ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 8| Issue: 3| March 2022|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2022: 8.205 || ISI Value: 1.188

A SURVEY ON DATA INTEGRATION IN DISTRIBUTED WEB INFORMATION SYSTEM USING MACHINE LEARNING TECHNIQUES

Jinduja. S[1] , Narayani. V[2],
[1]*Research Scholar, Manonmaniam Sundaranar University, Tirunelveli,*
[2]*Assistant Professor/Computer Science, St Xavier's College, Tirunelveli*

Article DOI: https://doi.org/10.36713/epra9739
DOI No: 10.36713/epra9739

ABSTRACT
Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention. In our current advanced digital era it is very tedious process to identify an efficient integration of distributed sources in web information with the supported methodology towards enhanced time and space complexities. The existing web data based models are not effective in terms of distributed information processing, lack of optimal techniques due to the complexity in handling different web data resources, checking the effective integration output is also not feasible in the existing web data handling system. This paper presents a survey on Data Integration in Distributed Web Information System using Machine Learning Techniques.

KEYWORDS—*Segmentation, Web domain, Client structure, Machine learning, Performance*

1. INTRODUCTION

Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Machine learning is so pervasive today that you probably use it dozens of times a day without knowing it. Many researchers also think it is the best way to make progress towards human-level AI.

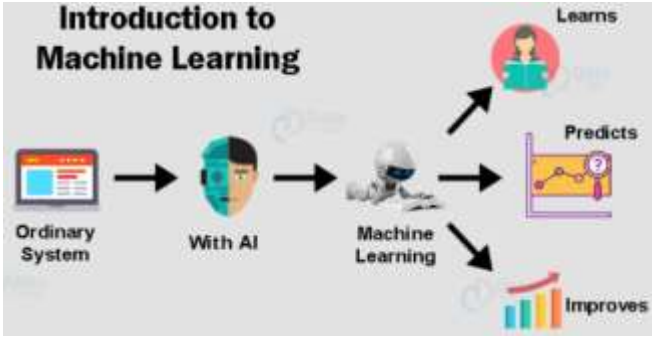

Fig-1: Machine Learning Processing [12]

2. DATA INTEGRATION

Data integration is the process of bringing data from disparate sources together to provide users with a unified view. The premise of data integration is to make data more freely available and easier to consume and process by systems and users. Data integration done right can reduce IT costs, free-up resources, improve data quality, and foster innovation all without sweeping changes to existing applications or data structures. And though IT organizations have always had to integrate, the payoff for doing so has potentially never been as great as it is right now.

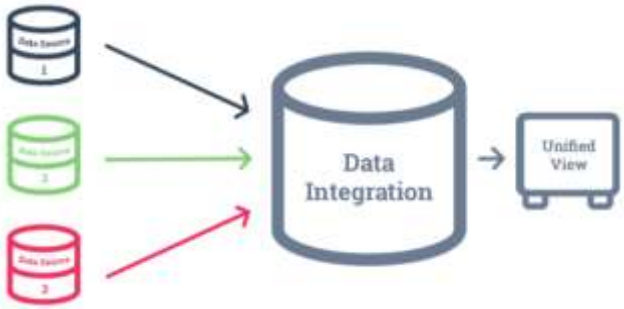

Fig-2: Data integration [13]

3. ELEMENTS OF MACHINE LEARNING

- **Data Set**

Machines need a lot of data to function, to learn from, and ultimately make decisions based on it. This data can be any unprocessed fact, value, sound, and image, text which can be interpreted and analyzed. A data set is a consolidated data of a similar genre that is captured in different environments. For example, a dataset of currency notes will have images of notes captured in different orientations,

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 8| Issue: 3| March 2022|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2022: 8.205 || ISI Value: 1.188

light, mobile cameras, and background so as to achieve maximum accuracy in notes classification and identification.

- **Algorithms**

Simply consider an algorithm as a mathematical or logical program that turns a data set into a model. There are different types of algorithms that can be chosen, depending on the type of problem that the model is trying to solve, resources available, and the nature of data. Machine learning algorithms use computational methods to "learn" information directly from data without relying on a predetermined equation as a model.

- **Models**

In Machine learning, a model is a computational representation of real-world processes. An ML model is trained to recognize certain types of patterns by training it over a set of data using relevant algorithms. Once a model is trained, it can be used to make predictions.

- **Feature Extraction**

Datasets can have multiple features. If the features in the dataset are similar or vary to a large extent, then the observations stored in the dataset are likely to make an ML model suffer from over fitting. Over fitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data.

- **Training**

Training includes approaches that allow ML models to identify patterns, and make decisions. There are different ways to achieve this including supervised learning, unsupervised learning, reinforcement learning.

