ISSN 2278-3091
Volume 9 No.2, March -April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse83922020.pdf
https://doi.org/10.30534/ijatcse/2020/83922020

Comprehensive study and Analysis of Extreme Multi-Label Classification Approach in Large Scale Recommendation System

Purvi Prajapati[1], Dr. Amit Thakkar[2]
[1]Department of Information Technology, CSPIT, CHARUSAT, India, purviprajapati.it@charusat.ac.in
[2] Department of Information Technology, CSPIT, CHARUSAT, India, amitthakkar.it@charusat.ac.in

ABSTRACT

Due to transformation of analog to digital world, huge volume of data is generated on daily basis. Millions of users are uploading, downloading and searching data on and from Social Media, Wikipedia, YouTube, Amazon, Websites etc. It is essential to analyze and retrieve significant information from millions of users with millions of categories. Hence, this is a challenging task to build classifier which will classify huge amount of data with its relevant subset of categories. In Recommendation System, the main goal is to recommend users based on the available data. However, this traditional recommendation system fails to deal with millions of items or labels in short time span. Extreme Classification approach is the recently introduced research area to tackle large amount of data with multi-label environment for the classification. Extreme Multi-Label classifier will construct model and predict the relevant categories from the large amount of available categories. This paper discussed different approaches for large scale Recommendation System using Extreme Multi-Label Classification Approach and empirical evaluation carried out on three multi-label datasets which handles large volume of the data.

Key words: Classification, eXtreme Multi-Label Classification (XMLC), Recommendation System, Machine Learning, Multi-Label Classification (MLC).

1. INTRODUCTION

Day by day number of users and information are increased drastically over internet, so there is a demand to extract and recommend relevant information to the users. Recommendation systems take advantage of user preferences, prioritization features, and recommend items that users would like. Basically Recommendation System is categorized in to content based recommendation and collaborative filtering based recommendation [1, 2, 23, 26]. In content based recommendation, mainly user preferences and item profiles are used to recommend items. Second approach is widely used in Recommendation Systems, in which item recommendation is based on user ratings or previous experience. This traditional Recommendation Systems are fail to deal millions of users with large amount of data.

Traditional Recommendation problem could be restructured into large scale classification approach which is referred as extreme multi-label classification problem. However, traditional Multi-Label Classification is used to classify multi label data (multiple items), but it fails to deal with millions of labels. Extreme Multi-Label Classification approach is used to overcome this limitation by selecting subset of items for the new instance from large amount of labels. Extreme Multi-Label Classification is also known as large scale classification approach [3, 4, 5, 7, 21, 24, 25].

As per the Wikipedia statistics, there are 5.764 million article available with average 1500 articles are added daily as of May 2018 [6]. Here eXtreme Multi-label Classifier (XMC) is used to predict subset of most relevant categories for Wikipedia article from large volume of categories [3, 4, 5, 8, 21, 27]. Same way in website or webpage classification, to recommend specific subset of categories such as sports, news, politics, business, technology etc. Product recommendation in online shopping is also one of the challenging task to recommend set of products within short time span.

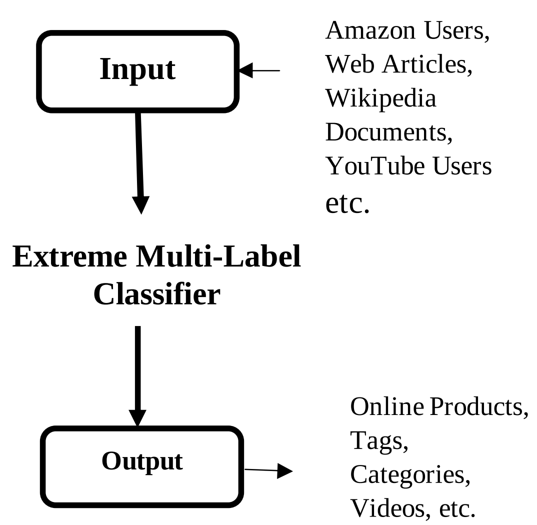

Figure 1: Extreme Multi-Label Classifier

Figure 1 represents Extreme Multi-label classifier for recommending data to millions of users with millions of categories.
Extreme Multi-Label Classifier is applicable to a product recommendation in e-commerce field, automatically recommending phrases from advertisement page, recommending set of relevant queries in search engine, recommending list of videos to YouTube users, music recommendation, article recommendation etc. [3, 4, 8, 22, 23]

