Indian Journal of Science and Technology, *Vol 9(30), DOI: 10.17485/ijst/2016/v9i30/95283, August 2016*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

ANN Models and their Implications in Content Extraction

B. S. Charulatha[1*], Paul Rodrigues[2] and T. Chitralekha[3]

[1]JNTUK, Kakinada – 533003, Andhra Pradesh, India; charu2303@yahoo.co.in
[2]King Khalid University, Saudi Arabia
[3]Central University, Kalapet – 605014, Puducherry, India

Abstract

Objectives: Internet is the repository of information, which contains enormous information about the past, present which can be used to predict future. To know the unknown users are inclined towards searching the internet rather than referencing the library because of ease of availability. This requirement initiates the need to find the content of a web page with in shortest period of time irrespective of the form the page is. So information and content extraction need to be at a basic generic level and easier to implement without depending on any major software **Methods:** The study aims on extraction of information from the available data after the data is digitized. The digitized data is converted to pixel- maps which are universal. The pixel map will not face the issues of the form and the format of the web page content. Statistical method is incorporated to extract the attributes of the images so that issues of language hence text-script and format do not pose problems, the extracted features are presented to the Back Propagation algorithm. **Findings:** The accuracy is presented and how the content extraction within certain bounds could be possible Tested using unstructured word sets chosen from web pages. The method is demonstrated for mono lingual, multi-lingual and transliterated documents so that the applicability is universal. **Applications/Improvement:** The method is generic, uses pixel-maps of the data which is software and language independent.

Keywords: Back Propagation, Content Extraction, Information, Statistical, Deterministic

1. Introduction

The basic requirement of traditional IR is that the query and the documents need to be in the same language. In traditional IR when the document content is present in a language other than English, is considered as “noise”[1]. Digitization of the documents is gaining popularity. Hence multi linguistic web pages are increasing in the repository. Due to this issue of noise the traditional IR cannot cope up for the retrieval of multi linguistic content. Hence introduction of new area to extract information irrespective of the language is the need. The solution is a variant of IR like Cross Lingual Information Retrieval (CLIR) and Multilingual Information Retrieval(MLIR). The working principle behind CLIR and MLIR is translation. The translation can be done either “to the query or document or both”. In case of query the query is translated to the language of the document. In case of document translation, the document is translated or both the query and the document can be translated. “The basic three methodology used for translation are a) Dictionary based b) Machine Translation and c) Parallel Corpora”. Query translation uses the method of (a) and (b) where as Document translation uses method (c).

Figure 1 describes the features of data in web and mobile communication. The current emphasis is on English. But unfortunately the web pages have multilingual format and content. In addition, the web pages have multimedia content also. Hence the web pages are heterogeneous due to the hybrid content present in the documents and unstructured nature due to language and or media which implies that there is no well-defined syntactic position.

