Information Retrieval in Conjunction With Deep Learning

Chapter · September 2020

DOI: 10.4018/978-1-5225-9643-1.ch014

3 authors:

Anu Bajaj
Machine Intelligence Research Labs (MIR Labs)
17 PUBLICATIONS **72** CITATIONS
SEE PROFILE

Tamanna Sharma
Guru Jambheshwar University of Science & Technology
11 PUBLICATIONS **6** CITATIONS
SEE PROFILE

Om PRAKASH Sangwan
Guru Jambheshwar University of Science & Technology
78 PUBLICATIONS **539** CITATIONS
SEE PROFILE

Information Retrieval in Conjunction With Deep Learning

Anu Bajaj
Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology, Hisar, India

Tamanna Sharma
Department of Computer Science and Technology, Guru Jambheshwar University of Science and Technology, Hisar, India

Om Prakash Sangwan
Department of Computer Science and Technology, Guru Jambheshwar University of Science and Technology, Hisar, India

ABSTRACT
Information is second level of abstraction after data and before knowledge. Information retrieval helps in fill up the gap between information and knowledge by storing, organizing, representing, maintaining and dissemination of information. For example, user's query to access information from a huge unstructured database. Manual information retrieval leads to underutilization of resources and take long time to process while machine learning techniques are implication of statistical models, which are flexible, adaptable and fast to learn. Deep learning is the extension of machine learning with hierarchical levels of learning which make it suitable for complex tasks. Deep learning can be the best choice for information retrieval as it has numerous resources of information and large datasets for computation. In this chapter, we are discussing applications of information retrieval with deep learning e.g., web search by reducing the noise and collecting precise results, trend detection in social media analytics, anomaly detection in music datasets and image retrieval etc.

Keywords: Information retrieval, deep learning, convolution neural networks, image retrieval, text retrieval, recurrent neural networks, autoencoder, deep neural network, long short term memory

Introduction

We come across with huge amount of data day by day, which is mainly because of the social media, web and mobile applications usage, e.g., 15 GB of data is generated by Facebook alone (Kanimozhi & Padmini, 2018). This exponentially growing unstructured data in the form of web logs, data records, and sensor data etc. need to be converted into useful information. The information is what we acquire from the unconstrained data to fill the knowledge gap. For example, we want to buy some product then we need to resort to the customers review about the particular product. On positive response of product we would be likely to purchase the product else not. This is just a small example why information retrieval is important. The archival of the inscribed information may be tracked from 3000 BC where the Sumerians deposited clay tablets with cuneiform inscriptions (Singhal, 2001). For proficient use of information even they also projected special classification for identification of each tablet and its contents. Hence, the information retrieval (IR) is the process of archiving, organizing, maintaining the information collected from a huge database and disseminating the same to fill the user's needs. In other words, IR system reads the user's query and look out for information in the documents (database and knowledge base) for image, text, sound, and sensing data etc. and this retrieved information is responded back to the

users (Guan & Zhang, 2008). The retrieved documents are ranked with their estimate of importance of document for a particular query. Intermediate stages are indexing, filtering, searching matching and ranking of the documents. In indexing the documents are indexed using signature files or inverted indices etc. and the filters remove all the stop words white spaces etc., finally the query in searched by using any brute force search, and linear search (Kanimozhi & Padmini, 2018), and the matched documents are ranked based on their similarity with the query. The user is responded back with the top ranked documents. Several models have been proposed for this purpose (Singhal, 2001) as shown in Table 1.

Table 1 IR Models

IR Models	Description
Boolean Model	In this model, the query is represented by Boolean expression of terms and the terms are connected with Boolean operators.
Vector Space Model	In this model, the word and phrases are known as terms and these terms are represented in form of vectors.
Probabilistic Model	It assesses the probability of significance to the query. The documents are ordered by decreasing probability of their significance known as probability ranking principle.
Inference Network Model	The documents are modeled using the inference process in the inference networks. The documents are ranked according to the term strength.

