International Journal of Engineering &Technology, 7 (3.12) (2018) 89-91

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Deep Learning and Indian Heritage

B.S.Charulatha[1], Arun Rajaraman[2]

[1]Associate Prof, [2]Retd Professor, Rajalakshmi Engineering College, IIT Madras
**Corresponding Author Email:[1]charu2303@yahoo.co.in, [2]arraman41@yahoo.com*

Abstract

Deep learning is becoming increasingly necessary as data, information and web and mobile interaction are proliferating with form, type and volume of data becoming easy to store, retrieve and process. This necessity is felt not only in science and engineering but also in social and commercial internet activity. The paper explores some ideas of deep learning to process large data in the context of content extraction in a heterogeneous web page containing multi-lingual information, in the Indian context. This basis is used to explore how Indian heritage over several thousands of years, is able to maintain information and knowledge in certain areas through oral, palm leaf and stone-cut data forms. The paper extends the proposed method to some heritage data and concludes how complex from computing point of view to extract the knowledge and essence of some texts like Bhagavad Gita.

Keywords: *content extraction, deep learning, Indian heritage, web pages.*

1. Introduction

The evolution of data from simple alphanumeric form to media form due to proliferation in communication and mobile technology, is becoming a complex and at the same time interesting research area. This is due to the explosive increase in web and mobile interaction in science, engineering, social and commercial activities. This has resulted in generation of large amounts of data in various forms from simple alphanumerical form to media form with language and location becoming immaterial. With so much of data coming in, methods to extract information and then knowledge started evolving leading to expert and smart computing. But these are not able to cope up with the volume, size, type and timing of data being generated from different sources [1] and new techniques like cloud computing and deep learning-a modified form of AI and soft-computing techniques started emerging. This paper is a study on development of a method [2] exclusively for handing Indian document forms to extract content from a web document and later attempting to get a hold on Indian heritage [3] and how content and knowledge have been passed on over several centuries.

2. Pixel-Based Content Extraction

Data in Indian context is quite complex, varied and multi-lingual and with mobile phone users increasing day by day, information and content extraction require an approach different from conventional data mining approaches and needed to account for oral and text variations in different language. Fig.1 gives an idea of the sources and forms of input data that is being generated and typically a document may be uni-lingual, multi-lingual, unstructured and heterogeneous. So a methodology is proposed using pixel-based representation so that processing is based on images of documents.

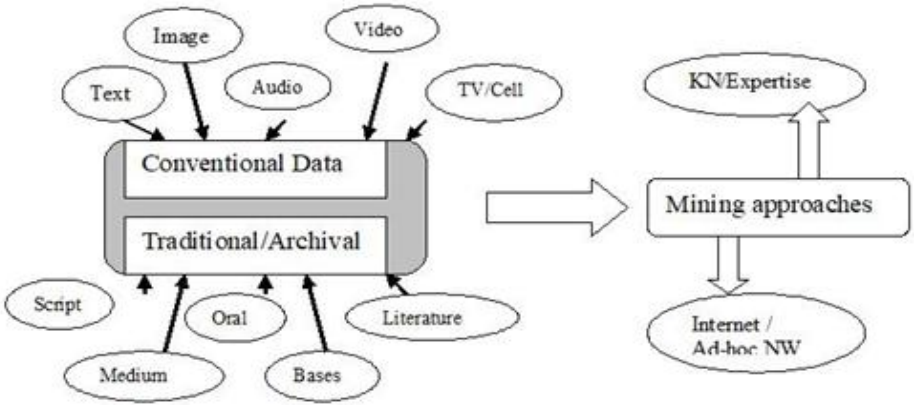


Fig. 1: Data forms -Indian context

Typically, a word or sentence or image is taken in a jpg form and converted to a 2D matrix using standard image processing means. The matrix is used to create different attributes using numerical processing and these attributes are used in clusters -given the

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

name key clusters-and ANN and statistical processing are used to extract content. Words can range from letters to texts that indicate object, subject and even abstract forms as shown in Fig.3.

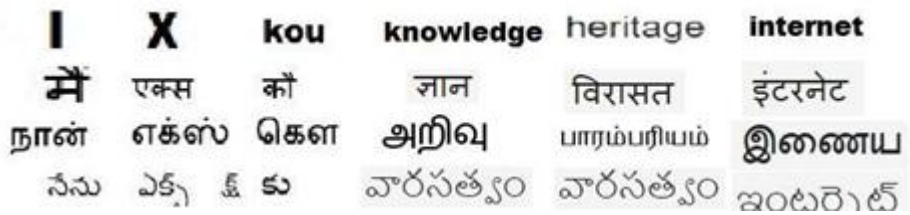


Fig. 2: Typical words in Indian context

After pixel processing and extracting attributes [3,4], bar charts of the wide variation are shown in Fig.3 along with patterns with basis English texts. ANN processing also can be done and Fig.4 gives a trend of base and trained sets.

