International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 01, 2020
ISSN: 1475-7192

A Literature Review on Different Types of Machine Learning Methods in Web Mining

Meera Alphy and Ajay Sharma

Abstract--- *The increase in the web usage in last two decades, tremendously enhance the research field to tackle the challenges faced by online user and browsing patterns to help the user by analysing the user clickstream from log file. This review gives leverage convenient web elements suitable in the web usage mining and concentrate on mining the web usage on the latest years. The classification of different algorithm on the common characteristics of its function and its learning mechanism is the highlight of this survey paper. World wide web systems become core competency of online transactions, online corporate existence which is prevalent to day to day life in the building and easy to access knowledge or products.*

Keywords--- *Supervised Learning, Unsupervised Learning, Semi Supervised Learning, Reinforcement Learning, Machine Learning, Deep Learning, Web Mining, Web Usage Mining.*

I. INTRODUCTION

Understanding the increase in the complex, dynamic, and uncertain intertwining of Internet technologies. Business systems creates new challenges and opportunities that necessitate the transcending of disciplinary boundaries. The size of web pages increasing day by day. The marketing, e-business and research increases day to day with the increasing number of users. It's a platform to connect with different people in different places with the 'tip of the finger' in milliseconds. As a result, the increase in network range, data complexity arises; and access of required data from the online raw data becomes more complicated. World wide web pages size according to the indexed web [1] contains at least 5.11 billion pages (Friday, 11 january,2019) and the Dutch indexed web [1] contains at least 351.79 million pages (Friday,11 january,2019). Challenges faced to retrieve the useful data from raw data contribute for the rise of web mining research. How the word "Web mining" originated and how it transformed to different forms and how the research of this field contributed to new technologies are discussed in this chapter. The web mining is classified in three' web content mining, web structure mining and web usage mining. Different users use patterns with wide range of similarity in their clickstream in World Wide Web. Identifying these patterns, forms the domain web usage mining which is a subdomain of web mining. There are different types of techniques used for pattern discovery and pattern analysis. A detailed review of techniques used for each phase of web usage mining. After passing through different phases of web usage mining, the similar patterns of user's clickstream are generated. Depend on input data the developer work with; the result can use for business improvement, advertising, medical sector, research, web personalisation and recommended system. Main contribution of worldwide web is online business, any user can buy any product from any place with sitting in the home; which can be said as "what a small world". To improve online business, the customer relationship

Meera Alphy, Research Scholar, Department of Computer Science and Engineering, SRM University, Delhi- NCR, Sonipat, Haryana, India. E-mail: meeraalphy.urumbath@gmail.com
Ajay Sharma, Associate Professor, Department of Computer Science and Engineering, SRM University, Delhi-NCR, Sonipat, Haryana, India. E-mail: ajaypulast@reffimail.com

DOI: 10.37200/IJPR/V24I1/PR200276
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

management in the organisation which deals with the customers and the product they buy which helps for future customers purchase. The pattern attained by different costumers buy the similar products can improve the website and the product recommendation. This results in the web usage mining importance in this era.

1.1. Machine Learning

The term Machine Learning is mentioned by Arthur Lee Samuel in his paper 'some studies in machine learning using game of checkers' in 1959 [11]. Machine Learning is the branch of computer science that has to do with building algorithms that are directed by facts or information. Machine learning [12][13] is an artificial intelligence (AI) technique to automatically learn and improve computer systems to provide better performance.Machine learning is defined as automatic programs for data organise, classify and evaluate. There are four types of machine learning methods. They are

a. supervised machine learning

b. unsupervised machine learning

c. semi supervised machine learning

d. reinforcement machine learning

Supervised method [14]

This kind of method can automatically generate output by analysing the techniques used in past. Here data is labelled. There is no human interference required in this method. The system will learn and analysis dataset to predict the output. The learning algorithm is used to check the predicted output is accurate. This method is fully programmed such a way where the desired output generates by automatic learn and improve technique. From the figure 1, shows how supervised learning method works, with an example. Here a picture is the input and algorithm analysed, learned and improved and shows the predicted output is a book.