It is different from the Database Management System (DBMS) in the sense that it is probabilistic (unstructured data) in nature while DBMS (structured data in tables) is deterministic. This is because the IR system searches the documents for the keywords provided by the query and in response give all the significant documents containing the required information. On the other hand, DBMS respond to exact match of the query and give the very specific response. It has started its journey from the library management which expands to office automation, knowledge extraction, multimedia management, medical information management, etc. applications. Due to the advent of internet, the data is growing exponentially day by day which needs to be handled efficiently. Hence some sort of automation is required to retrieve the information from such a large archive. Researchers started experimenting IR with machine learning techniques which is discussed in the section 3. The next section 4 explains why the deep learning techniques are more prominent than the machine learning techniques. Finally, we have concluded this chapter in the last section 5.

Information Retrieval with Machine Learning

IR becomes more efficient and effective with the help of machine learning algorithms. Figure 1 show the complete IR process. The machine learning algorithms are classified into two parts: supervised and unsupervised learning (Alzubi, Nayyar, & Kumar, 2018) as shown in Table 2.

Figure 1. Information Retrieval Architecture

IR is used in various domains like digital libraries, medical diagnostics, image retrieval, chemo-informatics, music systems, question answering, social media analytics etc. In this section we are providing the literature work done in the domain of IR using machine learning approaches.

Table 2 Machine Learning Approaches

Machine Learning	Methods	Description
Supervised Learning	Multiple Regression Analysis	It is a statistical approach that determines relation amongst independent variables and dependent variables.
	k-Nearest Neighbor (kNN)	It classifies the object surrounded by its k (integer) nearest neighbor using the majority rule.
	Naive Bayes	It is a probabilistic method which uses Bayes formula for predicting membership with the assumption that the features are independent.
	Random forest	It is used to classify the objects based on the majority voting rules and ensemble of the multiple decision trees.
	Neural network and deep learning	Neural network (NN) has layers of connected neurons which learns from with input layer, hidden layers (multiple in case of deep learning) and output layers.
	Support vector machine	It is a technique which draws input to high-dimensional space by constructing a set of hyperplanes through some non-linear mapping.
Unsupervised learning	k-means Clustering	It is a classification approach to classify observations into k clusters by minimizing the total intra-cluster variance.
	Hierarchical Clustering	It generates a hierarchical clusters for classification by agglomerative (merge up smaller clusters) clustering or divisive (split larger clusters) clustering.
	Principal Component Analysis	It is a technique for transformation of a set of interrelated features to principal components by using the orthogonal procedure.
	Independent Component Analysis	It is a statistical technique to separate the multivariable output from statistical independent additive components.

Simple query asked by the user is no longer provide appropriate results. Hence, the query needs to be modified using the similar terms for better results. The authors (Kim, Seo, & Croft, 2011) have proposed a Boolean query expansion method and exploited the decision trees for producing the Boolean queries by learning through pseudo labelled documents and ranking is done on the basis of query quality predictors. A query expansion method (Diaz, Mitra, & Craswell, 2016) was developed on the basis of local trained word embedding that captured the nuances of topic specific language. They suggested to use large topically-unconstrained corpora instead of topically

constrained corpora. Support vector machine (Drucker, Shahrary, & Gibbon, 2002) has been used to provide relevant feedback for the queries in which we have less number of retrieved documents to enhance the response.

Medical diagnostic systems have also started using machine learning techniques for decision making in emergency context e.g. for prioritization of requests. The authors (Pollettini, Pessotti, Filho, Ruiz, & Junior, 2015) have used the conceptual IR i.e. the textual processing of the clinical data e.g., diagnosis, chief complaint etc. to discover the similar cases in the databases. The patients' appropriate destination the priority of the request was automatically assigned by using top k cases and voting system. The researchers have found that the use of decision tree with random forest gave good results for the classification. They also found that the semantic approach is better and faster than the text mining approaches. In the study (Song, He, Hu, & He, 2015) two re-ranking method for medical decision support systems was developed. Content based image retrieval (Sinha & Kangarloo, 2002) was proposed with principal component analysis that reduced the search dimensionality using a set of prototype images. The match is found by producing the projection vector of query image and comparing to that of the database images.

