ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 1, January 2020

Machine Learning Approach to Read and Process Old Data by XML Mechanisms

Rohit Raturi

Associate Director Enterprise Solutions, Regional Development Center, KPMG USA LP, Montvale, NJ, USA

ABSTRACT: Day by day the authors increased a lot as a result the number of books in paper format or digital format has increased a lot for same subject or concept. Keeping all the data in a proper order and keeping the data in the proper set became the challenging task. This paper mainly focuses regarding different applications at digital library. The main reason in selecting machine learning is it reads any type of data such as text, image, audio and video. For developing the project we are using XML and Wisdom ++ software.

KEYWORDS: HTML, XML, Deep Learning, Machine Learning, Document Image Mapping,

I. INTRODUCTION

At present the people reading books was reduced tremendously day by day because of reasons like books are not available in hard copy mode, few releases in hard copy mode etc., so many are going towards kindle for reading books. In kindle the users will have different benefits like can read few pages to see how the book is and can purchase, to borrow the book form other kindle friends or users and also different books are available in the kindle for different ages from rhymes books to Engineering, Medical, Law books etc., so lack of availability of hard copy books and benefits of kindle edition books many of the users were moving to kindle editions. Machine learning has a unique procedure for data collection which we can see in below figure

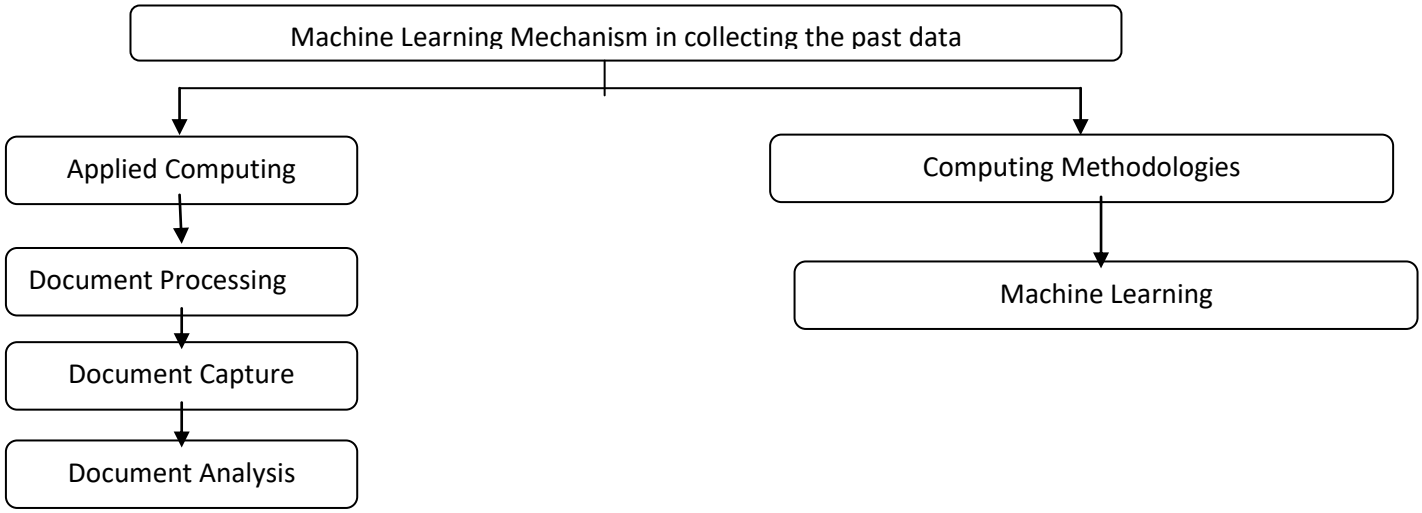

Figure 1: ML Procedure for collecting Historical data or information

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 1, January 2020

Many of the old books like at the time of 1920's which are in library were in the position of losing the papers by touching them because of old books and lost its strength in paper. [1] To read such kind of books it's not possible because those books were not given to users like normal books so the valuable information in the books were not utilizing by the present generation. So by identifying the problem many of the old books were making digitalized for making available of that particular data to all the users. For making digital we are using XML and Wisdom++ tool for data typing and storing.