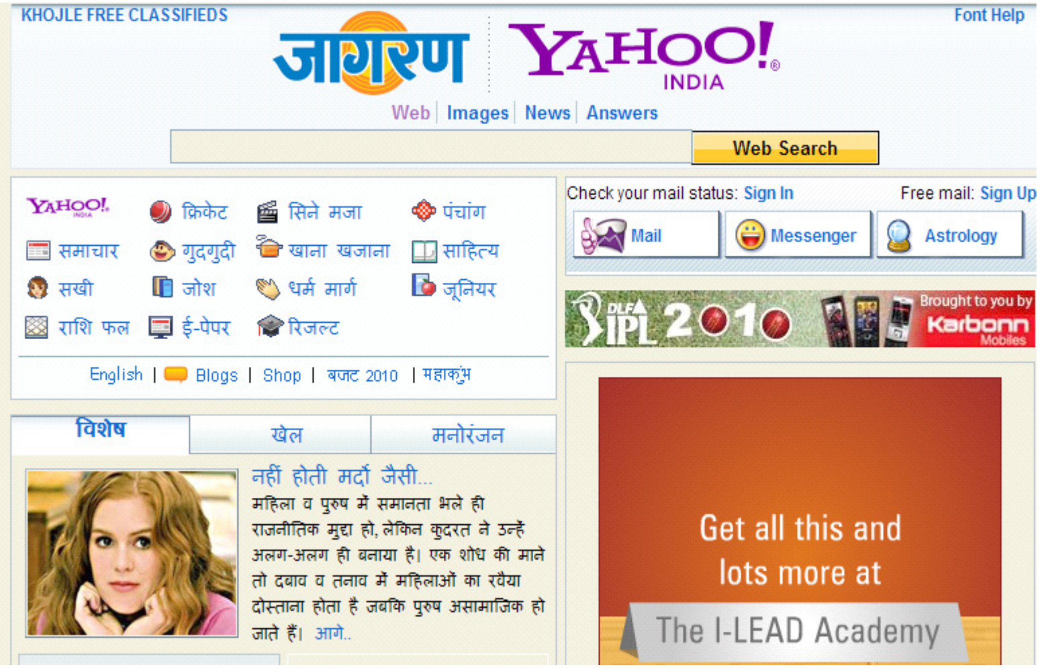
*Author for correspondence

Web pages now-a-days have different forms and types of content. In terms of form they may have pictures, video and audio files and text files in multi languages. Confining attention only to text files which are multi-lingual, many approaches exist which try to translate using different methodologies to arrive at the content. But at a web level, user may be interested in overall content rather than translation so that the user can pursue with the web page if need arises. This extraction has to be done in a short period of time as web page lives are short-lived! The present study is aimed at this content mining when the web page contains English, Sanskrit and Tamil. Here again the web image may contain texts, which are multilingual and has multimedia contents[2,3].

But even with these variations the web page may relate to an area or a subject and here an attempt is made to extract the content without translation so that a generalization of this approach is possible. The aim of the present work is to extract attributes found in the documents in Indian languages restricting with English, Sanskrit and Tamil. Alternate method to translation is proposed. Method should preferably be computer-understandable and not software-dependent. Pixel-based processing is proposed to assess the overall content is the focus of the study. The approach is based on converting to basic pixel-maps, so that issues of language, text-script and format do not pose problems, content is extracted with ease so that it is universal. Statistical and deterministic features of the images are extracted from the pixel map matrix of the image[4]. The extracted features are presented to the Back Propagation algorithm and studied for performance.

2. Content Extraction

From the web pages the content extraction is nothing but extracting the needed information after eluding the noisy data. The noisy data is that of advertisements, hyperlinks etc. The methodologies used to extract the content to name a few are using Tree, Wrapper and density graph. In the Tree method the DOM tree of the HTML page is transformed into a Block Tree. Each node has a vector associated with it. To find the content, the tree traversal is done. To find main block having the content vector values of the block nodes are compared, dominant is returned[5]. In the wrapper methodology, the wrapper is generated by wrapper induction system for an information source, only one wrapper is generated[6]. In this paper the authors, proposed a method to create a text density graph from the HTML webpage. The content is the region with the highest density[7]. For satellite image processing traditional methods use pixel for classification. The drawback of this method is that, it does not use the spatial information. Hence pixel groups were used to classify the data got from the satellite. This method uses the colour and spectral



नहीं होती मर्दों जैसी...

महिला व पुरुष में समानता भले ही राजनीतिक मुद्दा हो, लेकिन कुदरत ने उन्हें अलग-अलग ही बनाया है। एक शोध की माने तो दबाव व तनाव में महिलाओं का रवैया दोस्ताना होता है जबकि पुरुष असामाजिक हो जाते हैं। आगे..

Figure 1. Exact nature of web page.

attribute[8]. “In [9] author presents thresholding algorithm to be used in pixels for Optical Character Recognition System for Brahmi Script. The types of thresholding algorithms are

i) Global thresholding algorithms: A single threshold for all the image pixels is used. When the pixel values of the components and that of background are fairly consistent in their respective values over the entire image, global thresholding could be used.

ii)Local or adaptive thresholding algorithms: Different threshold values for different local areas are used.

Hence a method more generic and independent of language form or media type is needed. The proposed method is different from conventional mining approaches like translation and easily computer- understandable and software- independent and the approach uses the basic representation of data as pixel-maps. Internally data are represented only by pixels. Hence the pixel representation is considered for processing. Since pixel-maps are large datasets different reduction methods to extract features and attributes are considered using salient features of the pixel matrix, which may be binary or grayscale or even colored. The pixel map of the image is of higher dimension. Hence dimension reduction is done to the pixel map of the image using the nonzero elements in the matrix representation. The features extracted are the statistical properties of the images. Now the pixel map of the text or image is represented by the various attributes like mean, standard deviation, Eigen value, determinant, diagonal, rank and norm. Normalization of the attributes is done to so that data dependence of the attributes is taken care of [10]”.