Text retrieval using the Naïve Bayes method have been proposed by (Lewis, 1998) for independent text classification. The authors (Kumar, Ye, & Doermann, 2012) have applied learning techniques on the document images for retrieval and classification. The images have different patch code words which have been used for their retrieval. The labelled images are recursively partitioned into horizontal and vertical partitions and a histogram of patches of each partition is then calculated which gave very high precision and recall when trained with random forest classifier. Image retrieval was done by (Fu & Qui, 2012) using the random forest machine learning technique in which they used visual features to divide the tree nodes and image labels for supervising the division so that the similar semantic images may get located to same tree node. Semantic neighbor set of the querying image is found first and then the ranking is done with the help of semantic similarity measurement amongst the querying image and the image in its semantic neighbor set.

The authors (Maarek, Berry, & Kaiser, 1991) have constructed the software libraries by first aggregating the attributes from the natural language documents using indexing technique and then the hierarchy for browsing was automatically generated by clustering of these documents on the basis of the extracted attributes. Probabilistic model has been used to examine the lexical information for software clustering (Corazza, Di Martino, Maggio, & Scanniello, 2011). Information was retrieved from six different vocabularies, i.e., class, method, comments, and source code statements by applying automatic weight mechanism to check the contribution of each vocabulary. These weights are optimized by using iterative expectation-maximization algorithm. Vector space model has been applied for computing the similarity among classes which is useful in making the software clusters.

The authors (Sathya, Jayanthi, & Basker, 2011) projected intelligent cluster search engine by using k-means clustering because the traditional search engines used ranking algorithm which did not properly classify the web pages, therefore, did not provide the relevant web pages. The IR technique had been improved by comparing co-occurring terms and documents clustering. In the study (Yang & Lee, 2006) have used the machine learning technique for information searching. A

navigational map has been established for the World Wide Web by using the proposed learning technique. The web pages mapping was done with self-organization map and their relation with thematic keywords were established by the feature maps. These maps were then used to develop a structure for assisting the users in IR. Learning to rank method have been used for combining translation resources to cross-lingual IR (CLIR). CLIR search information in one language while the query is produced in other language. Monolingual IR features (Azarbonyad, Shakery, & Faili, 2019) have been used to map to the cross lingual IR by using translation information from different translation resources. It is showed that learning to rank techniques have improved the CLIR performance.

Another application of IR is community question answering where some people ask questions and anybody who has knowledge of the subject may answer the query. We have microblog platforms also where one can find answers to some specific event, subject or task. These knowledge bases are becoming larger day by day. For example twitter data is real time and social or geographical impact, therefore, valuable historic information is context rich and up to date also. Hence, there is need for some automation learning mechanism to answer the complex queries. Learning to rank (LTR) mechanism (Herrera, Poblete, & Parra, 2018) has been applied to re-rank the significant answers on the top locations. IR is also used for trend detection in social media analytics. The authors (Hore & Bhattacharaya, 2019) have used machine intelligence for classifying the opinions or suggestions of people in India on the hot topic of "save the girl child" under the slogan of "Beti Bachao Beti Padhao".

Chemo informatics extract, process and extrapolate the meaningful data from chemical structures. Due to huge amount of drug data it is impossible for drug designers to extract specific drug properties for specific drug designing. In the study (Lo, Rensi, Torng, & Altman, 2018) have used the machine learning technique and QSAR analysis for chemical fingerprints and similarity analysis. Several other works based on chemo informatics in combination with machine learning algorithms are there in the literature e.g. regression analysis (Eriksson et al., 2003), naïve base classifiers (Chen, Sheridan, Hornak, & Voigt, 2012), k nearest neighbors (Lo et al., 2015) etc.

Unsupervised learning was applied (Lu, Wu, Lu, & Lerch, 2016) for anomaly detection in music datasets. Music IR is the combination of various fields, i.e., electrical engineering, psychology, musicology and computer science and engineering. They have retrieved the music data by using statistical model and common features were extracted. The appropriate labelling of the anomalous music genre dataset is done by labelling them as corrupted, distorted or mislabeled clips without any requirement of training data. Probability density function has been used (Hansen et al., 2007) for clean the broadcast and CD collections on the basis of segmentation problems, missing and wrong meta-data. They found the relation between the music features and meta-data and also spotted the unlikely music features by training conditional and unconditional densities.