In Section 2 basic approaches are mentioned for extreme classification where section 3 discuss state of art algorithms for large scale classification of Recommendation System. Section 4 provides dataset statistics and performance measures for Extreme Classification approach. Section 5 discusses experimental evaluation on three multi-label dataset. Section 6 represents current research issues in large scale recommendation system. At last conclusion and future directions are provided in section 7.

2. EXTREME CLASSIFICATION APPROACHES

Basically classification is categorized into Binary Classification, Multi-Class classification and Multi-label Classification problem. Binary classification problem deals with two class labels and assign one label to each instance. Multi-class Classification problem deals with multiple class label and it assigns one label to each instance. Where as in Multi-Label Classification problem, it assigns multiple labels to each instance. In large scale classification problem, Multi-label Classification approach fails to handle extreme number of features and labels. Extreme Multi-Label Classification (or Large-scale Classification) is divided into four different approaches such as Linear Classifier, Embedding Based Classifier, Tree Based Classifier and Deep Learning Based Classifier [3, 4, 5, 8, 11, 12, 13, 21].

Linear Classification or One-vs-All (OVA) is one of the general approach for classification problem. In Linear classification, independent binary classifier is applied for individual label. Training and Prediction of millions of labels with millions of independent classifiers are highly expensive and time consuming. It also require large space to store the model.

In embedding based approach reduces data from high dimensionality into low dimensionality. This approach has exploit label co-relations and data sparsity to compress the number of labels. During training and prediction phase, compression and decompression techniques are applied respectively for mapping of high dimensionality to low dimensionality [13, 21]. Various compression and decompression techniques are available in embedding based approach.

Tree based approach constructs hierarchical structure for labels, where all the instances with all the labels are present at root node. Then node splitting criteria is applied to generate left tree and right tree from the root node. Continue this splitting process till the leaf level nodes. Recursively split the node until all the leaf nodes contain predefined number of labels. A tree based classifier is applied on leaf node for the classification, which gives logarithmic time prediction cost if generated tree is balanced. This approach reduces prediction time by recursively arranging label based tree or feature based tree [3, 4, 21].

Deep learning based approach is also applicable for handling large dimensionality of multi label classification problem. In Deep Learning based Extreme Multi-Label (DXML) [8] paper, deep neural network is applied on the label graph to solve the label embedding task. DXML also enhanced classification accuracy by enhancing neural network ability through incremental learning approach. Deep Learning based approach achieves success to extract automatically large number of semantics features and labels from large set of instances. [8, 11, 21, 28].

3. RELATED WORK

Sparse Local Embedding for Extreme classification (SLEEC) [13] is the standard embedding based algorithm for extreme classification approach. SLEEC maintained label co-relationships using pairwise distance among nearest labels. In this prediction process, A k-nearest neighbor (kNN) classifier was applied by the SLEEC algorithm in which nearest neighbors persevered during training. SLEEC can efficiently scale data with a million labels that go beyond leading embedding methods [13, 21].

Multi-Label Random Forest (MLRF) [3] algorithm is tree based classifier that used recommending phrases from web pages to bid against search engine queries. It achieves logarithmic time prediction in terms of large number of labels. The performance of this algorithm was better than ranking based techniques and NLP (Natural Language Processing) based techniques [3, 21]. This algorithm implemented Gini Index's based brute force optimization method for partitioning the tree into the left and right subtrees [18]. This Gini index or entropy optimization is complicated over each feature when there is a large amount of training data and labels. FastXML[4] is based on MLRF[3] but uses a different optimization technique based on nDCG (Normalized Discounted Cumulative Gain). FastXML algorithm constructs balanced tree structured classifier that performs logarithmic time prediction in milliseconds. It is new paradigm in multi-label classification oriented towards recommendation system [4, 21].