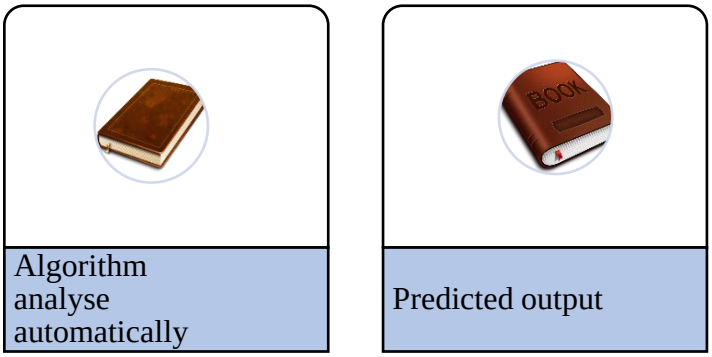

Figure 1: Supervised Learning analysis

Unsupervised method [14, 15]

This kind of method cannot automatically generate output. Its contrast to the supervised machine learning method. Here data is unlabelled. Its need human interference to analysis the dataset. The algorithm comes under this method need few human interferences to group the data and algorithm to proceed to get desired output.From the figure 2, shows how unsupervised learning method works, with an example. Here a picture is the input and teach the model by human interference to analyse algorithm and the model is trained.

DOI: 10.37200/IJPR/V24I1/PR200276
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

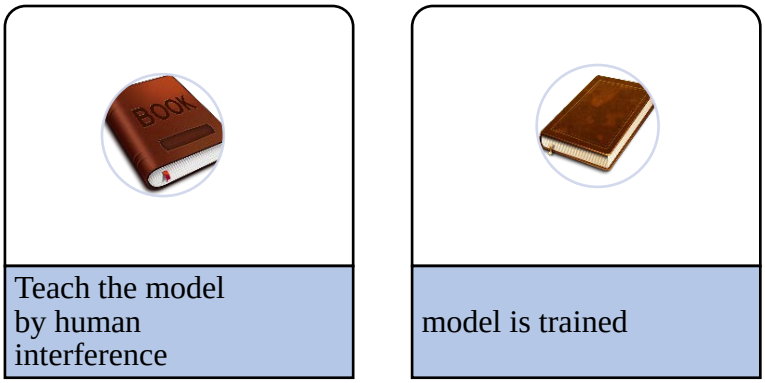

Figure 2: Unsupervised model

Semi supervised method [16]

This kind of method used in dataset where minority data are labelled and majority data is unlabelled. So, this method takes half supervised and unsupervised methods. It is applied where majority data are unlabelled such as image processing, pattern recognition and bioinformatics. From the figure 3, shows how semi-supervised learning method works, with an example. Here the input contains the labelled and unlabelled data and algorithm analyse starts by classify by using small amount of the labelled data and then uses this labelled data to classify the unlabelled data.This method increases the accuracy of the system by learn or training it. And handle the large dataset having semi-supervised learning with more accurate output. Error rates reduced by the learned extractors [17].