II. LITERATURE SURVEY

The procedure of generating of the information gathering was first started by using Natural Language Programming (NLP). [2] The NLP was used in digitalizing the information in the form of documents etc., . The NLP is also used in the case of Data collection, Data association and identifying the unstructured data. This procedure is also used in the collection of the large amount of information over digital library. Here the author designed the Key Phase Identification Program (KIP) [3] which works as a database for the key pairs for matching the sets of data. The key phase database helps in build the database based on the keywords in the textbook based on the keywords and matching terms it makes the reverent database so that it makes efficient for data retrieving and data storing makes more efficiently.

In the internet we have the large amount of data so identifying and making the data is a heavy task. To overcome this problem the concept of machine learning and pattern matching came into existence. The best mechanism is document summarization. Text summarization takes more appropriate text and proceeds.

To overcome the problem of text classification and identification in 2015 Fuzzy logic [4] came into existence. This algorithm used for text classification and semantic approach. This mechanism is used for extracting data from various sources in large databases. Later by using Fuzzy logic and HTML came into existence. [5] Recently the Fuzzy Logic mechanism came into existence for flexibility and easy usage. The fuzzy logic was designed with the help of XML instead of HTML due to various advantages over XML instead of HTML.

III. PROPOSED APPROACH

For the development we are using many applications such as Wisdom ++, fuzzy logic and XML programming. Compared to HTML over XML having many advantages such as:

- HTML is Pre-defined where as XML is a user defined tags.
- HTML is user side script where as XML is server side script.
- HTML is less secure where as compared to XML.

And for developing the project like typing the data instead of Microsoft office we are using Wisdom++ tool because of various reasons such as

- Sharing
- Concurrent editing
- Storing history of past typed data or files even we delete them
- User can comment on particular line by the edit option
- Managing

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 1, January 2020

The main agenda of this concept is to make search more efficiently and user friendly. This technique generally used for searching and generating results more efficiently and speedily. The results are generated by the help of browsing, texting, based on key words, based on repeated phases of data. The architecture of the project was designed based on the knowledge map. The ontology is a concept where the machine learning and E-learning concepts came into existence like recognition of images, text, audio and video by NLP programming. The concept of the data identification was classified into following steps namely:

The overall thing of classification approach is classified into following phases such as:

- Complete block registration.
- Data gathering.
- Data search.
- Information Retrieval.
- Generate metadata and
- Classifying data to the document.

Complete block registration [6]

Here the overall image is taken as a unit and the image is considered as a block and the read operation takes place for the process of text searching. Here the two types of classes were defined namely character class and word class. Automatic identification of book by its cover is only possible by machine learning with (Support Vector Machine) SVM algorithm. These algorithms mainly used in the pattern identification and paper detection.

Data gathering [7]

Generation of information from the raw data is one of the procedures of Machine Learning. The generated information is kept at servers and we can communicate with servers by the help of commands or queries. Metadata is provided by the digital library for the generation of the data. Some of the mechanisms and techniques are using for the data extraction and generation and making the data in a structured manner. The data is collected from web of sciences , Scopus articles , science articles.

Data search [8]

Different mechanisms of machine learning is using in this concept for the making the document in a most defined manner. The first order digital format of text books were designed in this order only. Machine learning is used for automatically search for the data from the database which makes easy for users to retrieve form bulk of databases.

Information Retrieval [9]

Generally the retrieval is possible for only the textual data dependent on the search key word. The mechanism in identifying of books, volumes and chapters was designed by textual manner only earlier now when machine learning came into existence the text search and image search also made possible in the year of 2007. The image search can be done by identifying the cover page of the book. Based on image processing it identifies the exactness of the image and retrieves to the user. The image recognition is presently using for the extraction of scientific documents.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 1, January 2020

Generate metadata [10]

This procedure is used for generating or extracting data from the large sources. Even though it extracts data from the large sources we can guarantee the data with accurate and with good quality. Some of the mechanisms such as Bayesian framework[13,14]

Classifying data to the document [11]

Machine Learning procedures used mainly for differentiating of documents identification which are on the table.[15,16,17 Different authors classify tables based on the author name, year of publication, volume number, issue number, ISSN number etc, here the author classify the tables based on the author details and year of the publication of the book. For making the archive of files of different size, different files etc., were needed. For data collection and data storing like Microsoft word document the same editor named Wisdom++ was introduced. This wisdom++ makes the document in processing automatic manner. By utilizing the positive unstructured learning methodology mechanism was proposed especially for document identification and definition.