4. TYPES OF MACHINE LEARNING

- **Supervised machine learning**.
  The most common form of machine learning, supervised learning involves feeding algorithm large amounts of labeled training data and asking it to make predictions on never-before-seen data based on the correlations it learns from the labeled data.
- **Unsupervised learning.**
  Unsupervised learning is often used in the more advanced applications of artificial intelligence. It involves giving unlabeled training data to an algorithm and asking it to pick up whatever associations it can on its own. Unsupervised learning is popular in applications of clustering (the act of uncovering groups within data) and association (predicting rules that describe data).
- **Semi supervised learning**.
  In semi supervised learning, algorithms train on small sets of labeled data and then, as in unsupervised learning, apply their learning's to unlabeled data. This approach is often used when there is a lack of quality data.
- **Reinforcement learning**.
  Reinforcement learning algorithms receive a set of instructions and guidelines and then make their own decisions about how to handle a task through a process of trial and error. Decisions are either rewarded or punished as a means of guiding the AI to the optimal solution to the problem.

5. LITERATURE SURVEY

In 2018, Ahmad et al. [11] performed an ad hoc literature review, by selecting 166 research papers on the SO that were mainly classified about software development life cycle from the start of the SO website till the year 2016 positively. Similarly, the work of Baltadzhieva and Chrupała [1] thoroughly reviewed and analyzed various questions quality posted on diverse community question answering (CQA) websites like SO. In 2013, Meth et al. [2] conducted an SLR on investigating the works on automated requirements elicitation. Later on, Binkhonain and Zaho [3] conducted an SLR on ML algorithms for identifying and classifying NFRs. Recently, Iqbal et al. [4] presented a survey on ML algorithms and requirements engineering. They provided a bird's-eye view of how ML algorithms are aiding different requirements engineering activities. Besides, there are some surveys, SLR's, and systematic mapping studies done in other areas on sentiment analysis of scientific citations [5], data preprocessing methods for class imbalance problem [6], ML algorithms or techniques based software development effort estimation models [7], usability in agile software development [8, 9], and requirements prioritization [10], among others.

6. ALGORITHMS IN MACHINE LEARNING

Machine Learning is the study of learning algorithms using past experience and making future decisions. Although, Machine Learning has a variety of models, here is a list of the most commonly used machine learning algorithms by all data scientists and professionals in today's world.

- Linear Regression
- Logistic Regression
- Decision Tree
- Bayes Theorem and Naïve Bayes Classification
- Support Vector Machine (SVM) Algorithm
- K-Nearest Neighbor (KNN) Algorithm
- K-Means
- Gradient Boosting algorithms
- Dimensionality Reduction Algorithms
- Random Forest

7. CHALLENGES FOR MACHINE LEARNING

- **Technological Singularity**
  This is also referred to as super intelligence, which defines as "any intellect that vastly outperforms the best human brains in practically every field, including scientific creativity, general wisdom, and social skills." Despite the fact that Strong AI and super intelligence is not imminent in society, the idea of it raises some interesting questions as we consider the use of autonomous systems, like self-driving cars. It's unrealistic to think that a driverless car would never get into a car accident,

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 8| Issue: 3| March 2022|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2022: 8.205 || ISI Value: 1.188

but who is responsible? But these are the types of ethical debates that are occurring as new, innovative AI technology develops.

- **AI Impact on Jobs**
  While a lot of public perception around artificial intelligence centers around job loss, this concern should be probably reframed. With every disruptive, new technology, we see that the market demand for specific job roles shifts.
- **Privacy**
  Privacy tends to be discussed in the context of data privacy, data protection, and data security and these concerns have allowed policymakers to make more strides here in recent years.
- **Bias and Discrimination**
  Instances of bias and discrimination across a number of intelligent systems have raised many ethical questions regarding the use of artificial intelligence. How can we safeguard against bias and discrimination when the training data itself can lend itself to bias? Bias and discrimination aren't limited to the human resources function either; it can be found in a number of applications from facial recognition software to social media algorithms.
- **Accountability**
  Since there isn't significant legislation to regulate AI practices, there is no real enforcement mechanism to ensure that ethical AI is practiced. The current incentives for companies to adhere to these guidelines are the negative repercussions of an unethical AI system to the bottom line. To fill the gap, ethical frameworks have emerged as part of collaboration between ethicists and researchers to govern the construction and distribution of AI models within society.

8. IMPLEMENTATION

A Decision Process: In general, machine learning algorithms are used to make a prediction or classification. Based on some input data, which can be labeled or unlabeled, your algorithm will produce an estimate about a pattern in the data.

An Error Function: An error function serves to evaluate the prediction of the model. If there are known examples, an error function can make a comparison to assess the accuracy of the model?

