Bi-LDR: A Bi-Classification Model for Legal Document Recommendation using Knowledge Synthesis Approach

Gerard Deepak[1], Vamsi S[2], M Goutham Siddharth,[2] M Ushasree[2], Ramanathan N[3], PSai Kesava[2],Narasihma R[2] Santhanavijayan[2]

[1,2]Department of Computer Science and Engineering
[1]Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India
[2]National Institute of Technology, India
[3]State University of New York, Buffalo, United States of America
[1]gerard.deepak.cse.nitt@gmail.com

Abstract

Legal document recommendation is a requisite for specialized domains like socio, law and legal studies is a mandatory requirement as there are no existing specialized search engines which are semantically inclined. In this paper, alegal recommendationsystem Bi-LDR framework has been proposed which is a semantically driven model and based on bi-classification technique. Bi-LDR framework which is a framework for legal document recommendations using knowledge synthesis driven by both the user preferences and the existential knowledge. Datasets are subjected to LogisticRegression and Long Short-Term Memory (LSTM) classifiers. By using logistic regression, we can find the best fit model to describe the relationship between independent and dependent variables of a dataset. The dataset is also classified using LSTM, which is a better recurring neural network over the existing traditional neural network accounting for memory efficiency. In order to enrich the relevance computation, semantic similarities are computed using NormalizedCompression Distance (NCD). Entity similarity computations are done with that of the obtained individual terms from the user preferences by using Normalized Compression Distance, Shannon's Entropy and K-L divergence. An overall Accuracy and F-Measure of 97.62 and 97.61 with the lowest FDR of 0.04 has been achieved by the proposed framework.

Keywords-Knowledge Synthesis, Legal Document Recommendation, LSTM, Semantic Similarity, Web 3.0

1. INTRODUCTION

The World Wide Web is a vast worldwide repository of knowledge. Lots of Information is added to it daily. But in recent years the amount of information added to the World Wide Web has rapidly increased, it is known as the Information Explosion. But on the other side, there is a lack of Domain-Specific Information for specialized domains such as Legal, Medicine, Pharmaceutical, and other allied Domains. By considering all of these our ultimate goal is to formulate a Domain Specialized Document Recommendation Framework in order to ensure expert assistance and intelligence frameworks for a specific framework like medicine, legal, pharmaceutical, and other allied domains has to be built. Here Legal Document Recommendation Framework is proposed. The recommendation framework is made to automatically suggest particular pieces of information. It can assist in effectively retrieving information and making decisions by recommending content to users based on personal profiles. Some believe that a recommendation framework should include three components: background information, user preferences, and an algorithm to combine the information in order to provide the user with a useful recommendation. Legal professionals prepare advantageous and favorable arguments for a case by analyzing prior rulings. In Legal Domain as there are enormous amounts of rulings, Recommendation framework efficiently locate and provide referentially and/or

semantically pertinent judgments in the legal domain.

Motivation:Our main motivation is the lack of Specialized Domain-Specific experts and Intelligence systems for document recommendation specifically in the legal domain. Additionally, the majority of the current frameworks for document recommendation, including those for legal document recommendation, are based on non-semantic models that are not knowledge driven. Therefore, just based on learning or clustering paradigms is best suited for web 2.0, which is shallow, and if it needs to be made compatible with web 3.0 with a high degree of cohesiveness then the knowledge-centric paradigm that is semantically needed to be incorporated.

Contribution:A legal document recommendation framework that incorporates a bi-classifier for legal document recommendation which is semantically driven and knowledge-centric has been proposed. The legal knowledge pool receives input from the wiki data as well as the knowledge graph produced by Google's knowledge graph API. This legal knowledge pool is further enhanced by static domain ontologies that were created automatically using OntoCollab for the legal and law as a domain of choice. The dataset is fed to a logistic regression classifier as well as Long Short-Term Memory (LSTM). Semantic similarity measures which are incorporated are Normalized compression distance (NCD), Shannon's entropy, and KL divergence. Performance metrics such as Precision, Recall, Accuracy, and F-Measure are increased in the proposed framework when compared with the other baseline models.