Information Retrieval in Conjunction with Deep Learning

As we have discussed in introduction section that the data is growing exponentially and to learn from such a huge database is cumbersome task. Hence, traditional machine learning algorithms are not suitable for big data which have volume (scalable data), velocity (data growth), variety

(diverse sources) and veracity (uncertain data). Therefore, deep learning approaches came into practice which vanishes the effects of gradients, so, more suitable to use with raw high-dimensional data (Hinton, Osindero, & The, 2006). Deep learning found its origin from Neural Networks in which feed forward NN combined with many hidden layers i.e. it learns from low features to high level features e.g. it can learn raw pixel input as color information and in the next layer it may go up to edge of the object using the previous layer information. Deep learning automatically select raw, heterogonous, high dimensional data, without manual selection. As a result, machine learning has shifted towards deep learning due to data driven and computational power driven activities. The major deep learning techniques are presented in Table 3. Here, we have reviewed the current status of deep learning techniques used in the IR studies and provide a brief summary of their advantages and future perspectives.

Table 3. Deep Learning Techniques

Deep Learning Techniques	Description
Deep Neural Networks (DNN)	It comprises of multiple hidden layers with each layer having hundreds of nonlinear processing components. It takes large number of input features and extract features automatically from various hierarchical stages by using neurons of different layers.
Convolution Neural Networks (CNN)	It consists of several convolutional layers (having a set of filters with small receptive fields and learnable parameters) and subsampling layers (reduces the feature map size). The feature maps are joined to get fully connected layers to get a final output.
Recurrent Neural Networks (RNN)	It takes sequential data as input and forms a directed cycle by allowing the connections among neurons in the same hidden layers. It may diminish the vanishing gradient problem by using long short term memory (LSTM).
Autoencoder	It has encoder NN which converts the information from input layer to some hidden layers and then fed to decoder NN which rebuild its own inputs using lesser number of hidden layers. Hence its basic purpose is dimensionality reduction.

Due to general constraint of available human labelled data for a huge database, the authors (Palangi et al., 2016) developed a model for sentence embedding using RNN with LSTM cells which extracted the information from each word of the sentence and embedded it into semantic vector. They have trained the model in weakly supervised manner i.e. with long term memory the model kept on accumulating the information entered by the user till the last word of the sentence. The hidden layer provided the semantic representation of the whole sentence by automatically

discarding unimportant words and keeping the important ones. It showed that the semantic vector evolved over time and took only important data from any new input. Also these detected words automatically activated the cells of RNN-LSTM to activate the cells of the similar topic. These automatic word detection and topic allocation supported LSTM-RNN for document retrieval. The proposed method outperformed the paragraph vector method. Deep learning vectors has been developed (Almasri, Berrut, & Chevallet, 2016) to train high quality vector representations for huge amount of unstructured terms. It also get a big number of term relationships for query expansion technique. The method was empirically analyzed and the results were proved to be better than other expansion models.

Medical researchers have started using deep learning algorithms for diagnostics of disease. For example, deep CNN showed great potential in classification of skin lesions by using input as pixels and disease labels of the images. It has been showed that the deep learning can classify the most common and most deadly skin cancer with the competence level equivalent to dermatologists (Esteva et al., 2016). The authors (Gulshan, Peng, Coram, Stumpe, & Wu, 2016) developed a deep NN for automatic finding of diabetic retinopathy and macular edema and classification using retinal fungus images.

Deep NN for software bug localization (Lam, Nguyen, Nguyen, and Nguyen, 2015) has been proposed to automatically locate the documents containing bug. The major challenge in software bug localization is the lexical mismatch. They have used vector space model for textual similarity and DNN for learning the lexical mismatch in bug reports and source files.

The authors (Zhuang, Liu, Li, Shen, & Reid, 2017) developed a weakly supervised deep learning approach to automatic label the web image information. They have used two strategies random grouping, which piled various images in one training instance to increase the labelling accuracy at group level and attention. It surpassed the noisy signals from the incorrectly labeled images. The authors (Jaderberg, Simonyan, Vedaldi, & Zisserman, 2016) developed end to end text reading pipeline-detecting and recognizing text from natural scene images. CNN was used for recognition and region proposal mechanism for detection purposes. The networks were produced wholly by a synthetic text generation engine without any human labelling. The proposed system worked well for both image retrieval and text spotting comparing to other methods for all standard datasets.

