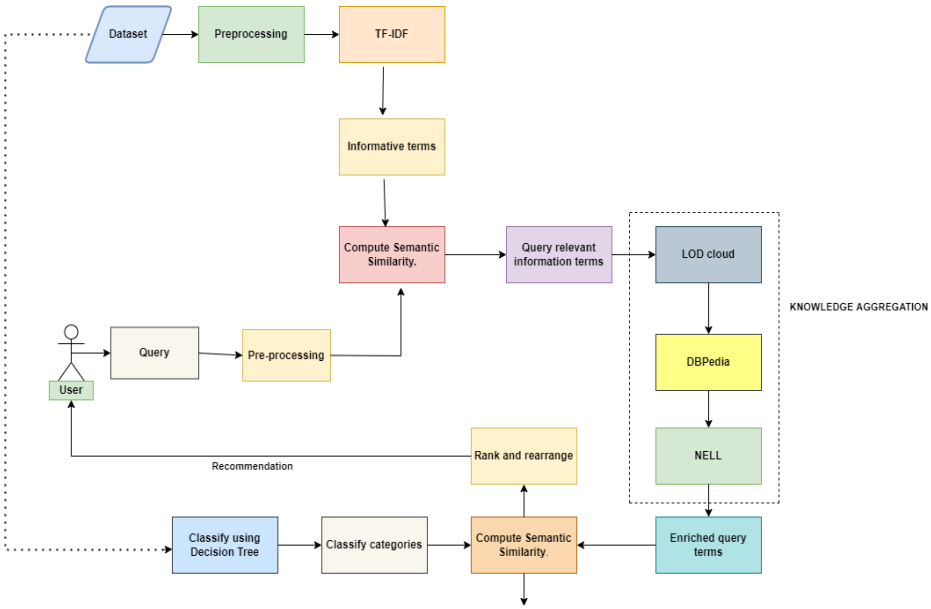

Fig. 1. Proposed system architecture of the knowledge-centric, semantically-driven document recommendation framework for recommending documents

Fig.1. portrays the architecture for the proposed KESDR model. It entails the query which is obtained from the user subjected to pre-processing. Pre-processing involves tokenization, lemmatization, removal of stop words and recognition of named entities. White space special character customized tokenizer was incorporated for tokenization. For lemmatization, the WordNet Lemmatizer was employed. RegEx based stop word removal customized algorithm was encompassed. GATE (General Architecture for Text Engineering) was utilized for named entity recognition. A categorical document dataset is taken into consideration for this framework which is subjected to pre-processing. The pre-processing involves elimination of redundant documents, after which the TF-IDF model is applied to the dataset.

TF-IDF (Term Frequency – Inverse Document Frequency) is an extensively used statistical method for retrieving information and Natural Language Processing. It reckons how significant a term is within a document relative to a set of documents. The TF-IDF score for the word t in document d, in the document set D is calculated using the equations (1),(2) and (3).

$$TF(t,d) = \log\left(1 + freq(t,d)\right) \quad (1)$$

$$IDF(t,d) = \log\left(\frac{N}{count(d \in D : t \in d)}\right) \quad (2)$$

$$TF - IDF = TF * IDF \quad (3)$$

The Semantic Similarity is computed using the identified most informative terms and pre-processed query terms . In this case, Normalized Google Distance is used to obtain the value of Semantic Similarity with a threshold of 0.5. A threshold of 0.5 is considered in order to increase the population of terms at the current scenario and to expand the initial informative term set in order to contribute feasible knowledge which is relevant to the query.The Normalized Google Distance is used to quantify Semantic Similarity obtained from the number of hits that Google Search Engine returns for a defined set of keywords. Keywords with kindred meanings in a natural language sense have closer Normalized Google Distance values. The Normalized Google Distance for two search terms x and y is calculated by equation (4).

$$NGD(x,y) = \frac{\max\{logf(x), logf(y)\} - logf(x,y)}{\log N - \min\{\log f(x), \log f(y)\}} \quad (4)$$

where N represents the total number of web pages searched multiplied by the average number of words per page. f(x) and f(y) denote the number of web pages with words x and y respectively. The number of web pages on which both x and y occur, is depicted by f(x,y).

Once the NGD is applied to the pre-processed query terms and the most informative terms, set of query relevant informative terms are formulated. Knowledge aggregation is followed. Query relevant informative terms are passed into the LOD (Linked Open Data) cloud via SPARQL endpoints. Further, the outcome of LOD cloud to the relevant query terms is passed to the DBpedia via SPARQL endpoints in order to aggregate the terms further and finally, all the entities yielded from the LOD cloud and DBpedia are sequentially passed into the NELL knowledge store through the NELL API. At the end of this phase, highly synthesized relevant yet undeviating knowledge is aggregated and the main reason for doing this is to eliminate or reduce the cognitive gap between the external knowledge from the World Wide Web as well as the intrinsic knowledge fed into the localized framework. Subsequently, the dataset is classified by using the features from the enriched query terms into several categories using the Decision tree classifier. Decision trees are favourable for accomplishing classification tasks. It creates a classifier structured like a tree, where the features of the dataset are represented by internal nodes, decision rules are represented by branches and final classes are denoted by leaf nodes.