Organization: The paper is organized in the following manner. Section 2 depicts theRelated Works. Proposed System Architecture is presented in section 3. Implementation is discussed in Section 4. ThePerformance Analysis and Results are shown in Section 5. Section 6 concludes the paper.

2. RELATED WORKS

Jenish Dhanani et al., [1] proposed a dictionary-based recommendation system for legal documents. The proposed method is an efficient and effective recommendation system of legal papers to find the most pertinent Indian rulings when a decision is issued. The suggested method solely took into account decisions made by the Indian Supreme Court, despite the fact that numerous other legislative bodies in India generate a wide range of legal material.Jenish Dhanani et al., [2] proposed another method for recommending legal documents using cluster-based pairwise similarity computing. The proposed method applies citation network clustering to create clusters on judgments that are referentially meaningful. It is possible that search space can be limited to the cluster itself rather than the complete collection by computing the similarity scores between pairs of each cluster that have strong significance amid all the judgments. KeetSugathadasa et al., [3] have introduced a method for retrieving legal documents that use both document vector embeddings, and deep learning. In this study, three innovative models were created and evaluated against a gold standard produced using internet legal domain resources. Their research has demonstrated the value of taking IR systems which are domain-specific into account, which cooperate in the development of the semantic web. Zhiqiang Zhang et al., [4] proposed a document retrieval technique based on graphs. The generic maximum common subgraph-based similarity calculating approach for graphs is presented, and experiments are carried out to demonstrate the improvement in results. However, as employing graphs increases complexity, new techniques are required for improvement. Jenish Dhanani et al., [5] proposed a practical and adaptable legal judgment recommendation. This method uses a pre-learned word embedding-based system for legal document recommendation to train the Doc2Vec embedding using a pre-learned word embedding that is particular to the Legal domain and contains the Legal semantic information. Bhaskar Mitra et al., [6] have proposed a model for document retrieval using Conformer-Kernel with query term independence. This model has many conformer layers which are memory-efficient and permit documents offline pre-encoding while indexing. Gaurav Kant Shankhdhar et al., [7] have introduced a legal semantic web recommendation system. This system uses the concept of the Semantic Web for proactive decision-making. This helps to prepare for similar cases beforehand which can guide for better judgments. ZichenGuo et al., [8] have introduced a content-based recommendation framework. The proposed framework targets producing recommendations that are related to the content of judicial cases. And it takes the information regarding the content using the combination of TF-IDF and LDA topic model,

ignoring usual methods like Collaborative Filtering Recommendations. HadasRaviv et al., [9] have proposed entity-based language document retrieval models that are persuaded using an entity linking tool. They showed that these language models are efficient for both query expansion, and cluster-based document retrieval, and their cons using performance comparison. FaezehEnsan et al., [10] have put forward a document retrieval model that uses semantic linking. In this model, the calculation of how much a question is related to a given document is done using semantic relatedness between the concepts. Experiments are conducted to prove that the performance of keyword-based systems will be improved by the proposed model. In [11-18] several Ontological Models and Semantic Frameworks have been depicted in support of the literature of the proposed work.

3. PROPOSED SYSTEM ARCHITECTURE

Following figure represents the blueprint for creating a Bi-LDR framework which is a framework for legal document recommendations using knowledge synthesis driven by both the user preferences and the existential knowledge. The user preferences in terms of queries and user clicks are subjected to preprocessing, and this preprocessing involves the processes such as tokenization, lemmatization, which are main processes including two other processes which are known as stop word removal (removing commonly occurring words) and named entity recognition in order to obtain the individual terms of the user preferences which will be further used in the framework. Simultaneously the knowledge synthesis takes place from heterogeneous pockets of knowledge and a knowledge stack is used namely the stack comprises of a directory called the Law next directory, E-book indexes with the indexes of the E-books which are belonging to the legal and law domain are taken and the indexes are crawled, and these indexes are stored in a knowledge stack and then the web indexes. The web indexes structurally represent the web usage data as well as the structural data of the web 3.0 which are pertaining and belong to legal and law as a prospective domain. These are used in the stack and this stack is imported into the legal and law domain and is subjected to preprocessing and preprocessing again involves tokenization, lemmatization, and NER, and from that, a taxonomy is formulated. The taxonomy is formulated based on the existing hierarchy as well as matching the terminal terms. To match the terminal terms, the concept similarity is made of and at least there should be one link established between the terminal nodes, and thereby a shallow taxonomy is formulated. Once the taxonomy is formulated, it is subsequently used in the further framework.