Cold start Recommendation System using XML (CS-XML) [10] has developed a recommendation system for online discussion forums where new posts are created every moment, as a result Collaborative Filtering (CF) based algorithms won't work. CF based recommendation system will not work well when data is sparse data means history is not available for all user and items. And CF do not work when new user or new item problem. Collaborative Variational Auto encoder (CVAE) is used to handle cold start problem using Bayesian generative model [10]. It is effective compare to most widely used recommendation approaches [19]. Collaborative Topic Regression (CTR) is used the text data for recommendation with probabilistic matrix factorization (PMF) [10, 20].

XML-CNN [11] is the deep learning based CNN (Convolutional Neural Network) for extreme multi-label data. It combines CNN and consider multi-label co-occurrence to handle large dataset in Extreme Multi-Label Classification. This XML-CNN used dynamic max pooling scheme that extracts richer information from the text document.

4. DATASET AND PERFORMANCE MEASURE

Data source is available on "The Extreme Classification Repository" [17]. Below Table 1 shows data set statistics for three different data sets with high dimensionality. Extreme classifier performance evaluation is distinct from the traditional multi-label classifier. Various loss functions such as Precision, Recall, Hamming loss, F-score, and Jaccard distance are used in the multi-label classifier to evaluate classifier performance [5, 17].

Table 1: Dataset Statistics

Dataset	EUR Lex4K	AmazonCat	Wiki10
Feature Dimensionality	5000	203882	101938
Label Dimensionality	3993	13330	30938
# Train Points	15539	1186239	14146
# Test Points	3809	306782	6616
Average Points per Label	25.73	448.57	8.52
Average Labels per Point	5.31	5.04	18.64

Precision@k and Recall@k are the standard measures for the recommendation problems. nDCG (normalized Dis-counted Cumulative Gain) is a normalization of the Discounted Cumulative Gain (DCG) measure. nDCG is the measure of ranking the document in information retrieval, machine learning, and recommendation field. Reciprocal Rank (RR) is the measure to calculate rank for the first relevant document. When averaged across all queries, the measure is called the Mean Reciprocal Rank (MRR) [5, 17]. Table 2 represents different measures based on Precision@k and nDCG@k [5].

Table 2: Evaluation Measures for Extreme Classification

P@k	$\frac{1}{k} \sum_{l \in rank_k(\hat{y})} y_l$
PSP@k	$\frac{1}{k} \sum_{l \in rank_k(\hat{y})} \frac{y_l}{p_l}$
DCG@k	$\sum_{l \in rank_k(\hat{y})} \frac{y_l}{\log(1+l)}$
PSDCG@k	$\sum_{l \in rank_k(\hat{y})} \frac{y_l}{p_l \log(1+l)}$
nDCG@k	$\frac{DCG@k}{\sum_{l-1}^{\min(k\|\|y\|\|0)} \frac{1}{\log(1+l)}}$
PSnDCG@k	$\frac{PSDCG@k}{\sum_{l=1}^{k} \frac{1}{\log(1+l)}}$

5. IMPLEMENTATION

Experimental results are taken on three dataset as mentioned in Table 2 for benchmark algorithms such as SLEEC, FastXML and XML-CNN. The code for the benchmark algorithms are collected from "The Extreme Classification Repository" [17]. Below Figure 2 and 3 represents performance of the three extreme classification algorithms [4, 11, 13].

We measured precision@k and nDCG@k as evaluation metrics to verify the performance of the three extreme classification algorithms. We found from the experimental analysis that embedding based approach performs efficiently for EURLex-4k (small) dataset. And for high dimensional AmazonCat-13k dataset tree based approach and deep learning based approach performs comparatively the same.