In community question answering the difficulty lies in semantic matching between question answer pairs and modeling of contextual factors. An attentive deep NN was proposed (Xiang, Chen, Wang, & Qin, 2017) to learn the deterministic facts for answering the question. The pros of this method is that it can support various input formats by using CNN, attentive based LSTM ad conditional random fields. The authors (Hoogeveen, Bennett, Li, Verspoor, & Baldwin, 2018) proposed the learning algorithms for detection of the misflagged duplicate questions i.e. the questions which were mistakenly flagged as duplicate from the archived ones but their meanings are actually different. The text features alone cannot model the duplicates, therefore, meta-information of user posting the question and the questions themselves is required. The researchers used various machine learning and deep learning approaches for classification or detection of duplicates and found that random forest worked well than other algorithms.

Deep learning outperformed machine learning techniques in designing chemically valid and synthetically accessible molecules with appropriate characteristics for drug discovery (Blaschke, Olivecrona, Engkvist, Bajorath, & Chen, 2018). In the study (Gómez-Bombarelli et al., 2018) developed a technique to automate molecular design by using auto-encoders. The network was fed with thousands of structures to derive a set of coupled functions. By using vector decoding, perturbing known chemical structures and interpolation among chemical structures, the DNN succeeded to develop new molecules with drug like properties.

In music genre system, the defined music features have been used for retrieving information but Deep Belief Networks (DBN) were used (Lee, Largman, Pham, & Ng, 2009) to scale spectrograms which automatically learned the features and outperformed the traditional audio features in music genre recognition. The learned feature representation from the unlabeled data showed very good performance for multiple music classification tasks. The authors (Hamel & Eck, 2010) have used the DBN on Discrete Fourier Transform (DFT) of audio data for the automatic feature extraction to solve the task of genre recognition and auto-tagging. Computational models (Pati, Gururani, & Lerch, 2018) has been built for automatically assessing the music performance, i.e., rating the performance on several criteria like musicality, etc. by analyzing the audio recording. The researchers used deep NN for feature learning instead of hand crafted feature. The results showed that supervised feature learning technique better characterize the music performance than the baseline approaches. We have summarized the purpose of IR along with the machine learning techniques in the Table 4.

Table 4. Summary of Machine learning techniques in association with IR based Applications

Authors	Purpose for information retrieval	Machine Learning/Deep Learning Technique Used	IR Model Used
Lam et al., 2015	Addressed the problem of lexical mismatch	Deep Neural Network	Vector Space Model
Herrera et al., 2018	Learning to rank framework was used for aggregation of microblog information and QA task	MART, RankNet, RankBoost, LambdaMart	tf-idf Model
Kumar et al., 2012	Efficient query generation	Decision Trees	-
Kim et al., 2011	Comparison between local and global embedding	Word2Vec, GloVe	Local Latent Semantic Analysis
Diaz et al., 2016	Query expansion using deep learning outperformed the language model based query expansion.	Word2Vec	Dirichlet Model
Almasri et al., 2016	Try to overcome the language barriers.	Learning to Rank	Cross-Language Information Retrieval

Azarbonyad et al., 2019	Used for audio classification	Convolutional Deep Belief Network	-
Lee et al., 2009	Auto tagging of musical data	Radial Basis Model	-
Zhuang et al., 2017	Text recognition task from the images	Convolutional Neural Network	-
Jaderberg et al., 2016	Automated music performance assessment	Deep Neural Networks	-
Xiang et al., 2017	Classifying misflagged questions from community question answering database	Convolutional Neural Network	tf-idf Model

This table represents machine learning techniques in association with IR. It summarizes potential of the machine learning algorithms at different stages of IR based applications like chemo-informatics, music genre system etc.

Conclusion

In this chapter, we discussed about how the IR process can be enhanced by using machine learning approaches. The learning algorithms help in better classification and ranking of the relevant documents. However, due to drastic growth of data, it is difficult to retrieve information from such a huge database. Researchers started using deep learning techniques for handling big data and automatic labelling of raw and heterogeneous data. These learning algorithms shown good performance for various applications of IR e.g. image retrieval, music information systems, chemo-informatics, medical management systems etc. We can conclude that deep learning algorithms have great potential in solving IR problems.