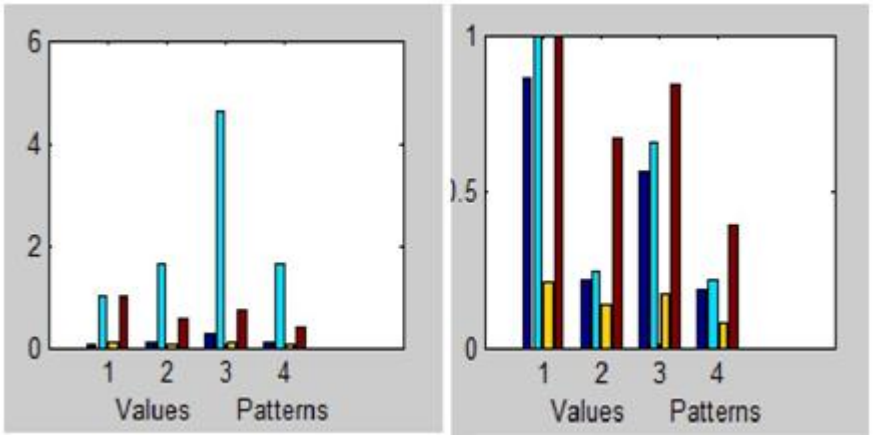


Fig. 3: Attribute values and patterns in four languages

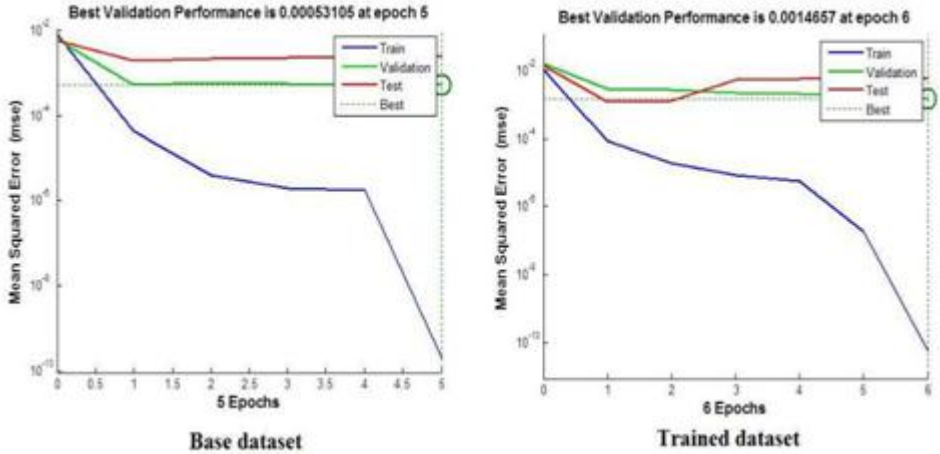


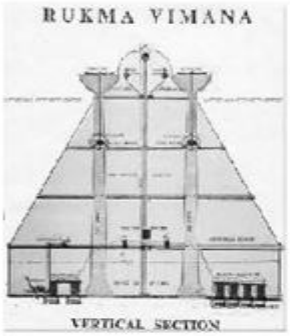
Fig. 4: ANN analysis for chosen attributes

3. Deep learning and Concept of key Clusters

The idea of pixel map processing is also extended to developing key clusters so that language and text anomalies can be taken care of [4] and some case studies on these aspects have been reported earlier [5,6] with applications different kinds of data having image language and text variations. These form the basis for studying archival and other forms of documents, and features of complexities in Indian heritage forms are discussed first.

4. Indian Heritage and Deep Learning

Indian heritage and tradition belong to one of the few thousand year-old civilizations existing as early as 2200 BC as evidenced in Mohanjadaro-Harappa excavations. Besides knowledge and expertise levels available at that time are very well documented or passed on orally through generations [7] so that data available is enormous, varied and complex. To bring out this, Fig.5 and 6 give two instances the first one on flying and the second one in chemistry.



Vega-Saamyaat Vimaano Andajaanaam. Sootra 1.

Lallachaarya says, "That which can fly in the sky with speed equal to that of birds, is called Vimaana."

Aachaarya Naaraayana says,

"That which can speed on earth, on water, through air, by its own power, like a bird, is a "Vimaana."