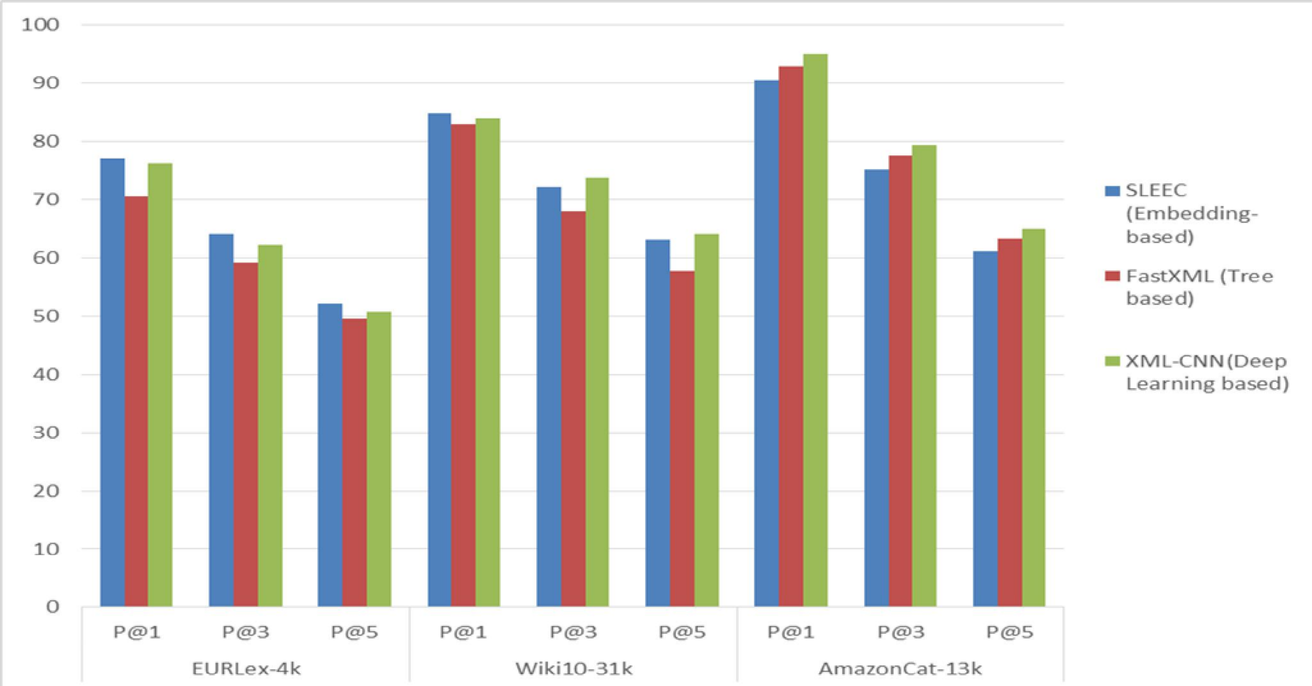

Figure 2: Results of Extreme Multi-Label Classifier using Precision@k Metric [4, 11,13]