References

Almasri, M., Berrut, C., Chevallet, J.P. "A Comparison of Deep Learning Based Query Expansion with Pseudo relevance Feedback and Mutual Information." In: *European Conference on Information Retrieval,* 2016, pp. 709–715.

Alzubi, J., Nayyar, A., and Kumar, A. "Machine learning from theory to algorithms: an overview." In *Journal of Physics: Conference Series,* 1142(1), IOP Publishing, 2018, p. 012012.

Azarbonyad, H., Shakery, A. and Faili, H., "A Learning to Rank Approach for Cross-Language Information Retrieval Exploiting Multiple Translation Resources." *Natural Language Engineering,* 2019, pp.1-22.

Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J. and Chen, H., "Application of Generative Autoencoder in De Novo Molecular Design. *Molecular informatics*, *37*(1-2), 2018, p.1700123.

Chen, B., Sheridan, R.P., Hornak, V. and Voigt, J.H., "Comparison of Random Forest and Pipeline Pilot Naive Bayes in Prospective QSAR Predictions", *Journal of Chemical Information and Modeling*, 52(3), 2012, pp.792-803.

Corazza A., Di Martino S., Maggio V., Scanniello G., "Combining Machine Learning and Information Retrieval Techniques for Software Clustering." In: Moschitti A., Scandariato R. (eds) Eternal Systems. EternalS, *Communications in Computer and Information Science*, 255, 2011, Springer, Berlin, Heidelberg, pp. 42-60.

Diaz, F., Mitra, B., Craswell, N.: "Query expansion with locally-trained word embeddings." *arXiv preprint arXiv:1605.07891 (2016).*

Drucker, H., Shahrary, B. and Gibbon, D.C., "Support Vector Machines: Relevance Feedback and Information Retrieval", *Information Processing & Management*, 2002, 38(3), pp.305-323.

Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T., McDowell, R.M. and Gramatica, P., "Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification and Regression-Based QSARs", *Environmental health perspectives*, *111*(10), 2003, pp.1361-1375.

Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H.M. and Thrun, S., "Dermatologist-Level Classification of Skin Cancer" *Nature*, 2016.

Fu, H. and Qiu, G., "Fast Semantic Image Retrieval Based on Random Forest", In *Proceedings of the 20th ACM International Conference on Multimedia*, 2012, pp. 909-912.

Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P. and Aspuru-Guzik, A., "Automatic Chemical Design Using a Data-Driven Continuous Representation Of Molecules", *ACS central science*, *4*(2), 2018, pp.268-276.

Guan, S. and X. Zhang. "Networked Memex Based on Personal Digital Library." *Encyclopedia of Networked and Virtual Organizations*. IGI Global, 2008, pp. 1044-1051.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., and Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J. and Kim, R., "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs", *Jama*, 316(22), 2016, pp.2402-2410.

Hamel, P. and Eck, D., "Learning Features from Music Audio with Deep Belief Networks." In *Proceedings of International Society for Music Information Retrieval Conference*, 2010, pp. 339–344.

Hansen, L.K., Lehn-Schiøler, T., Petersen, K.B., Arenas-Garcia, J., Larsen, J. and Jensen, S.H., "Learning and Clean-up in a Large Scale Music Database", In *2007 15th European Signal Processing Conference*, 2007, pp. 946-950.

Herrera, J., Poblete, B., Parra, D., "Learning to Leverage Microblog Information for QA Retrieval", In *European Conference on Information Retrieval*, 2018, Springer, Cham, pp. 507-520.

Hinton, G.E., Osindero, S. and Teh, Y.W., "A Fast Learning Algorithm for Deep Belief Nets", *Neural computation*," 18(7), 2006, pp.1527-1554.

Hoogeveen, D., Bennett, A., Li, Y., Verspoor, K.M. and Baldwin, T., "Detecting Misflagged Duplicate Questions in Community Question-Answering Archives", In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

Hore, S. and Bhattacharya, T., "Analyzing Social Trend Towards Girl Child in India: A Machine Intelligence-Based Approach," In Recent Developments in Machine Learning and Data Analytics, Springer, Singapore, 2019, pp.43-50.

Jaderberg, M., Simonyan, K., Vedaldi, A. and Zisserman, A., "Reading Text in the Wild with Convolutional Neural Networks", *International Journal of Computer Vision*, 116(1), 2016, pp.1-20.