Subsequently, the dataset is subjected to the topic extraction on preprocessing and the dataset is the legal domain categorical dataset. These topics will be further submitted to Google's knowledge graph search API in order to yield the relevant knowledge graphs and knowledge subgraphs of the extracted topics potent to the legal domain. Since the knowledge graphs become insufficient, the terminal nodes and leaf nodes of the knowledge graph are submitted to the wiki data API in order to yield the hierarchical auxiliary knowledge. The data from the wiki data as well as the knowledge graph yielded from Google's knowledge search graph API are fed into the Legal knowledge pool and this legal knowledge pool is further enriched by static domain ontologies which were automatically generated using OntoCollab for legal and law as a domain of choice.The static domain ontology enriched with the legal knowledge pool and the formulated taxonomy from the knowledge stack is computed with normalized compression distance (NCD) which is a semantic similarity measure with a threshold of 0.5 in order to yield the matching terms. The reason for keeping the threshold at 0.5 is to ensure a large number of nodes are aggregated and matched. The matched terms yielded from the normalized compression distance i.e., the terms match from the ontology enriched legal knowledge pool and the taxonomy. These terms are fed as features to the logistic regression classifier in order to classify the existing dataset.

Logistic regression is one of the famous machine learning classifiers, it follows the principle of manual feature selection, therefore the matched terms are fed as the features to classify the dataset using the logistic regression classifier.According to the concept of machine learning, Logistic Regression is classified as the supervised learning classification algorithm. It predicts observations by classifying them into discrete categories.It is also useful in classifying observations into two or more classes. By using logistic regression, our goal is to find the best fit algorithm to describe the relationship between independent and dependent variables of a dataset.However, the dataset is also classified using LSTM. which is a better recurring

neural network over the existing traditional neural network accounting for memory efficiency. Since the general neural networks do not have a good hold over memorizing certain patterns, LSTM performs better. LSTM has multiple hidden layers (short-term memory) which held important and keywords, which will later be useful for the program to decide and suggest better recommendations for the user, while the irrelevant information is discarded after processing every cell. LSTM consists of 3 gates mainly, the first one is FORGET GATE and the second one is the INPUT GATE, and lastly OUTPUT GATE. The Forget gate decides whether the information is relevant for future purposes or not. The input gate constantly updates after deciding whether the existing memory can be discarded or not as the relevance of particular data/text might lower as we progress through the dataset/document respectively. Finally, the Output gate decides the next hidden state

