INTELLIGENCE EXTRACTION USING VARIOUS MACHINE LEARNING ALGORITHMS

Atul Mokal[1], Dipali Pawar[2], Mayuri Nikam[3], Varun Gaikwad[4]

[1]*Asst. Professor, Dept. of Computer Engineering, ISB&M School of Technology, Pune, Maharashtra, India*
[2,3,4]*Student, Bachelor of Engineering, Dept. of Computer Engineering, ISB&M School of Technology, Pune Maharashtra, India*

---***---

Abstract - *Intelligence Extraction is a process of arranging data proper manner by using some machine learning technique. Structured information is nothing but a data that can be easily understandable by human being or machine Unstructured information is opposite of structured information it is in dynamic form. So extracting structured information from this unstructured data is very tedious task.*

Key Words: **Intelligence Extraction, Structure Information, Unstructured Information.**

1. INTRODUCTION

Every day we process large amount of data so processing and analyzing this data which is unstructured is complex task. Millions of Documents is uploaded everyday on cloud and to handle those data we require system which is easy to handle, reliable, efficient, user friendly and through which we can get structured data.

World Wide Web is a central location in which data is stored and managed, so this organization which contain huge amount of information which is in the form of pdf, images, text, number, videos etc. from this huge data user wants only relevant data.

1.1 Intelligence Extraction

Intelligence Extraction is software system which extract data automatically. As we know machine learning is the field of artificial intelligence (AI) that uses statistical techniques to gives computer system the ability to learn from data without being explicitly programed

1.2 Extraction Techniques

There are three techniques of data extraction – Supervised, Unsupervised, Reinforcement learning algorithms. Supervised machine learning divided into two subgroups- **Regression** and **Classification.** Regression is the problem of predicting a continuous quantity and classification deals with assigning observation into different categories rather than predicting continuous quantities. Supervised Learning Algorithm is the machine learning task of finding a function that maps and input to an output based on examples input output pairs. For example-Decision table, Naïve Bayes, Support Vector Machine (SVM). Unsupervised learning algorithm is branch of machine learning that learns from test data that has not been classified or categorized.

Unsupervised machine learning divided into two subgroups- **Clustering** and **Association.** Clustering is technique that involves the grouping of data points. Association is rule-based machine learning for discovering relation between variables in the large dataset. For example- K-Mean, hierarchical etc. Reinforcement learning (RL) is an area of machine learning concerned with how software agents take action in an environment so as to maximize some desire. It is broadly divided into three types- **Q-learning**, **Temporal Difference**, **Deep Adversarial Network**.

2. Working

The system is shown in fig.1. This system uses various Extractions techniques to get structured data. As the extracted data may contain unwanted data which may not useful for the user. System works on four stages-

2.1 Data Collection

It collect data from different sources which is in the form of raw data which is difficult to maintain and modify over time.

2.2 Data Preprocessing

It involves transforming raw data into understandable format it contain four steps:-

A. Data Cleaning- it detects and correct inaccurate records from database.
B. Data Integration- it combine data into meaningful into valuable information.
C. Data Transformation- it convers data from one format or structure into another.
D. Data Reduction- it is process transforming numerical, digital information, into corrected, ordered and simplified form.
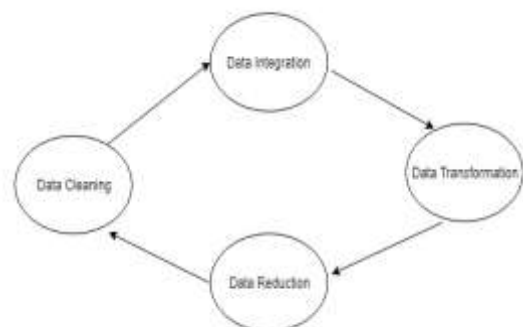

Fig -1: Data Preprocessing

2.3 Data Storage

If data is not stored properly can be lost, retrieving data is a costly proposition. The storage place is called memory data can be stored is measured in bytes if we store data systematically it can be easily understood we can process data quickly, it reduces number of errors. If we store data systematically we require less resources and it is easy to analyze.