Kanimozhi, S. and Padmini Devi, B., "A Novel Approach for Deep Learning Techniques Using Information Retrieval from Big Data" *International Journal of Pure and Applied Mathematics*, 118(8), 2018, pp. 601-606.

Kim, Y., Seo, J., Croft, W.B., "Automatic Boolean Query Suggestion for Professional Search." In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2011, pp. 825–834.

Kumar, J., Ye, P. and Doermann, D., "Learning Document Structure for Retrieval and Classification", In *Proceedings of the 21st International Conference on Pattern Recognition*, 2012, pp. 1558-1561.

Lam, A.N., Nguyen, A.T., Nguyen, H.A. and Nguyen, T.N., "Combining Deep Learning with Information Retrieval to Localize Buggy Files for Bug Reports (N)." In *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2015, pp. 476-481.

Lee, H., Largman, Y., Pham, P. and Ng, A.Y., "Unsupervised Feature Learning for Audio Classification Using Convolutional Deep Belief Networks," *Advances in Neural Information Processing Systems*, 2009, pp. 1–9.

Lewis, D.D. "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval", In *European conference on machine learning*, Springer, Berlin, Heidelberg, 1998, pp. 4-15.

Lo, Y.C., Rensi, S.E., Torng, W., and Altman, R.B., "Machine learning in Chemoinformatics and Drug Discovery", *Drug Discovery Today*, Elsevier, 2018.

Lo, Y.C., Senese, S., Li, C.M., Hu, Q., Huang, Y., Damoiseaux, R. and Torres, J.Z., "Large-Scale Chemical Similarity Networks for Target Profiling of Compounds Identified in Cell-Based Chemical Screens", *PLoS Computational Biology*, 11(3), 2015, p.e1004153.

Lu, Y.C., Wu, C.W., Lu, C.T. and Lerch, A., "An unsupervised approach to anomaly detection in music datasets." In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 749-752.

Maarek, Y.S., Berry, D.M. and Kaiser, G.E., "An Information Retrieval Approach for Automatically Constructing Software Libraries." *IEEE Transactions on Software Engineering*, 17(8), 1991, pp.800-813.

Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X. and Ward, R., "Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24(4), 2016, pp.694-707.

Pati, K., Gururani, S. and Lerch, A., "Assessment of Student Music Performances Using Deep Neural Networks", *Applied Sciences*, 8(4), 2018, p.507.

Pollettini, J.T., Pessotti, H.C., Filho, A.P., Ruiz, E.E.S., Junior, M.S.A., "Applying Natural Language Processing, Information Retrieval and Machine Learning to Decision Support in Medical Coordination in an Emergency Medicine Context", *IEEE 28th International Symposium on Computer-Based Medical Systems, 2015*, pp. 316-319.

Sathya, M., Jayanthi, J. and Basker, N., "Link based K-Means Clustering Algorithm for Information Retrieval." In *International Conference on Recent Trends in Information Technology*, 2011, pp. 1111-1115.

Singhal, A., "Modern Information Retrieval: A Brief Overview." *IEEE Data Eng. Bull.*, 24(4), 2001, pp.35-43.

Sinha, U. and Kangarloo, H., "Principal Component Analysis for Content-Based Image Retrieval", *Radiographics*, 22(5), 2002, pp.1271-1289.

Song, Y., He, Y., Hu, Q. and He, L., "ECNU At 2015 CDS Track: Two Re-Ranking Methods in Medical Information Retrieval." In *Proceedings of the 2015 Text Retrieval Conference*, 2015.

Xiang, Y., Chen, Q., Wang, X. and Qin, Y., "Answer Selection in Community Question Answering via Attentive Neural Networks", *IEEE Signal Processing Letters*, 24(4), 2017, pp.505-509.

Yang, H.C., Lee, C.H., "Mining Unstructured Web Pages to Enhance Web Information Retrieval", *International Conference on Innovative Computing, Information and Control*, IEEE, 1, 2006, pp. 429-432.

Zhuang, B., Liu, L., Li, Y., Shen, C. and Reid, I., "Attend In Groups: A Weakly-Supervised Deep Learning Framework for Learning from Web Data." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1878-1887.