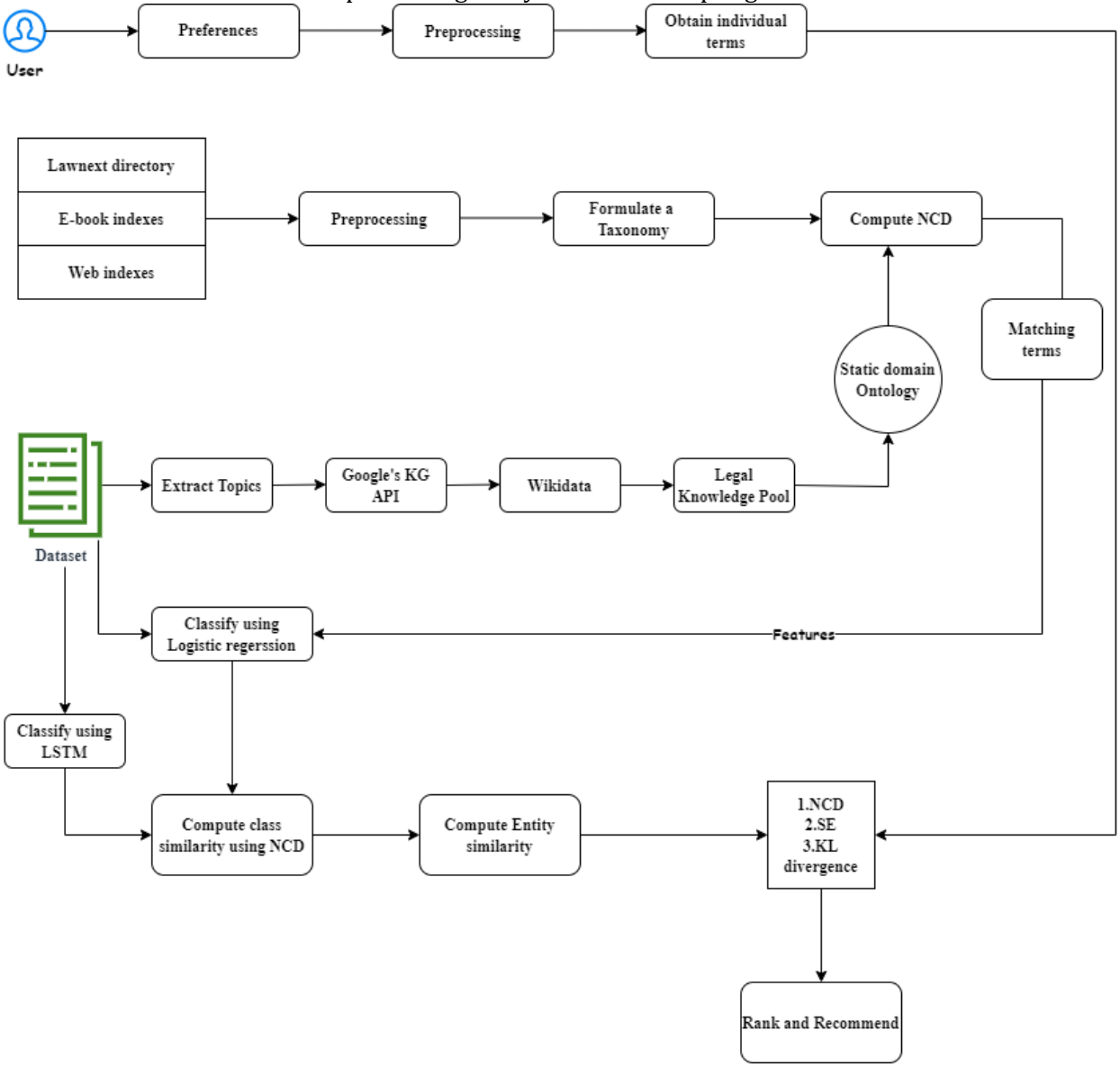

Fig.1. System Architecture Diagram

All the matching classes and the entities under each class are further used for entity similarity computation. Entity similarity computations are done with that of the obtained individual terms from the user preferences by using Normalized compression distance, Shannon's entropy and kl divergence. At this phase the NCD threshold is set as 0.5 because furthermore we are using stronger association measures and Shannon's entropy set deviation at 0.35 and KL divergence at 0.35,

relaxing the final aggregation as the NCD threshold is already set with a higher value of 0.5.Normalized compression distance is a measure of the similarity between two things. It could be two documents, two texts, two programs, two data points (similar type), etc..,. In NCD we use Kolmogorov's complexity K, which is the length of the shortest piece of program which produces the output. So, for two different strings x, y we first compute NID (Normalized information distance) as shown in Equation (1).

$$\square\square\square(\square,\square) = \frac{\square\square\square\{\square(\square/\square),\square(\square/\square)\}}{\square\square\square\{\square(\square),\square(\square)\}} \quad (1)$$

Equation (1) computes NID between x and y.Where K (x | y) is basically an algorithmic information of x when y is given as input.Now by approximating K by real-world compressors Z,

$$\square\square\square\square(\square,\square) = \frac{\square(\square,\square)-\square\square\square\{\square(\square),\square(\square)\}}{\square\square\square\{\square(\square),\square(\square)\}} \quad (2)$$

NCD is given by Equation (2).

$$\square(\square) = -\sum_{\square\square\square}^{N} P(x)\, l\square\square\ \square(\square) \quad (3)$$

Shannon's entropy is given by Equation (3).Where Σ represents the summation and logarithmic of base 2 is generally used and p(x) is defined as the probability of the occurrence of a certain event.