2.4 Data Sets

There are two types of data sets:
A. Traditional: - In this data is stored in centralized location, single database and data is in structured format.
B. Big data: - The amount of data stored here is massive it can be from different sources. It is complex process here data can be in structured or unstructured format and it is stored at different location.

3. Working Steps

1) Take data from different Sources.
2) Apply data preprocessing technique.
3) Apply machine learning algorithm.
4) Traverse that data into Graph format.
5) Perform extraction to get data
6) Store that data in database.

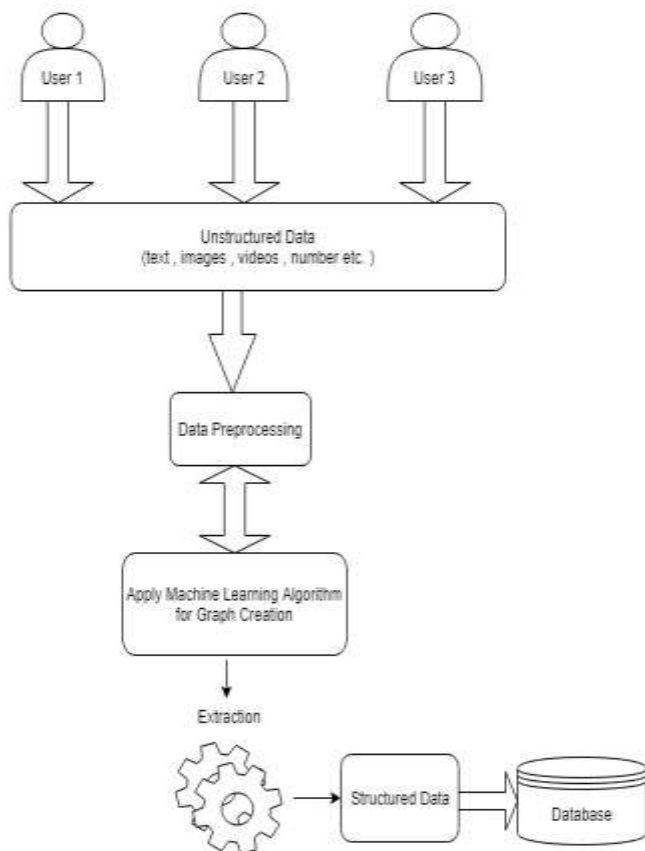**Fig -1**: System Working

4. Results and Discussion-

This system work on data collected from various users such as company, organization data etc. As user wants only relevant data so our proposed system categories that data so it can be easily accessible by user. We use different types of machine learning algorithms for extraction and categorizing that data into graph format. After we store that data into database or cloud for future use.

5. CONCLUSION

We implement a system in which extraction is based on the different machine learning algorithms which sort an unstructured data into structured format so it may be user friendly for user.

ACKNOWLEDGMENT

REFERENCES

[1] Vidya V L, "A Survey of Web Data Extraction Techniques", International Journal of advance research in computer science and management studies, vol. 2, Issue 9, Sep. 2014.

[2] Information Extraction on Novel Text using Machine Learning and Rule-based System , Ria Chaniago School of Electrical Engineering and Informatics Bandung Institute of Technology Bandung, Indonesia

[3] 2018 12th IEEE International Conference on Semantic Computing, Data Acquisition and Information Extraction for Scientific Knowledge Base Building Piotr Andruszkiewicz Institute of Computer Science Warsaw University of Technology Warsaw, Poland.

[4] S.S.Bhamare, Dr. B.V.Pawar" Survey on Web Page Noise Cleaning for Web Mining" International Journal of Computer Science and Information Technologies, Vol. 4 (6), 2013.

[5] Yanhong Zhai, Bing Liu," Web Data Extraction Based on Partial tree alignment", ACM 1-59593-046-9/05/0005.