A Model Optimization Process: If the model can fit better to the data points in the training set, then weights are adjusted to reduce the discrepancy between the known example and the model estimate. The algorithm will repeat this evaluate and optimize process, updating weights autonomously until a threshold of accuracy has been met.

9. TOOLS PERFORMANCE

The following data illustrates machine learning tools with its implementation scope for the betterment of data integration for the distributed web information system.

1) Scikit-learn
2) PyTorch
3) TensorFlow
4) Weka
5) KNIME
6) Colab
7) Apache Mahout
8) Accord.Net
9) Shogun
10) Keras.io
11) Rapid Miner

10. RECENT TRENDS

- No-Code Machine Learning
- TinyML
- AutoML
- Machine Learning Operationalization Management
- Full-stack Deep Learning
- Generative Adversarial Networks
- Unsupervised ML
- Reinforcement Learning

11. APPLICATIONS

- **Web Search Engine**

One of the reasons why search engines like Google, bing etc. work so well is because the system has learnt how to rank pages through a complex learning algorithm.

- **Photo tagging Applications**

Be it Facebook or any other photo tagging application, the ability to tag friends makes it even more happening. It is all possible because of a face recognition algorithm that runs behind the application.

- **Spam Detector**

Our mail agent like Gmail or Hotmail does a lot of hard work for us in classifying the mails and moving the spam mails to spam folder. This is again achieved by a spam classifier running in the back end of mail application.

12. CONCLUSION

The detailed research survey in the field of distributed web information system towards data integration with traditional approach when compared to the machine learning approaches and recent techniques with advanced tools shows that the higher level of impact in the field of machine learning in the data integration for distributed web information system with the cope up towards latest trends and systematic pathways for the improvement progress of several advanced strategies. The approaches for machine learning dealt with the various levels of implications towards the selection strategies for the analysis and prediction of data integration for the distributed web information system implementations. The tools performance and applications of machine learning provides the several directions for the development of different

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 8| Issue: 3| March 2022|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2022: 8.205 || ISI Value: 1.188

methodologies to implement in the better way. In future this research will lead the direction of data integration in distributed web information system in an effective way.

REFERENCES

1. *A. Baltadzhieva and G. Chrupała, "Question quality in community question answering forums," ACM SIGKDD Explorations Newsletter, vol. 17, no. 1, pp. 8–13, 2015.*
2. *H. Meth, M. Brhel, and A. Maedche, "The state of the art in automated requirements elicitation," Information and Software Technology, vol. 55, no. 10, pp. 1695–1709, 2013.*
3. *M. Binkhonain and L. Zhao, "A review of machine learning algorithms for identification and classification of non-functional requirements," Expert Systems with Applications, vol. 1, 2019.*
4. *T. Iqbal, P. Elahidoost, and L. Lúcio, "A bird's eye view on requirements engineering and machine learning," in Proceedings of the 2018 25th Asia-Pacific Software Engineering Conference (APSEC), pp. 11–20, Nara, Japan, December 2018.*
5. *A. Yousif, Z. Niu, J. K. Tarus, and A. Ahmad, "A survey on sentiment analysis of scientific citations," Artificial Intelligence Review, vol. 52, no. 3, pp. 1805–1838, 2019.*
6. *H. Ali, M. N. M. Salleh, K. Hussain et al., "A review on data preprocessing methods for class imbalance problem," International Journal of Engineering & Technology, vol. 8, pp. 390–397, 2019*
7. *J. Wen, S. Li, Z. Lin, Y. Hu, and C. Huang, "Systematic literature review of machine learning based software development effort estimation models," Information and Software Technology, vol. 54, no. 1, pp. 41–59, 2012.*
8. *D. A. Magües, J. W. Castro, and S. T. Acuna, "HCI usability techniques in agile development," in Proceedings of the 2016 IEEE International Conference on Automatica (ICA-ACCA), pp. 1–7, Curico, Chile, October 2016.*
9. *D. A. Magües, J. W. Castro, and S. T. Acuña, "Usability in agile development: a systematic mapping study," in Proceedings of the 2016 XLII Latin American Computing Conference (CLEI), pp. 1–8, Valparaiso, Chile, October 2016.*
10. *P. Achimugu, A. Selamat, R. Ibrahim, and M. N. r. Mahrin, "A systematic literature review of software requirements prioritization research," Information and Software Technology, vol. 56, no. 6, pp. 568–585, 2014.*
11. *A. Ahmad, C. Feng, S. Ge, and A. Yousif, "A survey on mining stack overflow: question and answering (Q&A) community," Data Technologies and Applications, vol. 52, no. 2, pp. 190–247, 2018.*
12. *http://www.inspirationalladies.org/training/ai-machine-learning/*
13. *https://blog.hurree.co/blog/data-integration-in-marketing*