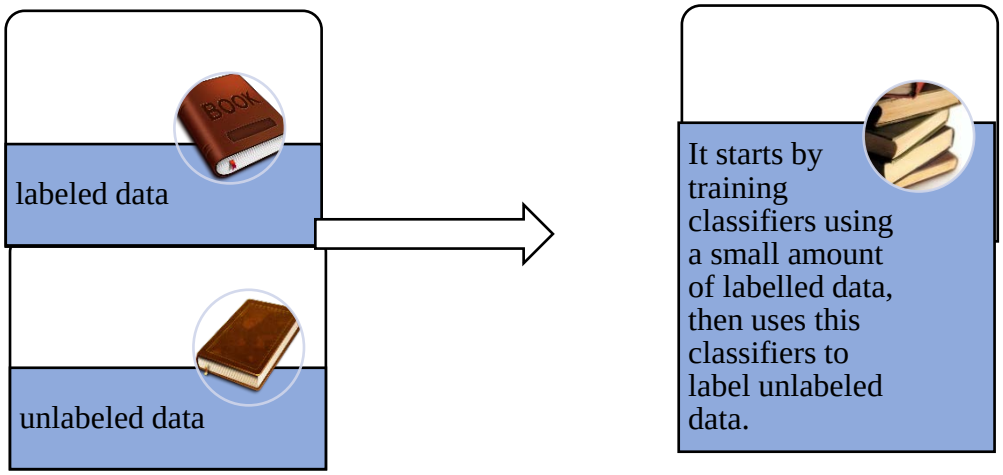

Figure 3: Semi-supervised Learning method

Reinforcement method[18]

This method uses to identify the best function for productive action. It can attain by learn and error search and delayed rewards. The agent helps to reinforcement method is reward feedforward. The current output is determined by previous and next output. For example, to detect the unauthorized sites by analysing precious and next actions. Its helps in security and fraud analysis in web mining. And mainly used for chess game. Reinforcement method finds the output from sequence of actions.

DOI: 10.37200/IJPR/V24I1/PR200276
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

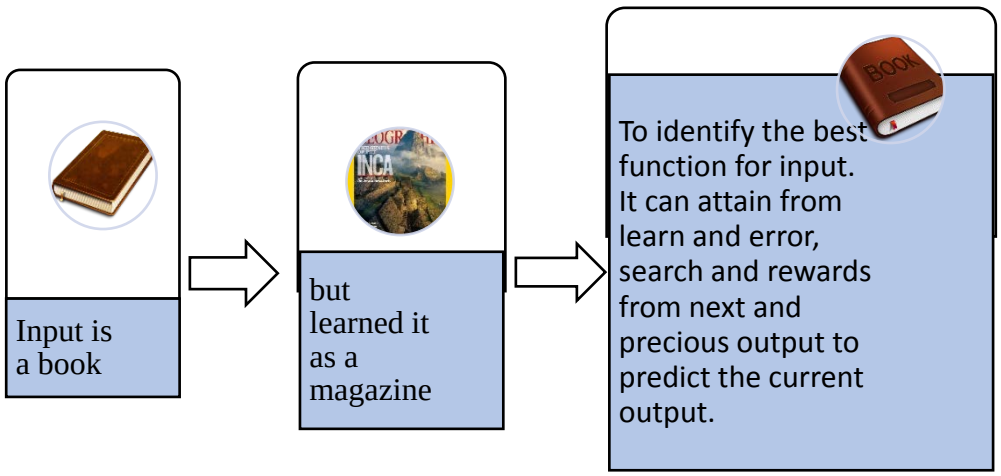

Figure 4: Reinforcement Learning

They are two types of reinforcement Learning [19], they are

i. Positive Reinforcement Learning

Increase in the forte and the regularity of the action, take place due to a specific action in an event which is known as Positive Reinforcement Learning. It can also be said as favourable and supportive actions for Learning. Qualities of this method are making a best use of performance and help to adapt for a lasting session of time. Drawback of this method is, an excess production of favourable analysis can weaken the results (if the method uses beyond need).

ii. Negative Reinforcement Learning

The solidification of an action due to negative state is over or sidestepped is known as Negative Reinforcement Learning. Qualities of this method are improvement in actions and deliver the bold disobedience to least performance standard. Drawback is link to least actions.

Deep Learning

The deep learning is introduced by Rina Dechter [3] to machine learning. It is also referred as Deep Hierarchical learning and Deep Structural learning. Deep learning [20,21,22] is a subset of machine learning which used for representation to illustrate the facts that can be learned by supervised, semi-supervised and unsupervised form of the Human brain's behaviour and function [23,24,25].The deep learning algorithm contains deep NN algorithm, deep BN algorithm [26,27] and recurrent NN algorithm [28,29,30]. The deep learning is a part of machine learning where refer to the deep artificial neural network (deep ANN). Deep artificial neural network can contain more than one hidden layer. There are input layer, hidden layer and output layer. Deep learning supervised learning examples are feature engineering and unsupervised learning example are deep BN algorithm. The inputs can be analysed using supervised or unsupervised Deep learning.