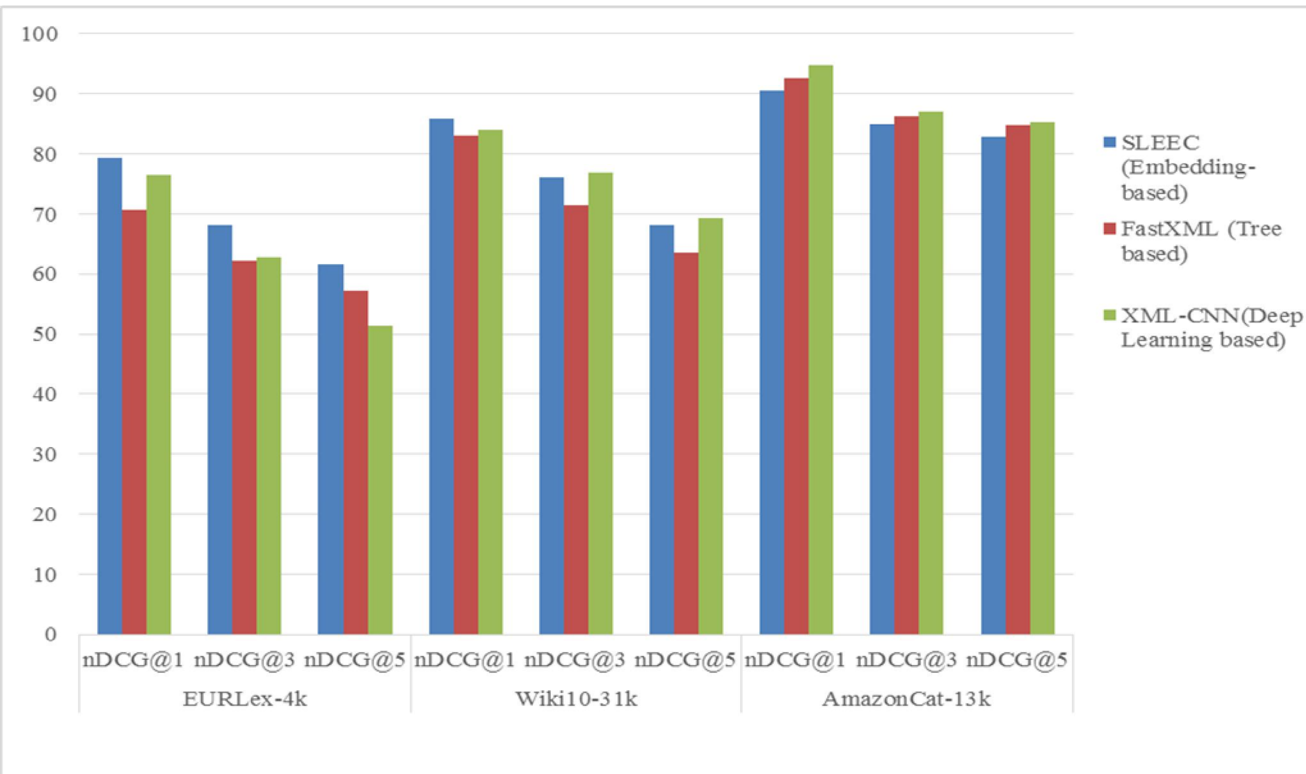

Figure 3: Results of Extreme Multi-Label Classifier using nDCG@k Metric [4, 11, 13]

6. RESEARCH ISSUES

Rareness

Rare item or category recommendation is one of the challenging issue in large scale classification approach. Physical item supply systems are categorized by a shortage of resources. Local physical shops have inadequate storage space and it can display only a small fraction of the items that exist in shop to the customers [2]. But on-line shopping can make everything as accessible to the customer. The dissimilarity between physical space and online systems was called the long tail effect. In long tail phenomenon, physical shops have few popular items while on-line companies can make everything possible to the customers [1, 2].

In extreme classification approach few of the labels have less number of training instances. Hence, it is crucial to classify inconsistently or rarely occurring labels compare to consistently occurring labels. These inconsistent labels are labeled as tail labels, which are challenging to predict but sometimes they are informative in few applications compared to frequently occurring labels [4, 7, 9, 21]. Figure 4 shows tail label occurrences in Wikipedia and Amazon [4, 5].

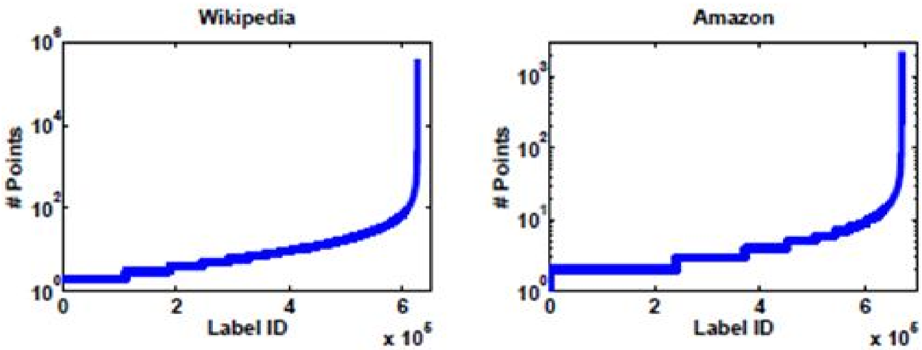

Figure 4: Long Tail Distribution [4, 5]

Novelty

New item or new user prediction is the well-known problem in Recommendation approach, which is referred as a cold-start problem. This can be possible for new users as well as for a new item, which has no any past ratings or history. Means there aren't enough user actions for a particular item in cold start problem. Content-based filtering is successfully addressing this challenge. It initially uses the metadata of new items while creating recommendations, while user's preferences go on the second phase for a specific period of time. Collaborative Filtering approach fails to recommend newly posted item or newly joined user [1, 2, 10]. Now cold start problem is treated as eXtreme Multi-Label Classification, where for new item or new user problem, extreme classifier will predict subset of users or items respectively from large amount of data.