Shankha says, "Experts in the science of aeronautics say, "That which can fly through air from one place to another is a Vimaana"

And Vishwambhara says, "Experts say that that which can fly through air from one country to another country, from one island to another island, and from one world to another world, is a "Vimaana"."

a) Image of Vimana b) Translated text of Vimana
Ref: http://www.hinduwebsite.com/sacredscripts/hinduism/vimana/vchp01.asp
Fig. 5: Traditional data on aircraft

Similarly texts are done on stone as shown in Fig.6 (a) and here this is for an incomplete structure and how it needs to be rebuilt. In another amazing Indian text , Fig.6(b), AgastryaSamhita gives the precise directions for constructing electrical batteries.



*"Place a well-cleaned copper plate in an earthenware vessel. Cover it first by copper sulfate and then moist sawdust. After that put a mercury-amalgamated-zinc sheet on top of an energy known by the twin name of Mitra-Varuna. Water will be split by this current into Pranavayu and Udanavayu. A chain of one hundred jars is said to give a very active and effective force." –*Agastya Samhita (Indian Princes' Library)

a) Stone cutting texts b) Palm leaf text extractions
Fig. 6: Traditional text data in different forms

Besides these type of data on heritage, there are customs and practices passed on over the years through generations and these are in the form of oral and image forms in art and culture and these vary region to region and even in one region different language and dialects. Different dance forms and paintings like in Ajanta and Ellora are typical examples**.**

Availability in web and computer image forms is being done in the last decade and the examples chosen are from such instances. Now the complexities of these types of data are in medium of availability-stone, palmleaf or oral- lingual variations-same language in different dialects- and heterogeneity in that sketches and forms of objects, present in documents. Another major feature which calls for deep learning is in the different interpretations of content. The famous image or letter or text used in various forms shown in Fig.7.



Fig. 7: Script complexity in heritage

This neither belongs to lingual or image classification and the implication is well known to many Indians and Asians. If one uses the present method and relates to an abstract content, the document

nature can be obtained to extract the features of these web documents. Another example is the famous Bhagavad Gita, which is a narration of what one sees virtually-gnaana dhrishti- without giving his interpretation unless asked for! That is why so far many bhaashyams-content extraction of the Gita- are in vogue starting from Aadi Sankara. If one has to extract content in a computer, deep learning techniques are needed, as direct translations do not give what is implied. An example is on indriyas-faculties- and they are used as gnaanendriya , karmendriya etc. meaning the use and application of indriyaas and some abstract words like bhakthi, punya, bhutham, etc need an in-depth understanding calling for use of soft-computing as well. Some studies are initiated by classifying into object, subject, abstract, logic groups and using these as clusters different interpretations given are taken as input sets so that computer-based understanding of Gita is possible

5. Summary

The study presents a pixel-based approach for content extraction with emphasis on Indian web pages having multi-lingual and heterogeneous characteristics. Later after presenting the complexities of Indian heritage, proposal to use the approach with development of different clusters for different content interpretations is made with a follow-up needed for deep learning.

References

[1] Li, Y, C, C, Kuo, J, Wan, X : Introduction to content-based image retrieval — Over view of key techniques, in Image Databases: Search and Retrieval of Digital Imagery, eds. , JohnWiley, New York, pp. 261–284(2002)

[2] Bhanu Prakash, Kolla, Dorai RangaSwamy, M, A, Raja Raman, Arun : ANN for Multi-lingual Regional Web Communication, ICONIP 2012, Part V, LNCS 7667, pp. 473-478(2012)

[3] Bhanu Prakash, Kolla, Dorai RangaSwamy, M, A, Raja Raman, Arun , Performance of Content Based Mining Approach for Multi-lingual Textual Data, International Journal ofModern Engineering Research, Vol.1, No. 1, pp. 146-150 (2011)

[4] B.S.Charulatha, Paul Rodrigues, T.Chitralekha, Arun Rajaraman Mining Ambiguities using Pixel-based Content Extraction, Advances in Intelligent Systems and Computing DOI 10.1007/978-81-322-2674-1,Springer India (2016)

[5] B.S.Charulatha, Paul Rodrigues, T.Chitralekha, Arun Rajaraman Clustering for knowledgeable web mining, Advances in Intelligent Systems and Computing 324,DOI 10.1007/978-81-322-2126-5_54,Springer India (2015)

[6] B.S.Charulatha , Thesis titled Generic and Novel Approach for Web Content Mining for Indian Languages, July 2016.

[7] Rajaraman,A, Paradigm changes needed for Infra structural systems in Indian context, talk in NIT, Tiruchy, January 2017.