1.1.1. Soft Computing And Hard Computing [31]

Soft computing is the operational solutions to complex computational problems arising from the use of approximate calculations. Soft computing any technique derived from human mind. Its characteristics are

DOI: 10.37200/IJPR/V24I1/PR200276
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

impression, partial truth and uncertainty. It contains components such as fuzzy logic, machine learning, genetic algorithms, swarm intelligence, ant colony optimization, particle optimization, Bayesian networks, artificial neural networks and expert systems.Neural network is a computer system modelled on the human brain and nervous system. Within the field of machine learning, neural networks are a subset of algorithms built around a model of artificial neurons spread across three or more layers. In 1990's soft computing term coined [32].Hard computing is conventional computing technique where exact input data, model and output have real solution. Hence it takes lot of time to execute. Hard computing has characteristics are precision and deterministic. It contains components such as binary logic and numeric analysis. Hard computing input data should be exact data but in soft computing input data can be noisy data.

II. SURVEY ON LEARNING TECHNIQUES IN WEB USAGE MINING

2.1. Recommendation system using web usage and domain knowledge [4]

Three phases have been implemented for the proposed recommender system. In the first phase, ontology to represent the domain knowledge. Second phase, semantic network for domain terms and third phase, automatically create a semantic recommendation system using web usage and domain knowledge. The proposed work starts with construction of domain ontology, which mainly deal with extraction of useful data from the log file (pre-processing). And then defining the concepts or terms which have features of same concept sharing. Finally, the define the taxonomic and non-taxonomic relationships which have three types of taxonomic relationship' top-down development process, bottom-up development process and hybrid development process. And the author taken hybrid development process for taxonomic for proposed system. The non-taxonomic relations method such as 1-M and M-N relationship. The author taken M-N relationship for relational database. Author generated this ontology domain in three levels' general level, specific level and web page level. In general level gives a relationship between the ontology terms with web page and terms definition. In specific level, relationship between domain terms related to the domain concepts. In web page level, it holds the relation between the web page in the web site and web page to the domain terms. Queries were 'isAbout' and 'hasPage' object property used. The input has taken from public dataset. Performance of new recommender system has calculated by precision and satisfaction. Advantage of this method is better performance in terms of precision and satisfaction. This method can be improved by depending more on Minsup compared to PLWAP-Mine and more parameters can used. Future work can be done for improving term extraction method. And the author suggested key extraction algorithm for this future work.

2.2. A graph illustration for the associations between HTTP requests [5]

The author proposed request dependency graph for better modelling of the similarity between HTTP requests to extract the user similar clickstream by features of web usage. Pre-processing is processed to extract useful HTTP request. Author proposed two-step algorithm for the analyse the strength of the node by the accessed time, define each node whether its predecessor node or successor node and calculate edge weight. First step was launching the request dependency graph from obtained HTTP and second step was statistical inference approach to recognize the key requests. The input was real network data where features of this dataset consists of 339.2 GB, 49,103 of unique devices, 26,059 of hosts, 2,025,994 of HTTP requests. Primary object and secondary objects' two types of web

DOI: 10.37200/IJPR/V24I1/PR200276
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

object where primary object is accessed mostly, secondary object is accessed for few times from the web page. Primary are accessed URL by opening a hyperlink in a web page and it's a main feature to know which device accessed. Secondary requests are remaining requests other than primary requests. The advantage of this proposed method was improved F1 score where it attains by precision and recall' defines accuracy of experiment results. To improve the proposed method by using more parameters used for distribution of objects in a web page. Future work suggested by author was different applications this work can be used, different distribution of objects in a web page and extraction & visualization of large and complex dataset.