3. DataSet

The method is tested with actual content and how far they are effective in extracting the content. The deterministic words that represent the cat like meaw, 9lives, purr, cuddle are taken. The statistical features are extracted using the pixel map of the image. The ambiguous word set of cat is chosen like ‘fur’, ‘fourlegs’, ‘brown’ and ‘claws’ for which linear membership of +/- 10% is considered. The fuzzy dataset like ‘clean’, ‘pet’, ‘soft’ and ‘cunning’ with linear membership with +/- 15% in linear distribution is considered. The fourth as a combination of ambiguous and fuzzy data sets are considered. Table 1 represents the words taken under various categories.

Table 1. Words taken in various cadre

	Ambiguous	Fuzzy
Meaw	Fur	Clean
9lives	Fourlegs	Pet
Purr	Brown	Soft
Cuddle	Claws	Cunning

Table 2 represents the sample statistical features of the chosen words. Nine attributes are found out of which only one is shown in the table. The attribute shown in Table 2 is that of the pixel mean. The pixel mean is normalized in order to take care of data dependence.

Table 2. Sample numerical equivalent for the words taken

Words	PixelMean
Meaw	0.9979516
9lives	0.997943
Purr	0.9972719
Cuddle	0.998435
Fur	0.9982177
Fourlegs	0.9988695
Brown	0.9986654
Claws	0.99714
Clean	0.9972261
Pet	0.9926979
Soft	0.9943133
Cunning	0.9952956

As the input data and the content they refer to do not have any direct relation neural model is used for establishing the relation and the algorithms used is back propagation learning algorithm. The algorithm adjusts the initially randomized set of synaptic weights so as to maximize the difference between the network’s output of each input fact and the output with which the given input is known to which it belongs to. For this application a three layered back propogation network is developed with “input layer, hidden layers and output layer”. The learning rate η is taken as 0.9. Error function used is RMSE.

The algorithm of Back propagation is taken from [11] and implemented.

This model is used for four datasets and one hidden layers and sigmoid functions.

4. Result

In [12] the author has taken the data set for the Indian Languages like Tamil, Telugu and Hindi. In this working

English language is taken. For testing a paragraph from the web page http://bigcatrescue.org/domestic-cat-facts/ is taken. This a clear one like 'cat' words describing cat are given in Table 1[10]. It is represented only using English Language. A web page is chosen and subjected to removal of tags, special characters, stop word removal and stemming algorithm. Using the words chosen from the domain the web page content can be classified using the Back Propagation method. The output values for nearly 53 words the graph looks alike that as shown in the Figures 1 to 3 where variation of output with x-axis being the words and the y axis being the guessed output are shown under various scenarios. In Figure 2 label reaches the guessed output from lower values while in Figure 3 it is from higher values showing clearly that more training is needed. In Fuzzy case of Figure 4 the choice of equivalent memberships using linear correlation gives same values from the beginning. Hence additional training is needed. The results are shown below:

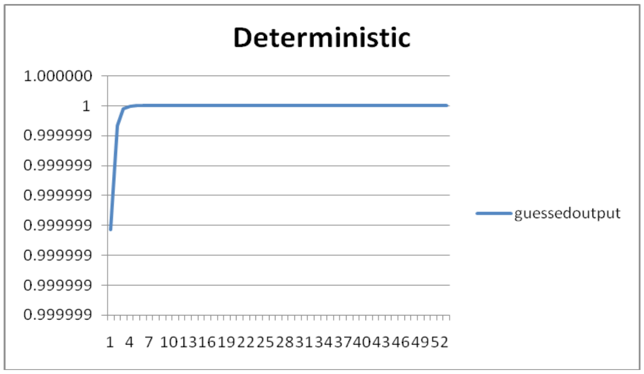


Figure 2. Guessed output for deterministic data.

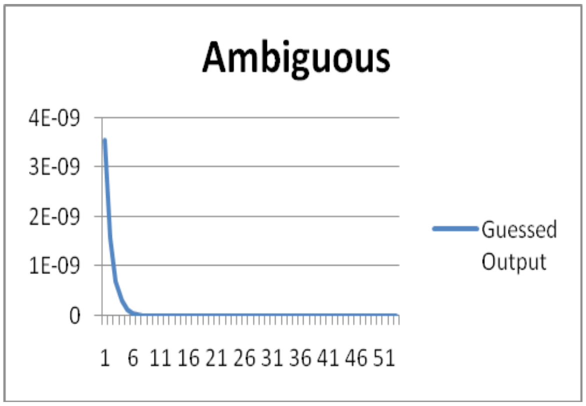


Figure 3. Guessed output for ambiguous data.