$$\square_{\square\square}(\square||\square) = -\sum_{\square\square\square}^{N} P(X)\ \square\square\square\left(\frac{\square(\square)}{\square(\square)}\right)_{N} \quad (4)$$

KL divergence is given by Equation (4).Where P(x) and Q(x) is the probability distribution over the event x. Simply KL divergence means the difference of logarithmic values of the probabilities P and Q in which the expectations are takes using the probabilities P Ultimately the matching instances are ranked in the ascending order of Normalized compression distances and are recommended to the user as presets and also the legal documents containing these presets are also suggested to the user, if the user is satisfied then there won't be any further clicks recorded and if the user is not satisfied the user clicks are again submitted as current user preferences and this process is repeated until the user is satisfied.

4. IMPLEMENTATION

The implementation of the proposed framework was carried out using python and google Collaboratory is used as the Integrated development environment (IDE). Python's NLTK (natural language toolkit library framework was used for conducting the preprocessing tasks. In order to carry out the implementation, an i7 processor with 16gb ram with a clock speed of 3.6GHz was used. The experimentation was conducted on three different datasets which were merged into a single dataset, namely the dataset of Legal documents [19], Legal documents entity recognition dataset [20], andLegal Citation Text Classification i.e., Legal industry-citations Text classification dataset [21]. The merging of datasets is done by converting them into csv and annotating them along with the keywords of the document. Apart from these 17424 Legal documents based on Indian context comprising Legal documents crawled from the world wide web. Indian context, Indian court of Law, Indian judgment documents, as well as Indian Lawsuit files for both criminal and civil cases and several textbooks of Law and the E-books were sliced randomly into individual documents which resulted in 17424 Legal documents and these documents were annotated based on the categories and were merged into the dataset which synthesized. The experimentations were conducted for this synthesized and annotated dataset and experimentations was done for 4882 queries whose ground truth has been collected from 1027 Law experts ranging from, Lawyers, judicial experts, law professors, law students of various universities and colleagues, and people who are pursuing masters and phd in Law were all taken into consideration as law experts and the ground truth for queries was generated. All the results yielded by the proposed model and the baseline model were validated for the exact same dataset and for the exact same number of queries and for the exact same ground truth, which was collected, and the results were tabulated. The proposed Bi-LDR Algorithm is depicted as Algorithm 1.

Algorithm 1: Proposed Bi-LDR Algorithm for Legal Document Recommendation

Input: Dataset Ds, Knowledge stack Ks, and user preferences Pu given by the user
Output: Ranked and highly relevant legal documents.

Step 1: Input user preferences Pu.

Step 2: Preprocessing of Pu is done and are broken into tokens by a) tokenization, b) lemmatization, c) stop word

d)removal, named entity recognition.

HashSet I□← insert (Terms).

Step 3: preprocessing of knowledge stack Ks, formulate a taxonomy

HashSet Tx ← extract (keywords)

Step 4: HashSetTp ← extract (topics from dataset)

Step 5: Mapped using Google's Knowledge graph search API and fed to Wikidata. (Enriched by static domain

ontologies)

Step 6: computing NCD for the obtained enriched HashSetTp and HashSetTx

Step 7: for each topic

p←Tp.iterator()

t←Tx.iterator()

$$NCDz(p,t) = \frac{z(p,t) - min\{Z(p), Z(t)\}}{max\{Z(p), Z(t)\}}$$

end for

Step 8: NCD threshold= 0.5

if(calc_ncd>ncd_threshold)

Matching HashSet H□ ←(p,t)

Step 9: Push matching terms H□ into the logistic regression classifier.