2.3. An improved method for web navigational usability by differentiating actual and predicted usage [6]

The proposed method is used for web navigational usability by comparing actual and predicted output. Author divided into three phases' pre-processing data, applying web usage algorithm and the anticipated usage by intellectual specialists based on their understandinguser behavior. The pre-processing tasks are sub divided into data cleansing, user identification, user session identification and path completion. The proposed algorithm called Ideal User Interactive Path Model (IUIP model) is an unsupervised learning algorithm which include the ideas of ACT-R (Adaptive Control of Thought-Rational)model and gives importance to both path and benchmark interactive time. The input to the proposed work is a small service-oriented website which consist 3000 entities. And output obtained 58 unique users and 81 sessions. The proposed work compared with traditional method by task success rate, average effort and average time. Advantage of this method gives better results than traditional methods in terms of handling large and complex dataset. This method can be improved by using more importance to user satisfaction and its performance facets of usability. Future work can be done by additional tools used for proposed IUIP modeling architecture with more improved and optimizedmultifaceted tasks.

2.4. A dependency graph method to label the similarity between web requests access [7]

The proposed method contains architecture and experimentally generate a heuristic parallel algorithmto differentiate user clickstream with the supportof cloud computing technology which is hybrid learning method.User clicks identificationdefines the procedure that recognise set of requests generated by users' online actions from a huge quantity of HTTP requests took. The author executed three steps in the proposed work. In step 1, data pre-processing and setting request sequence defines to remove logs that areunusable or half-finished actions. In step 2, generating dependencygraph defines represent the similarity between the user clickstream from logs with dependencygraph model. In Step 3,identifying user clicks defines vertex and weight given to the dependency model to recognize the similar user clickstream. Author applied parallel algorithm to improve these steps. Parallel algorithm contains MapReduce paradigm which operates with map stage and reduce stage. The input data is real massive data from traffic monitoring system (TMS) with a size of 228.7GB which have three million users. The characteristics of the dataset are 89,956userswith file size 288.7GB, 3,491,280 no of users, 453,181,959 no of request, 89,956 identified main request.The advantage of this method shows increase in precision, coverage, F1 function and speed evaluation than previous methods. This method can be improved by giving more importance to user behaviour and web page characteristics. According to author, future work can be constructing an unsupervised learning model for repository covering all accessed URL.

DOI: 10.37200/IJPR/V24I1/PR200276
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

2.5. An improved web usage service graph of recommending user's interest in web service [8]

Author proposed new method of web service recommendation which includes user's most likely favourite quality of service and variety of user on Web services. User's likes and quality of service priority on web usage are mostly done first by analysing the Webusage history. Then author gives weights to the web usersusage of online data by calculating their importance with past and most preferred user likes, and their useful quality of service. And developed aweb usage service graph based on the well-organized relationship between usage of web services. And finally generate a ranking algorithm for ranking web usage service. The input used is real-world Web service dataset named WS-DREAM project. The advantage of this proposed algorithm is increase in web usage service recommendation system performance similarity measures such as diversity, functional relevanceand QoS utility combination, and valuation of the diversified ranking.This method can be improved by diversified ranking measure using k-hop nearest neighbours. Future work can be done by clusteringmethods to improve the similarity computation and analysing real time dataset.

III. APPLICATION OF WEB USAGE MINING

Applications of web usage mining are mainly used in advertisement, investigation, scam detection and research. The importance of web usage mining is to improve the design of website & online search engines by analysing user's behaviour patterns., customer relationship management play an important role. An improved understanding of the customer's emotions, requirements, and interests for particular time period can help to increase the business profit by correlating cross selling or selling items which the purchaser wants to buy. Customer relationship management allows understanding who all are the customers where they belong to do, what they are interested in, and how their interest changes with respect to time, season and emotion. This proposed method includes an improvement in terms of similarity measures' precision, coverage and scalability. There are plenty of application in real life with regard with world wide web. Web usage mining is extracting useful information from raw data by analysing user click stream, navigation and user behaviour on online resources. Some other applications are e-commerce[12], personalisation website, system improvement, security[10], site design support, e-learning, online business improvement, social intelligence, speech recognition, online fraud detection, online health care such as fraud detection, dosage error detection, connected machines, clinical trial participation, preliminary diagnosis and cybersecurity.