Scalability and Sparsity

In eXtreme Multi-Label Classification, label space and user space both are very large. In this high dimensional environment, it is crucial to perform efficient and fast training and prediction [4, 5, 7, 24]. Data sparsity problem is occurred due to large amount of labels and to extract relevant subset of labels from high dimensional environment [11].

Label synonym and Label polysemy

In Label synonymy, item or label name is dissimilar but meaning of label is same. So in eXtreme Multi-Label environment, it is very crucial to predict synonyms of the each item or label. One item or label with multiple meaning is referred as Label Polysemy. To extract the meaning of the label and find its synonym or polysemy is one of the challenging task in eXtreme Classification with millions of label. Using label synonym we can build effective recommendation system to recommend items to user [15, 16].

Implicit Ratings

Basically Collaborative Filtering and Content Based Recommendation System are using explicit ratings of the users for items. These approaches are fails to handle implicit ratings [12].

Real Time Recommendation

Traditional Recommendation System is unable to perform training and prediction in milliseconds. Collaborative Filtering approach worked on Low Rank Assumption which have poor prediction in extreme setting [10, 13].

7. CONCLUSION AND FUTURE WORK

This paper represented different Extreme Multi-label Classification approaches to solve large scale recommendation problem. Empirical evaluations are performed on the three benchmark dataset to the most widely used extreme multi-label classifiers, where label space and user space both are high-dimensional. Results demonstrated that the Recommendation System using Extreme Multi-Label Classification strategy is effective for handling large volume of the data. Collaborative Filtering based Recommendation System worked on low rank assumption which has poor performance in extreme environment.

Our future work will plan to model how interest of communities are changed over time. Over a long time, not all people will remain interested in the same item as their experiences change. Also, in future work will concentrate on deep learning based model for the on line event Recommendation System.