Security approaches in this research domain [12]

Machine Learning concept is used in providing privacy for the data which is in the digital library. The mechanism of Machine Learning and Deep Learning helps in identifying the any type of malicious insiders and keeps data more secure. Some authors proposed some techniques for making data more resolve the privacy risks. They have developed the project with PDF documents more secure.

Figure 2 shows how the same data is stored in the same place with same key words.

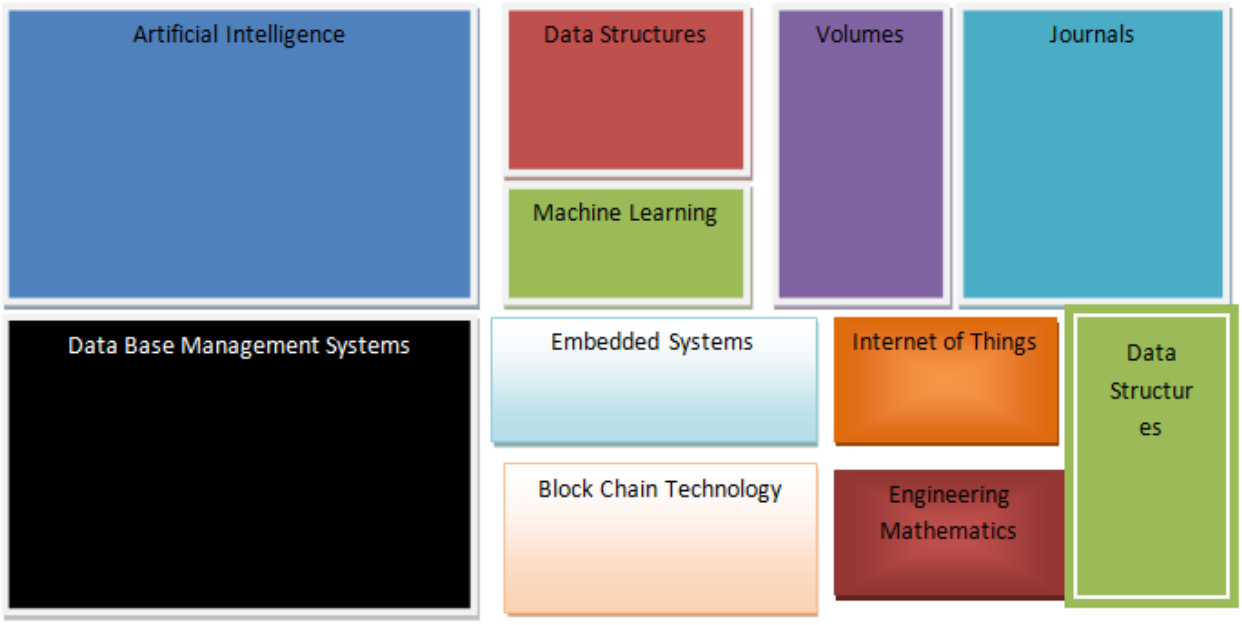

Figure 2: Making similar data at one place based on key words

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 1, January 2020

IV. RESULTS

This paper focuses on the how data is stored in the digital format in the digital libraries. With improving the soft copies instead of hardcopies so that it makes available for all the users. The making of digitalized format is only possible with Artificial intelligence and Machine learning. This technology helps a lot in defining, classifying, retrieving and extracting the data.

This concept in the network is done in the different citations such as web of science, Scopus with related to artificial intelligence and machine learning. The terminology of network definition and departments were defined by the different technologies. The definition and the network analysis are done in the digital manner.

The experimental results of digitalization and normal manner were shown in the graph. Here the graph was designed based on the past information up to 6 years in using digital books and normal books readers. When the Amazon kindle came into existence the usage of soft copies were increased

The results obtained by generating the project by making the data available as soft copy is as follows:

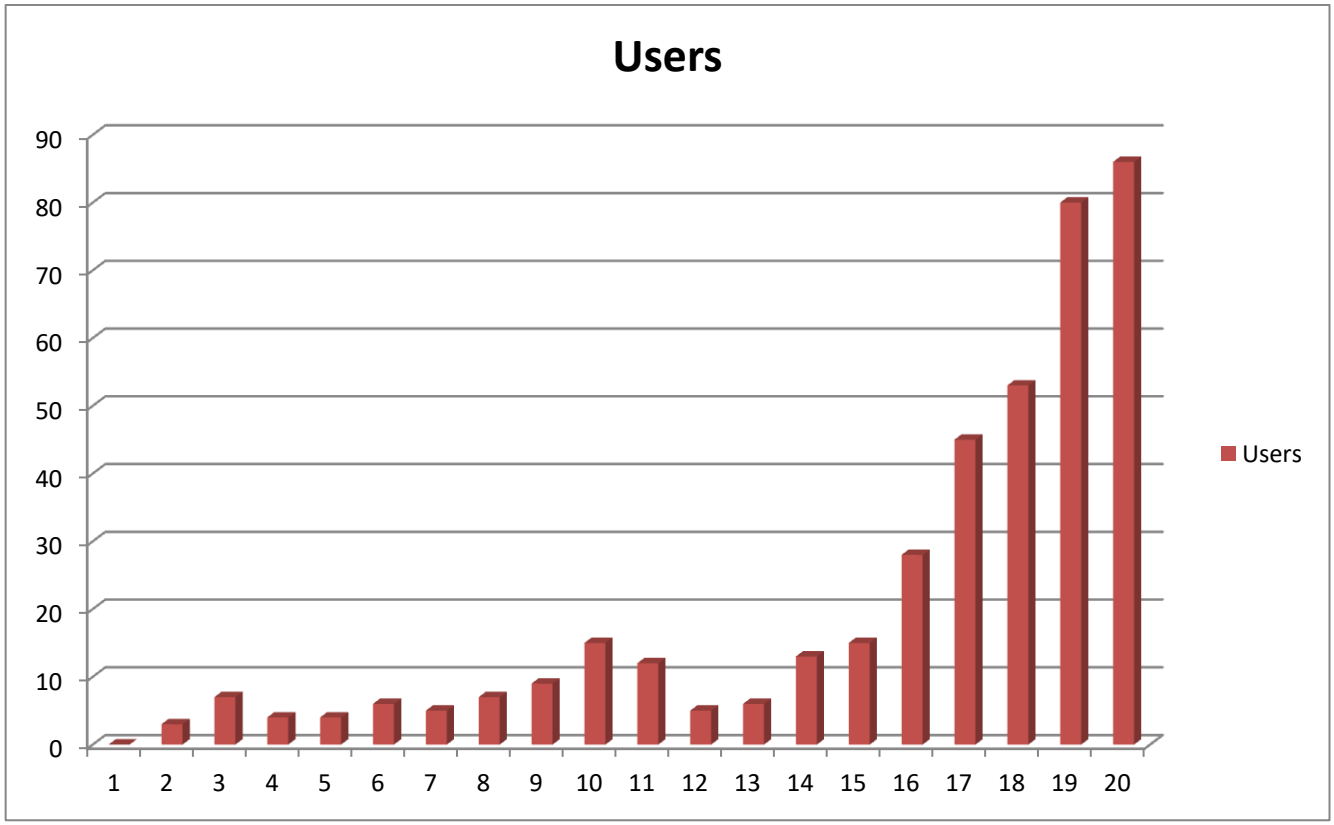

Figure 3: Graph showing the increasing the usage of digitalized books

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 1, January 2020

Here the above graph shows the improvement in usage of digital books reading by people from past 8 to 9 years. The average users searching and retrieving books is 9 per minute at the time of 2019.

Overall from this concept by making final analyzing of results we can say that due to people increase in number so the resources, things etc., demand was increased in larger ratio. Instead of printing and making the resources like books etc., by making the things digitalization the resources can be shared and helpful for many of the people. By bringing out and implementing the task we can observe in the graph that the usage of people reading Ebook was increased in larger number.

V. CONCLUSION AND FUTURE WORK

Overall in this paper we have presented the mechanism in designing document using Wisdom++ software in the digital libraries. By giving input automatic processing of data and extraction of the required data is formed. The main agenda of our work is to collect and make historical data in digitalized manner to available to all people to read the data and gain knowledge. Here the different methods and techniques are made for the developing of the project by digitalized manner using Machine Learning. Here we have used Image Processing technology for identification of the image. While transforming the data there are few problems in reading the data which is in the image format. Even though two major problems were left over such as the Wisdom++ algorithm doesn't works perfectly over the colored images it is working on normal images and the other is we embedded the light with OCR by the help of Wisdom++ software for identifying as well as for interaction purpose.