IV. PROGRAMMING LANGUAGES USED FOR WEB USAGE MINING

The programming languages used for pattern analysis in-demand are

- Python,
- Java,
- C,
- C++
- Ruby and
- JavaScript

DOI: 10.37200/IJPR/V24I1/PR200276
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

V. CONCLUSION

The web contains a huge amount of data and make accessible in any part of the world. Web mining is a challenging task due to its huge size and semi structure of the web data. Web mining is multidiscipline field such as data mining, machine learning, artificial intelligence, deep learning and natural language. Main challenges of web mining are to make easily access the information from web (from huge amount of information), redundant (repeating the same information in many pages), dynamic (information in web changes) and noisy (mixture of different variety of unwanted data). The web mining term created by Etzioni [9]in his paper (1996), web mining is the application of data mining techniques to automatically unearth and separate information from World Wide Web.

Majority of work done in supervised learning than unsupervised learning. By analysing recent research, we have found out most work on unsupervised 90%. And very few researches only on supervised learning approximately 5%. Research on Semi-supervised and reinforcement learning remaining 5%. And most of the research work on web usage mining analysed by precision, recall and F1 function. Other similarity measures used in web usage mining are cohesion-separation: inter and intra cluster similarity,silhouette coefficient,similarity matrix approach, robust cardinality, dumn and dumn like indics, entropy, purity and rate of contractility and veracity.

REFERENCES

[1] https://www.worldwidewebsize.com
[2] https://www.expertsystem.com/machine-learning-definition/
[3] Rina Dechter"Learning while searching in constraint-satisfaction problems" *AAAIproceedings of the 5th national conference on artificial intelligence, volume: 1, page(s): 178-185*, 1986.
[4] Thi Thanh Sang Nguyen, Hai Yan Lu and Jie Lu, "Web-Page Recommendation Based on Web Usage and Domain Knowledge"*IEEE Transactions on Knowledge and Data Engineering*, volume: 26, issue: 10, page(s)**:** 2574 – 2587, 2014.
[5] Jun Liu, Cheng Fang, Nirwan Ansari, "Request Dependency Graph: A Model for Web Usage Mining in Large-Scale Web of Things" *IEEE Internet of Things Journal*, volume: 3, issue: 4, page(s):598 – 608, 2016.
[6] RuiliGeng and Jeff Tian "Improving Web Navigation Usability by Comparing Actual and Anticipated Usage" *IEEE Transactions on Human-Machine Systems*, volume: 45, issue: 1, page(s): 84 – 94, 2015.
[7] Cheng Fang ; Jun Liu ; Zhenming Lei " Parallelized user clicks recognition from massive HTTP data based on dependency graph model" *Wireless Communication over ZigBee for Automotive Inclination Measurement China Communications IEEE*, volume: 11, issue: 12, page(s): 13 – 25, 2014.
[8] Guosheng Kang, Mingdong Tang, Jianxun Liu, Xiaoqing (Frank) Liu and BuqingCao "Diversifying Web Service Recommendation Results via Exploring Service Usage History" *IEEE Transactions on Services Computing*, volume: 9, issue: 4, page(s): 566 – 579, 2016.
[9] Oren Etzioni "The World Wide Web: quagmire or gold mine?" *Appears in Comm. of ACM*, 1996.
[10] S. SobithaAhila and K. L. Shunmuganathan "Role of Agent Technology in Web Usage Mining: Homomorphic Encryption Based Recommendation for E-commerce Applications", *Springer Science+Business Media*, 2015.
[11] Ian Witten, Eibe Frank and Mark A Hall "Data Mining: Practical Machine Learning Tools and Techniques" *Elsevier Morgan Kaufmann Series in Data Management Systems*, third edition, 2010.
[12] *Tom M Mitchell "Machine Learning", McGraw-Hill Science/Engineering/Math, 1997.*
[13] https://www.expertsystem.com/machine-learning-definition/
[14] Pavel Laskov, Patrick Dussel, Christin Schafer and Konrad Rieck "Learning Intrusion Detection: Supervised or Unsupervised*?" International Conference on Image Analysis and Processing*, pages: 50-57, 2005.
[15] Jiliang Tang and Huain Liu "An Unsupervised Feature Selection Framework for social Media Data" IEEE Transactions on Knowledge and Data Engineering, volume: 26, issue: 12, pages: 2914- 2927, 2014.