Dataset Ds is pushed into logistic regression classifier, put into a set R□

Step 10: Dataset Ds is pushed into LSTM classifier

Step 11: class similarity is computed with R□ using NCD and also entity similarity is computed

HashSet F□← insert (final terms)

f←F□.iterator();

Step 12: for each term in I□ and F

Calculate NCD

if(calc_ncd>ncd_threshold)

Pass to next similarity calculator

Calculate Shannon's entropy

sh_threshold=0.35

if (calc_sh>sh_threshold)

Pass on to next similarity calculator

Calculate KL divergence

kl_threshold=0.35

if(calc_kl>kl_threshold)

Push into resultant set R□

Step 13: Sort R□ in terms of rank and higher relevance and produce output to the user in sorted order

Step 14: if (user satisfied== true)

Terminate process

else

Retake user input and repeat process until user satisfied=true

5. RESULTS AND PERFORMANCE EVALUATION

The performance of the proposed Bi-LDR framework which is a knowledge-centric semantically driven module with a bi-classifier for legal document recommendation is computed by using Precision, Accuracy, Recall, F-Measure percentages, and False Discovery Rate (FDR) as potential metrics. Accuracy, Precision, Recall, and F-Measure percentages indicate the relevance of results, and the FDR computes the number of false positives which are captured by the model. For performance comparison, the proposed Bi-LDR framework is baselined with CRJC, LSW, and LJRWE as potential models. CRJC, LSW, and LJRWE all are legal document or judicial document recommendation frameworks.

Table 1. Performance Comparison of the proposed Bi-LDR with other approaches

Model	Average precision %	Average Recall %	Average Accuracy %	Average F-Measure %	FDR
CRJC [8]	90.24	92.73	91.48	91.46	0.10
LSW [7]	88.45	90.12	89.28	88.27	0.12
LJRWE [5]	92.25	94.69	93.47	93.45	0.08
Proposed Bi-LDR	**96.81**	**98.43**	**97.62**	**97.61**	**0.04**

The experimentations are performed on 1839 queries that have their ground truth collected. To collect the ground truth several law and judicial case history, and research papers, as well as knowledge basis, were collected and given to several law students in their third year and Master of Law Students of three distinct universities, and nearly 184 law students were involved in collecting and formalizing the ground truth. Specialized legal domain ground truth is used to benchmark the

results of the Bi-LDR and the baseline models. The baseline models are also queried for exactly the same no. of queries in the exact same environment as the proposed Bi-LDR model. From Table 1 it is seen that the proposed Bi-LDR yields an overall average precision of 96.81%, and an overall average recall of 98.43%, an average F-measure of 97.61% with an overall accuracy of 97.62%, and FDR of 0.04. CRJC furnishes an average precision of 90.24%, an average recall of 92.73%, an average accuracy of 91.48%, and an F-measure of 91.46% with an FDR of 0.1. LSW yields an average precision of 88.45%, an average recall of 90.12%, an average accuracy of 89.28%, and an F-measure of 88.27% with an FDR of 0.12. LJRWE yields an average precision of 92.25%, an average recall of 94.69%, an average accuracy of 93.47%, an average F-measure of 93.45% with an FDR of 0.08. By comparing the values from Table 1 it is evident that the Bi-LDR performance is ahead of CRJC, LSW, LJRWE with the highest precision, recall, accuracy, and lowest FDR value.

The Bi-LDR model is better than other baseline models because of its knowledge induction model where knowledge stack is used comprising of a law next directory which is a directory for legal resources then indexes of e-books as well as web indexes are used to form data taxonomy. Apart from this dataset knowledge is grown by using knowledge graphable, knowledge graph API and wiki data. Apart from this the static domain ontology for legal domain matching, the dataset is generated and aggregated. Most importantly the usage of two classifiers, the bi-classification framework of which one is feature controlled strong and stringent machine learning logistic regression classifier and other is a deep learning strong LSTM classifier. The class similarity and entity similarity are also computed using Normalized Google Distance at several phases and final recommendation is based on normalized compression distance, Shannon entropy and KL Divergence and the initial taxonomy formalization will happen using cosine similarity. So, usage of several semantic similarity models with varied threshold and differential step deviation measures ensures the proposed Bi-LDR performance is finer than the baseline models. The reason CRJC framework which is a judicial cases content-based recommendation model doesn't perform as expected is because it only depends upon TF-IDF for knowledge selection and LDA model with collaborative filtering which further depends on rating but rating computation metrics are to be formulated and item relevance metrics where in this combination requires every entity for the domain variated so rating cannot be effective. Therefore, the CRJC framework which is a content-driven model for judicial case recommendation does not perform as expected. The reason LSW model doesn't yield efficacious results compared the proposed model is because it is based on the phenomenon of using RDF to generate a semantic web with static ontology which is a legal ontology environment that is advocated and once it is advocated for similar cases it is used directly for making decisions where in the relevance computation mechanism is not very strong and since the knowledge which is used is readily available, the behavior is based on the static knowledge of the model and so the decision making is not instant because of no learning paradigm. Therefore, the LSW model doesn't perform as expected. The reason why the LJRWE model also does not perform as expected is because it is using Doc2Vec learning using vector space semantics and apart from this pre learned word embeddings are used. The pre learned word embeddings are quite light weighted in nature and they cannot be compared to auxiliary knowledge generation. They can only account for a tenth of the auxiliary knowledge which is bound and fused into the model. And also, they do not have very strong relevance computation models of the framework for final recommendations. As a result, the LJRWE model does not perform as expected.