Based on the focal classes identified from the categories, the Semantic Similarity is computed initially using the Normalized Google Distance with a threshold of 0.75. All the matching classes are taken into consideration and all the instances inside every class are further substituted into Semantic Similarity computation with the enriched query terms using the Simpson's Diversity Index and the Cosine similarity. The Simpson's Diversity Index is computed with a step deviation of 0.25 and Cosine Similarity is computed with a threshold of 0.75. The reason for using a higher threshold here is to narrow down the feasible solution set to a much more optimal solution set and it is recommended to the user. So ranking is done based on the increasing order of Cosine Similarity and it is yielded to the user. If the user is not gruntled, the process proceeds until there are no further clicks by the user.

4 Results and Discussions

The semantically enriched document recommendation KESDR framework has been implemented using Python 3.10.4 using Google's Colaboratory as a tool with Intel Core I7 processor having a GPU speed of 400-1350 MHz. A customized and benchmarked dataset having documents related to Pharmacology and Pharmaceuticals are used. The document corpus comprises of 24962 documents which are aggregated and it is ensured that each document is labelled and annotated. Atleast, one label is present for each document and it was ensured that the full coverage of documents were present and the dataset was balanced by adding another 24862 documents which don't strictly belong to pharmaceutical domain. They might be slightly deviated from the pharmaceutical domain and another 24174 documents were completely irrelevant to the pharmaceutical domain. The dataset was crawled using a customized crawler. For pre-processing, Python's NLTK (Natural Language Tool Kit) was used and for lemmatization, the WordNet 3.0 Lemmatizer was used. A white space special character and period punctuated tokenizer was used. For stop word removal, RegEx based stop word matching algorithm was incorporated. Thesaurus based NER (Named Entity Recognition) was achieved by collecting Pharmaceutical index terms.

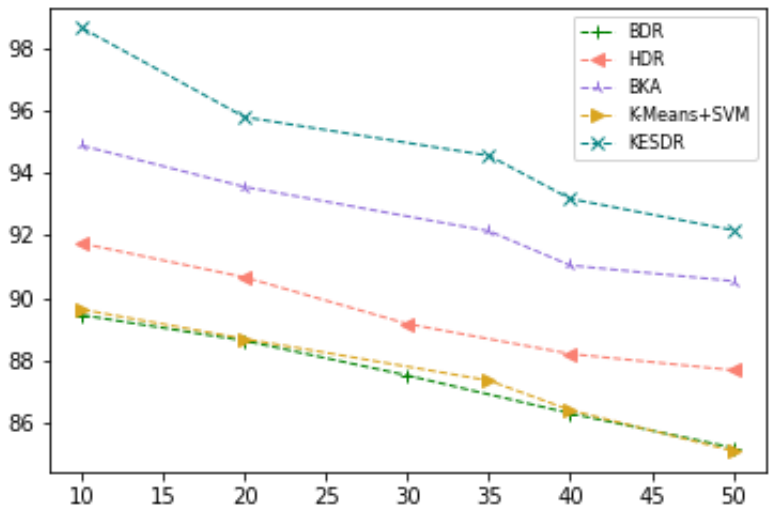

Fig. 2. Precision percentages of several approaches

From the Fig 2, it is clear that the proposed KESDR model performs better than the baseline models and occupies the highest position in the hierarchy, followed by the BKA model, then HDR model, then the K-Means+SVM hybrid model and finally the BDR model.
Precision, Recall, F-measure and Accuracy indicate the relevance of results furnished by the framework and the False Discovery Rate quantifies the number of False positives which are yielded by the KESDR framework. The key SDR was baselined with HDR, BDR and BKA so as to gauge the performance. Also, the combination of K-means clustering with SVM (Support Vector Machines) was also used in order to compare the performance of the proposed KESDR framework.