DOI: 10.37200/IJPR/V24I1/PR200276
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020

[16] Chapelle O, Scholkopf, Zien A and Eds "Semi-Supervised Learning" *IEEE Transactions on Neural Networks*, volume: 20, issue: 3, pages 542- 542, 2009.
[17] Andrew Carlson, Justin Betteidge, Richard C Wang and Estevam R Hruschka Jr and Tom Mitchell"Coupled Semi- Supervised Learning for Information Extraction" *third ACM International conference on Web search and data mining*, pages: 101-110, 2010.
[18] Lucian Busoniu, Robert Babuska, Bart De Schutte "Multi-agent Reinforcement Learning: An Overview" *Innovations in Multi-Agent Systems and Applications-1 Springer*, pages: 183-221, 2010.
[19] www.greeksforgreeks.org / what is reinforcement learning.
[20] YoshuaBengio, Aaron Courville and Pascal Vincent "Representation Learning: A Review and New Perspectives" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume: 35, issue:8, pages: 1798-1828, 2013.
[21] Jurgen Schmidhuber "Deep Learning in Neural Networks: An Overview" *Elsevier Neural Network*, volume: 61, pages: 85-117, 2015.
[22] Yann LeCun, YoshuaBengio and Geoffrey Hinton "Deep Learning" *Nature*, volume: 521, issue: 7553, pages: 436-444, 2015.
[23] Adam H Marblestone, Greg Wayne and Konrad P Kording "Toward an Integration of Deep Learning and Neuroscience" *Frontiers in Computational Neuroscience*, volume: 10, issue: 5, pages: 1-60, 2016.
[24] Bruno A Olshausen and David J Field "Emergence of simple-cell receptive field properties by learning a sparse code for natural images" *Nature*, volume: 381, issue: 6583, pages: 607-609, 1996.
[25] YoshuaBengio, Dong-Hyun Lee, Jorg Bornschein and Zhouhan Lin Bengio "Towards Biologically Plausible Deep Learning" 2015.
[26] Hinton G "Deep belief networks" *Scolarpedia*, volume: 4, issue: 5, page: 5947.
[27] Geoffrey E Hinton, SiminOsindero, Yee WhyeTeh "A Fast learning algorithm for deep beliefs nets" *Neural Computation*, page: 18, issue:7, pages 1527-1554, 2006.
[28] Alex Graves, Marcus Liwicki, Santiago Femandez Roman Bertolami, Horst Bunke and Jurgen Schmidhube "A Novel Connectionist System for Improvedd Unconstrainedd Hand writing Recognition" *IEEE Transactions on Patterns Analysis and Machine Intelligence*, volume :31, issue: 5, pages: 855-868, 2009.
[29] HasimSak, Andrew Senior and Francoise Beaufays "Long Short- Term Memory recurrent neural network architectures for large scale acoustic modeling" *INTERSPEECH*, pages: 338-343, 2014.
[30] Xiangang Li and Xihong Wu "Constructing Long Short-Term Memory based Deep Recurrent Neural Networks for Large Vocabulary Speech Recognition" *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages: 4520-4524, 2015.
[31] Yu ChiHo "An explanation of ordinal optimization: Soft computing for hard problems" *Elisvers Information Science*, volume: 113, issue: 3-4, pages 169-192, 1999.
[32] Xiangang Li and Xihong Wu "Constructing Long Short-Term Memory based Deep Recurrent Neural Networks for Large Vocabulary Speech Recognition" *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages: 4520-4524, 2015.

DOI: 10.37200/IJPR/V24I1/PR200276
Received: 22 Nov 2019 | Revised: 26 Dec 2019 | Accepted: 10 Jan 2020