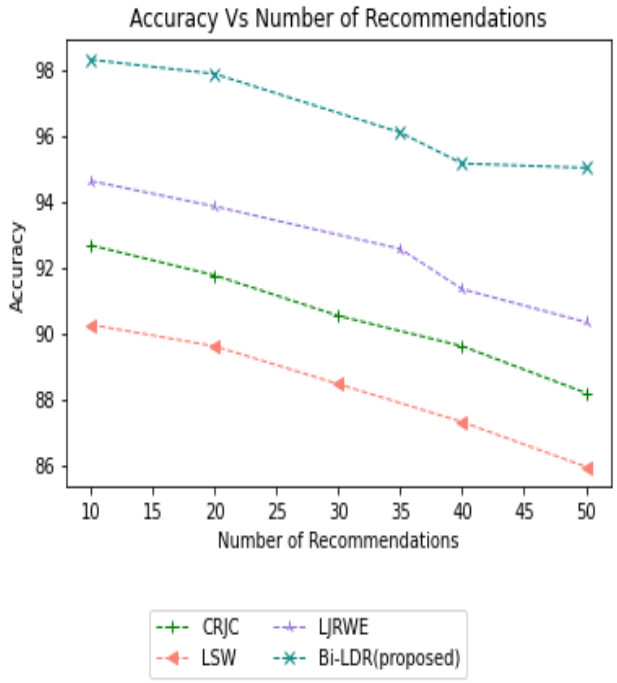

Fig.2. Precision % vs Number of Recommendations

Figure 2 depicts the line graph which is the Precision Percentage vs Number of recommendations distribution among the proposed Bi-LDR approach and other baseline models. It is indicative that the top position of hierarchy is occupied by the proposed Bi-LDR model, which is next followed by LJRWE and then followed by CRJC and the lowermost hierarchy is occupied by the LSW framework. Using several semantic similarity models, Bi classification, and being knowledge induction model and using knowledge graphable, knowledge graph API, and wiki data makes the proposed Bi-LDR model perform better than baseline models and the highest position of the hierarchy. LJRWE is behind the proposed Bi-LDR model because it uses pre-learned word embeddings which are lightweight and do not have strong relevance computation. The CRJC model does not perform well because it depends on the rating which is not effective, and its dependency only on TF-IDF for knowledge selection. These criteria make the CRJC model stand behind the proposed Bi-LDR model. The LSW model advocates a legal ontology environment for making decisions directly. But the relevance mechanism is not strong here and there is no learning paradigm that makes the decision-making non-instant. These reasons make the LSW model stand at least in the position of hierarchy.

6. CONCLUSIONS

In this paper, A Bi-Classification Model for Legal Document Recommendation (Bi-LDR) framework has been proposed for Legal Document Recommendation which uses Knowledge Synthesis Approach. In order to classify the dataset, Logistic Regression classifier was employed integrated with Long Short-Term Memory (LSTM)Classifer at different instances.. Semantic Similarities were computed using Normalized compression distance, Shannon's entropy, and kl divergence techniques with differential thresholds. The experimentations are conducted for 1839 queries. We used the matching classes and the entities under each class for entity similarity computation. Then we rank the matching instances using NCD and are recommended to the user as presets as well as the legal documents that include these presets. If the user is satisfied, then no additional clicks will be logged, but if the user is not, then the user clicks are again submitted as current user preferences and this process is repeated until the user is satisfied. The overall precision yielded by the framework is 96.81%, and yielded recall is 98.43%, and with an accuracy of 97.62%, and yielded F-measure is 97.61, with the lowest FDR of 0.04 which outperformed the corresponding values of remaining baseline models.