From Table 1., it is evident that the proposed BDR yields least values for all the performance measures with an extortionate value for FDR because although Bidirectional Encoder Representations from Transformer is applied for document retrieval which is integrated with BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), the entire learning is reliant solely on the dataset, i.e., the learning is dataset driven. Supplemental knowledge is not made inclusive into the framework. Although BIRCH is used, it helps in supporting NLP tasks, it increases the overall efficacy of BERT. However, due to the dearth of auxiliary knowledge and lack of strong relevance computation paradigms, the baselined BDR model has lagged to a significant extent. Although, HDR performs slightly better than the BDR model, it also yields much lesser precision, recall, accuracy and F-measure and much higher FDR when compared to the proposed KESDR model. The reason being, although HDR uses semantic and lexical matching, the matching takes place using key words alone. Addition of auxiliary knowledge is comparatively low and hence semantic deep neural network is used. The deep neural network learns from the dataset and causes overfitting. Therefore, the relevance is lost. But, if the deep neural network was semantically attenuated by first creating the knowledge graph or a knowledge model, then the performance would have been much better for the HDR. Also, the relevance computation mechanisms are not at all strong, being dependent on the deep learning paradigm. The BKA model has produced the immediate highest Precision, Accuracy, F-measure and Recall and the immediate lower FDR in contrast to the baseline model. The main reason behind good performance of the BKA model is it incorporates auxiliary knowledge with linked data and knowledge sources like DBpedia and Yahoo being used along with strong relevance computation mechanisms. However, still the strength of the relevance computation mechanism in the model can be made higher and the strength of the auxiliary knowledge can also be augmented by making it more heterogeneous. Although it is performing, still it can be improvised for better performance. The hybrid K-Means+SVM has also not yielded good results as SVM is a classifier and K-Means is a clustering algorithm. Even in this model, the auxiliary knowledge is absent, the classifier is not efficient and the relevance computation mechanism is also not good. The KESDR outperforms the baseline models mainly because of its heterogeneous knowledge aggregation by selecting knowledge from LOD cloud, DBpedia and NELL. Apart from this, a Decision tree classifier is used to classify the dataset where feature

selection is controlled and most importantly, in order to compute the relevance of results, the heterogeneous Semantic Similarity models are used. Relevance computation is strong owing to the inclusion of NGD. The inclusion of Cosine Similarity with NGD for computing Semantic Similarity along with the incorporation of SDI (Simpson's Diversity Index) also ensures that the proposed model has a strong relevance computation mechanism. Apart from that, the classifier is also strong which doesn't perform overfitting and knowledge leveraging from several heterogeneous knowledge sources – mainly, LOD cloud, DBpedia and NELL ensure a high knowledge density to the model, where a large number of entities are leveraged. Hence, this causes the proposed KESDR model to outstrip the baseline models.

Documents related to Pharmacology and Pharmaceuticals are used. The document corpus comprises of 24962 documents which are aggregated and it is ensured that each document is labelled and annotated. Atleast, one label is present for each document and it was ensured that the full coverage of documents were present and the dataset was balanced by adding another 24862 documents which don't strictly belong to pharmaceutical domain. They might be slightly deviated from the pharmaceutical domain and another 24174 documents were completely irrelevant to the pharmaceutical domain. The documents were crawled from several sources from the World Wide Web and they were spliced randomly in order to generate highly rich document corpus. Hence, a customized and annotated dataset was prepared. For the KESDR approach and the baseline models, an experimentation was held. A total of 4184 queries were addressed. Then, the ground truth was collected and validated based on ontologies and indexes belonging to pharmaceutical domain. If there were any deviants, the item was rejected ad if there was a correlation, the item was accepted. 50% deviants for rejected and 75% deviants for acceptance for the ground truth was followed.

5 Conclusion

In order to compute the Semantic Similarity, the proposed model uses the pre-processed query terms and the identified most informative terms using the Normalized Google Distance with a threshold of 0.5 to generate a set of query relevant informative terms. The query terms are enriched using the knowledge aggregation sources like DBpedia, LOD cloud and NELL, and further used to obtain the Semantic Similarity. Later, the dataset is classified by using the features from the enriched query terms into several categories using the Decision tree classifier. Based on the focal classes identified from the categories, the Semantic Similarity is calculated using the Normalized Google Distance with a threshold of 0.75. All the matching classes are taken into consideration and every instance inside every class is further substituted into Semantic Similarity estimation with the enhanced query terms using the Cosine Similarity and Simpson's Diversity Index. Based on the ascending order of Cosine Similarity values, ranking is done and given to the user. The KESDR model along with HDR, BDR, BKA and K-

Means+SVM models were evaluated based on the Precision, Accuracy, F-measure, Recall and FDR score. The proposed KESDR model gives the least FDR score of 0.0551 along with the highest F-measure of 95.42 % in comparison to other models. As a part of future work, we plan to enhance the proposed framework by further integrating it with nature inspired optimization algorithms, so that much more optimized and feasible results can be obtained.