REFERENCES

[1] S. A. Babar and P. D. Patil, "Improving Performance of Text Summarization", Procedia Computer Science, vol. 46 no. 2015, pp. 354 – 363, International Conference on Information and Communication Technologies(ICICT2014), doi: 10.1016/j.procs.2015.02.031.
[2] F. Ciravegna, S. Chapman, A. Dingli, Y. Wilks , "Learning to Harvest Information for the Semantic Web", pp. 312-326, 2004, doi: 10.1007/978-3-540-25956-5_22.
[3] S. Mukherjee, I.V. Ramakrishnan and A. Singh, "Bootstrapping Semantic Annotation for Content-Rich HTML Documents, Proceedings of the 21st International Conference on Data Engineering (ICDE 2005). 1084-4627/05.
[4] Y.L. Chi, T.Y. Hsu and W.P. Yang, "Building Ontological Knowledge Bases For Sharing Knowledge In Digital Archive", Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005, IEEE. 0-7803-9091-1/05
[5] Swathi Rathi,(2019) Attribute Evaluation of Dataset Using Chi-Square Test in Rapid Miner Studiovolume 7, issue 5,IJIRCCE SSN(Online): 2320-9801 ISSN (Print): 2320-9798
[6].] Swathi Rathi (2019), Estimating and Diverting the Route by Using Traffic Data over Big Data vol8,issue 9, IJIRCCE ISSN(Online): 2319-8753 ISSN (Print): 2347-6710
[7] Abhishek.et.al (2020) A novel hybrid approach of SVM combined with NLP and probabilistic neural network for email phishingInternational Journal of Electrical and Computer Engineering 10(1):44600DOI: 10.11591/ijece.v10i1.pp486-493
[8] Esposito F., Malerba, D., Semeraro, G., Fanizzi, N., Ferilli, S. (1998). Adding Machine Learning and Knowledge Intensive Techniques to a Digital Library Service. *International Journal on Digital Libraries*, 2(1): 1-17, Springer Verlag, Berlin.
[9] European-Union: The european year of cultural heritage 2018
[10] Grilli, E., Dininno, D., Petrucci, G., Remondino, F.: From 2d to 3d supervisedsegmentation and classification for cultural heritage applications. ISPRS - InternationalArchives of the Photogrammetry, Remote Sensing and Spatial InformationSciences XLII-2, 399–406 (2018).
[11] Kurniawan, H., Salim, A., Suhartanto, H., Hasibuan, Z.A.: E-cultural heritageand natural history framework: an integrated approach to digital preservation. In:International Conference on Telecommunication Technology and Applications. pp.177–182. Proc .of CSIT vol.5, IACSIT Press (2011)
[12] Li, J., Ding, J., Yang, X.: The regional style classification of chinese folk songsbased on gmm-crf model. In: Proceedings of the 9thInternational Conference onComputer and Automation Engineering. pp. 66–72. ICCAE '17, ACM, New York,NY, USA (2017).
[13] R. T. Mylavarapu and B. K. Mylavarapu, "Huge information extraction techniques of Data Security," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, 2018, pp. 179-183.
doi: 10.1109/ICICCT.2018.8473017
[14]B. K. Mylavarapu, "Implementing Machine Learning in Intensive Care Units: For Avoiding Over Medication," (2018) International Journal of Pure and Applied Mathematics, Volume 118 No. 20 2018, 4799-4811
URL: https://acadpubl.eu/hub/2018-118-21/articles/21f/33.pdf
[15] R. T. Mylavarapu, "A Method for Approximated Deep Learning Towards Dynamic Sharing from Big-Data Analysis," 2018 International Conference on Research in Intelligent and Computing in Engineering (RICE), San Salvador, 2018, pp. 1-6. doi: 10.1109/RICE.2018.8509060
[16] Sreenivas Sasubilli, Kumar Attangudi Perichiappan Perichappan, P. Srinivas Kumar, Abhishek Kumar, An Approach towards economical hierarchic Search over Encrypted Cloud, pages 125-129; Annals of Computer Science and Information Systems, Volume 14. ISSN 2300-5963.
[17] Gopinadh Sasubilli,Uday Shankar Sekhar, Ms.Surbhi Sharma, Ms.Swati Sharma, "A Contemplating approach for Hive and Map reduce for efficient Big Data Implementation" 2018 Proceedings of the First International Conference on Information Technology and Knowledge Management pp. 131–135 DOI: 10.15439/2018KM20