REFERENCES

Dhanani, J., Mehta, R., & Rana, D. P. (2021). Legal document recommendation system: a dictionary-based approach. International Journal of Web Information Systems.

Dhanani, J., Mehta, R., & Rana, D. (2021). Legal document recommendation system: A cluster based pairwise similarity computation. Journal of Intelligent & Fuzzy Systems, 41(5), 5497-5509.

Sugathadasa, K., Ayesha, B., Silva, N. D., Perera, A. S., Jayawardana, V., Lakmal, D., &Perera, M. (2018, July). Legal document retrieval using document vector embeddings and deep learning. In Science and information conference (pp. 160-175). Springer, Cham.

Zhang, Z., Wang, L., Xie, X., & Pan, H. (2018, May). A graph-based document retrieval method. In 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD)) (pp. 426-432). IEEE.

Dhanani, J., Mehta, R., & Rana, D. (2022). Effective and scalable legal judgment recommendation using pre-learned word embedding. Complex & Intelligent Systems, 1-15.

Mitra, B., Hofstatter, S., Zamani, H., &Craswell, N. (2020). Conformer-kernel with query term independence for document retrieval. arXiv preprint arXiv:2007.10434.

Kant, G., Singh, V. K., Darbari, M., Yagyasen, D., & Shukla, P. (2014). Legal semantic web-a recommendation system.

International Journal of Applied Information Systems (IJAIS), 7.

Guo, Z., He, T., Qin, Z., Xie, Z., & Liu, J. (2019, September). A content-based recommendation framework for judicial cases. In International Conference of Pioneering Computer Scientists, Engineers and Educators (pp. 76-88). Springer, Singapore.

Raviv, H., Kurland, O., & Carmel, D. (2016, July). Document retrieval using entity-based language models. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (pp. 65-74).

Ensan, F., & Bagheri, E. (2017, February). Document retrieval model through semantic linking. In Proceedings of the tenth ACM international conference on web search and data mining (pp. 181-190).

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.

Deepak, G., & Santhanavijayan, A. (2020). OntoBestFit: a best-fit occurrence estimation strategy for RDF driven faceted semantic search. Computer Communications, 160, 284-298.

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.

Santhanavijayan, A., Naresh Kumar, D., & Deepak, G. (2021). A semantic-aware strategy for automatic speech recognition incorporating deep learning models. In Intelligent system design (pp. 247-254). Springer, Singapore.

LeenaGiri, G., Deepak, G., Manjula, S. H., &Venugopal, K. R. (2018). OntoYield: a semantic approach for context-based ontology recommendation based on structure preservation. In Proceedings of International Conference on Computational Intelligence and Data Engineering (pp. 265-275). Springer, Singapore.

Roopak, N., and Gerard Deepak. "OntoKnowNHS: ontology driven knowledge centric novel hybridised semantic scheme for image recommendation using knowledge graph." In Iberoamerican Knowledge Graphs and Semantic Web Conference, pp. 138-152. Springer, Cham, 2021.

Rithish, H., Deepak, G., & Santhanavijayan, A. (2021, January). Automated assessment of question quality on online community forums. In International Conference on Digital Technologies and Applications (pp. 791-800). Springer, Cham.

Yethindra, D. N., & Deepak, G. (2021, September). A semantic approach for fashion recommendation using logistic regression and ontologies. In 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES) (pp. 1-6). IEEE.

Legal Documents Dataset "https://paperswithcode.com/dataset/dataset-of-legal-documents"

(2021). Legal Documents Entity Recognition Dataset [Dataset]. https://paperswithcode.com/dataset/legal-documents-entity-recognition

Shivam Bansal (2021). Legal Citation Text Classification [Dataset]. https://www.kaggle.com/shivamb/legal-citation-text-classification