References

1. Yilmaz, Z. A., Wang, S., Yang, W., Zhang, H., & Lin, J. (2019, November). Applying BERT to document retrieval with birch. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations (pp. 19-24).
2. Sakib, N., Ahmad, R. B., Ahsan, M., Based, M. A., Haruna, K., Haider, J., & Gurusamy, S. (2021). A hybrid personalized scientific paper recommendation approach integrating public contextual metadata. IEEE Access, 9, 83080-83091.
3. Guo, L., Cai, X., Hao, F., Mu, D., Fang, C., & Yang, L. (2017). Exploiting fine-grained co-authorship for personalized citation recommendation. IEEE Access, 5, 12714-12725.
4. Dai, T., Gao, T., Zhu, L., Cai, X., & Pan, S. (2018). Low-rank and sparse matrix factorization for scientific paper recommendation in heterogeneous network. IEEE Access, 6, 59015-59030.
5. Bai, X., Wang, M., Lee, I., Yang, Z., Kong, X., & Xia, F. (2019). Scientific paper recommendation: A survey. Ieee Access, 7, 9324-9339.
6. Kong, X., Mao, M., Wang, W., Liu, J., & Xu, B. (2018). VOPRec: Vector representation learning of papers with text information and structural identity for recommendation. IEEE Transactions on Emerging Topics in Computing, 9(1), 226-237.
7. Xia, F., Liu, H., Lee, I., & Cao, L. (2016). Scientific article recommendation: Exploiting common author relations and historical preferences. IEEE Transactions on Big Data, 2(2), 101-112.
8. Lee, Y. C., Yeom, J., Song, K., Ha, J., Lee, K., Yeo, J., & Kim, S. W. (2016, October). Recommendation of research papers in DBpia: A Hybrid approach exploiting content and collaborative data. In 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 002966-002971). IEEE.
9. Zhao, W., Wu, R., Dai, W., & Dai, Y. (2015, November). Research paper recommendation based on the knowledge gap. In 2015 IEEE International Conference on Data Mining Workshop (ICDMW) (pp. 373-380). IEEE.
10. Sharma, R., Gopalani, D., & Meena, Y. (2017, December). Concept-based approach for research paper recommendation. In International Conference on Pattern Recognition and Machine Intelligence (pp. 687-692). Springer, Cham.
11. Wu, Z. W., Chen, C. T., & Huang, S. H. (2021). Poisoning attacks against knowledge graph-based recommendation systems using deep reinforcement learning. Neural Computing and Applications, 1-19.
12. Kuzi, S., Zhang, M., Li, C., Bendersky, M., & Najork, M. (2020). Leveraging semantic and lexical matching to improve the recall of document retrieval systems: A hybrid approach. arXiv preprint arXiv:2010.01195.

13. Rospocher, M., Corcoglioniti, F., & Dragoni, M. (2019). Boosting document retrieval with knowledge extraction and linked data. Semantic Web, 10(4), 753-778.
14. Deepak, G., & Priyadarshini, J. S. (2018). Personalized and enhanced hybridized semantic algorithm for web image retrieval incorporating ontology classification, strategic query expansion, and content-based analysis. Computers & Electrical Engineering, 72, 14-25.
15. Kumar, A., Deepak, G., & Santhanavijayan, A. (2020, July). HeTOnto: a novel approach for conceptualization, modeling, visualization, and formalization of domain centric ontologies for heat transfer. In 2020 IEEE international conference on electronics, computing and communication technologies (CONECCT) (pp. 1-6). IEEE.
16. Deepak, G., Ahmed, A., & Skanda, B. (2019). An intelligent inventive system for personalised webpage recommendation based on ontology semantics. International Journal of Intelligent Systems Technologies and Applications, 18(1-2), 115-132.
17. Deepak, G., & Santhanavijayan, A. (2020). OntoBestFit: a best-fit occurrence estimation strategy for RDF driven faceted semantic search. Computer Communications, 160, 284-298.
18. Pushpa, C. N., Deepak, G., Thriveni, J., & Venugopal, K. R. (2015, December). Onto Collab: Strategic review oriented collaborative knowledge modeling using ontologies. In 2015 Seventh International Conference on Advanced Computing (ICoAC) (pp. 1-7). IEEE.
19. Deepak, G., & Kasaraneni, D. (2019). OntoCommerce: an ontology focused semantic framework for personalised product recommendation for user targeted e-commerce. International Journal of Computer Aided Engineering and Technology, 11(4-5), 449-466.
20. Deepak, G., & Santhanavijayan, A. (2022). UQSCM-RFD: a query–knowledge interfacing approach for diversified query recommendation in semantic search based on river flow dynamics and dynamic user interaction. Neural Computing and Applications, 34(1), 651-675.
21. Jain, S. (2021). Semantic intelligence: An overview. Web Semantics, 1-4.
22. Peis, E., del Castillo, J. M., & Delgado-López, J. A. (2008). Semantic recommender systems. analysis of the state of the topic. Hipertext. net, 6(2008), 1-5.