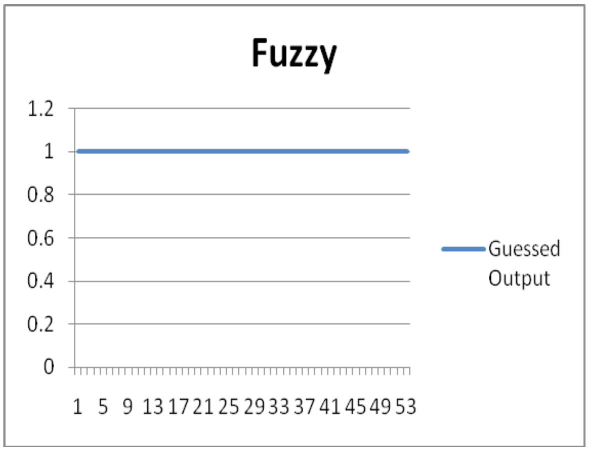


Figure 4. Guessed output for Fuzzy data.

5. Conclusion

The method described earlier is used with pattern recognition to compare whether any new input is related to the existing patterns. The proposed technique is data driven and not domain dependent. This method is generalized and is based on pixel maps of images or words. This concept can be used to determine the content in the web pages. The results are being satisfactory hence can be tested with different web pages. The idea can be implemented with multilingual web page with heterogeneous content. The images are restricted to black and white while the text is restricted to computer generated ones. This needs to be extended to various other images and hand written formats. The algorithm used is Backpropogation with transformation being Sigmoid function and RMSE is used to calculate the error. It can be extended to multilayered network and various other transformation functions.

6. References

1. Abusalah M, Tait J, Oakes M. Literature review of cross language information retrieval. World Academy of Science, Engineering and Technology. 2005; 4:175–7.
2. Charulatha BS, Rodrigues P, Chitralekha T. Automatic and adaptive clusters for information extraction. International Conference on Soft Computing and Machine Intelligence, New Delhi: India; 2014. p. 60–3.
3. Charulatha BS, Rodrigues P, Chitralekha T, Rajaraman A. Heterogeneous clustering. International Conference

in Information Communication and Embedded Systems ICICES; 2014.

4. Charulatha BS, Rodrigues P, Chitralekha T, Rajaraman A. Mining ambiguities using pixel-based content extraction. Proceedings of the International Conference on Soft Computing Systems; 2015 Dec 8. p. 537–44. DOI: 10.1007/978-81-322-2674-1_50.
5. Asfia M, Pedram MM, Rahmani AM. Main content extraction from detailed web pages. International Journal of Computer Applications. 2010; 4(11):18–21.
6. Sirsat S. Extraction of core contents from web pages. International Journal of Engineering Trends and Technology (IJETT). 2014; 8(9):6.
7. Arias J, Deschacht K, Marie-Francine M. Language independent content extraction from web pages, 9th Dutch-Belgain information retrieval workshop, Enschede: The Netherlands; 2009.
8. Xiaoxia S, Jixian Z, Zhengjun L. A comparison of object-oriented and pixel-based classification approachs using quickbird imagery. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science. 2010; XXXVIII, Part 8, 1–3.
9. Devi HK. Thresholding: A pixel-level image processing methodology preprocessing technique for an OCR System for the Brahmi Script. Journal of the Society of South Asian Archeology; 2006.
10. Charulatha BS, Rodrigues P, Chitralekha T, Rajaraman A. Clustering for knowledgeable web mining. A springer International Conference on Advances in Intelligent Systems and Computing, ICAEES; 2014. p. 491–8.
11. Mitchell T. Machine learning. Tata McGraw-Hill Education India; 2013 May 01.
12. Prakash KB, Rangaswamy MAD, Ananthan TV, Rajavarman VN. Information extraction in unstructured multilingual web documents. Indian Journal of Science and Technology. 2015; 8(16):1–8. DOI: 10.17485/ijst/2015/v8i16/54252.