REFERENCES

1. R. Francesco, L. Rokach, and B. Shapira, "**Introduction to recommender systems handboo**k", In Recommender systems handbook, pp. 1-35, Springer US, 2011. https://doi.org/10.1007/978-0-387-85820-3_1
2. MOOC course on "**Massive Mining Datasets**" offered by Stanford University. http://infolab.stanford.edu/~ullman/mmds/book.pdf.
3. R. Agrawal, A. Gupta, Y. Prabhu, and M. Varma, "**Multi-label learning with millions of labels: Recommending advertiser bid phrases for webpages**", In Proceedings of the 22nd international conference on World Wide Web, pages 13-24, ACM, 2013. https://doi.org/10.1145/2488388.2488391
4. Y. Prabhu and M. Varma, "**Fastxml: Afast, accurate and stable tree-classifier for extreme multi-label learning**", In Proceedings of the 20th ACMSIGKDD international conference on Knowledge discovery and data mining, pages 263-272, ACM, 2014. https://doi.org/10.1145/2623330.2623651
5. H. Jain, Y. Prabhu, and M. Varma, "**Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications**", In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 935-944, ACM, 2016. https://doi.org/10.1145/2939672.2939756
6. Wikipedia statistics. http://expandedramblings.com/index.php/wikipedia-statistics/.
7. K. Jasinska and N. Karampatziakis, "**Log-time and log-space extreme classification**", arXiv preprint arXiv:1611.01964, 2016.
8. W. Zhang, J. Yan, X. Wang and H. Zha, "**Deep extreme multi-label learning**", arXiv preprint arXiv:1704.03718, Jun 2018. https://doi.org/10.1145/3206025.3206030
9. C. Xu, D. Tao and C. Xu, "**Robust extreme multi-label learning**", In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA August, pages 13-17, 2016.
10. K. Halder, L. Poddar and M. Kan, "**Cold Start Thread Recommendation as Extreme Multi-label Classification**", In International World Wide Web Conference, 2018. https://doi.org/10.1145/3184558.3191659
11. J. Liu, W. Chang, Y. Wu, and Y. Yang, "**Deep learning for extreme multi-label text classification**", In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 115-124, ACM, 2017. https://doi.org/10.1145/3077136.3080834
12. Y. Prabhu, A. Kag, S. Gopinath, K. Dahiya, S. Harsola, R. Agrawal and M. Varma, "**Extreme Multi-label Learning with Label Features for Warm-start Tagging, Ranking & Recommendation**" , WSDM Feb-2018. https://doi.org/10.1145/3159652.3159660
13. K. Bhatia, H. Jain, P. Kar, M. Varma, and P. Jain, "**Sparse local embedding's for extreme multi-label classification**", In Advances in Neural Information Processing Systems, pages 730-738, 2015.
14. P. Yashoteja, K. Anil, H Shrutendra, R. Agrawal and M. Varma, "**Parabel: Partitioned Label Trees for Extreme Classification with Application to Dynamic Search Advertising**", World Wide Web Conference (WWW), 2018.
15. J. Wei and F. Meng, "**Personalized information recommendation based on synonymy tag optimization**", Cluster Computing, Pages 1-12, Springer 2017.
16. M. Lombardi, A. Marani, "**SynFinder: a system for domain-based detection of synonyms using WordNet and the web of data**", In Mexican International Conference on Artificial Intelligence, pp. 15-28, Springer, Cham, Oct 2015. https://doi.org/10.1007/978-3-319-27060-9_2
17. The Extreme Classification Repository https://manikvarma.github.io/downloads/XC/XMLRepository.html
18. L. Breiman, "**Random Forests**", Machine Learning, pp 5-32, Oct 2001. https://doi.org/10.1023/A:1010933404324.
19. X. Li , J. She, "**Collaborative Variational Autoencoder for Recommender Systems**", In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13-17, 2017, Halifax, NS, Canada.
20. C. Wang and D. Blei, "**Collaborative topic modeling for recommending scientific article**s", In Proceedings of SIGKDD, ACM, 448– 456, 2011. https://doi.org/10.1145/2020408.2020480
21. P. Prajapati, A. Thakkar, "**Extreme multi-label learning: A large scale classification approach in machine learning**", Journal of Information and Optimization Sciences, June 2019.
22. ZZ. Lin, BR. Dai, "**Reweighting Forest for Extreme Multi-label Classification**", In International Conference on Big Data Analytics and Knowledge Discovery, Aug 28 pp. 286-299, 2017, Springer, Cham. https://doi.org/10.1007/978-3-319-64283-3_21
23. S. Sohail, J. Siddiqui and R. Ali, "**Classifications of Recommender Systems: A review**", Journal of Engineering Science & Technology Review, 2017 Jul 1; 10(4).
24. K. Jasinska and N. Karampatziakis, "**Log-time and log-space extreme classification**", arXiv preprint arXiv: 1611.01964, 2016.
25. H. Yu, P. Jain, P. Kar, and I. Dhillon, "**Large-scale multi-label learning with missing labels**", In

International Conference on Machine Learning, pages 593-601, 2014.

26. Adeagbo M. A., Akhigbe B. I. and Afolabi B. S., "**Towards A Job Recommender Model: An Architecture-Based Approach**", International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No. 6, November-December 2019. https://doi.org/10.30534/ijatcse/2019/94862019
27. Sang Boem Lim, "**Classification and Big Data Usages for Industrial Applications**", International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No. 4, July - August 2019. https://doi.org/10.30534/ijatcse/2019/18842019
28. Sheetal S. Pandya , Nilesh B. Kalani, "**Review on Text Sequence Processing with use of different Deep Neural Network Model**", International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No. 5, September-October 2019. https://doi.org/10.30534/ijatcse/2019/56852019