ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(22), DOI: 10.17485/ijst/2016/v9i22/95165, June 2016* ISSN (Online) : 0974-5645

Content Extraction Studies using Neural Network and Attribute Generation

Kolla Bhanu Prakash[1,2] and M. A. Dorai Rangaswamy[1]

[1]Faculty of Computer Science Engineering, Sathyabama University, Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai - 600119, Tamil Nadu, India; bhanu_prakash231@rediff.com, drdorairs@yahoo.co.in
[2]Faculty of Computing, Chirala Engineering College, Ramapuram Beach Road, Chirala - 523157, Andhra Pradesh, India; drkbp1981@gmail.com

Abstract

Objectives: The amount of information available on web today is more than at any point in history, and greater challenges arouse due to this huge wealth of information available. Also to deal with this information overload, challenging tools are required. **Method of Analysis:** Internet in the present day especially in India is spreading both in rural and urban areas. Bilingual and Multilingual websites are increasing to a larger extent. Even websites are becoming multitasking. Our main problem is to deal with multilingual web documents and ancient documents. Because, content extraction becomes difficult when such documents are considered. The present paper proposes a neural network approach and attribute generation to justify the content extraction studies for multilingual web documents. **Findings:** Results obtained are well defined and a thorough analysis is done. **Novelty/Improvement:** The method is versatile in using pixel-maps, analytically stable in that the matrix input is used and is demonstrated for adoption to different models.

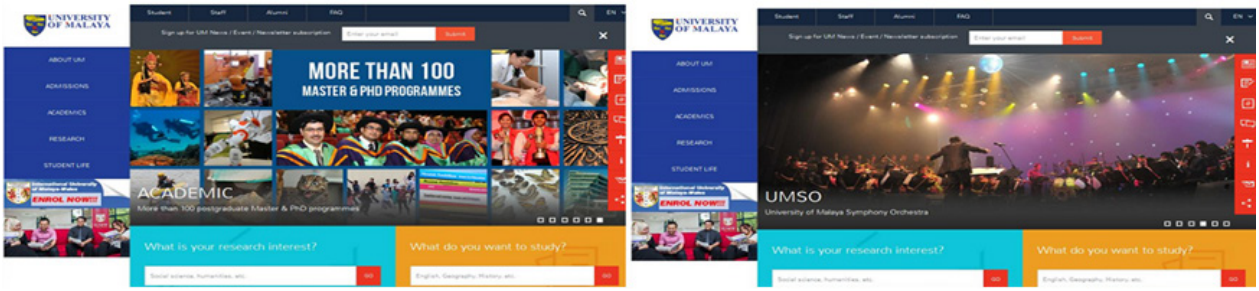
Keywords: Attribute, Content Extraction, Mining, Multi-Lingual, Neural Network, Pattern, Pixel

1. Introduction

Recent developments in communication and internet have brought in significant changes in scientific, engineering and societal context and wide range of user-oriented mobile applications like whatsapp, twitter etc. have added new dimension to modern living and thought process. Simultaneously, the reach of these developments is still a long way to go as long as the gap between human communication and computer-based communication is not bridged fully. There are many barriers to overcome like language, dialect, tradition, way of living etc. This is where; conventional data mining approaches need to be elevated to media-mining or content extraction approaches. Content extraction in general is the process of identifying main content of a web document which may consist of different forms of data in unstructured and non-homogeneous manner[1,2,3].

Added to this is the ability of including region and language based information, thanks to the exponential growth in use of cellular communication. Text based information has reached different levels with different languages forming the text either as a computer-generated data or acquired data through images forming most of the pages. All these aspects bring in a necessity of using a more general approach to extraction of information and it has become very important to consider different kinds of web pages. A typical web page in present day context is shown in Figure 1 and both relate to University of Malaya web site with most of the text and form remaining same except that the pictures on left are static while the one on the right is dynamic-temporal-with flicker of lights captured. These pictures make the document unstructured with varying contents. The picture on left gives the general features while the right one highlights a specific event and to know more about this, the user can play it to get an idea. A different type of unstructured

* *Author for correspondence*

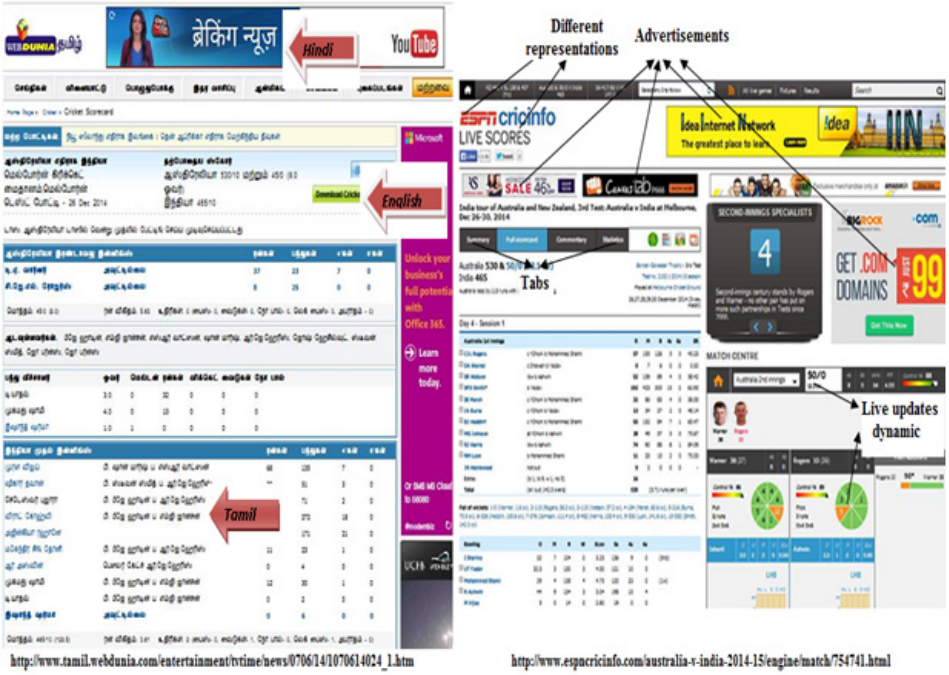


http://www.um.edu.my/

Figure 1. Variations in form, text and language levels in web page.

web document related to a particular game is shown in Figure 2 in a particular regional language on the left and the same in English is shown on the right.

The web page in Figure 2 has text-based information in two different languages–content may or may not be just translated one- and also different kinds of images which may be a photo or computer-generated drawings. This web page conveys information in the form of content even to a person who does not know any of the languages as the images convey more than the text. So, content extraction is a very important technique used commonly nowadays in text mining and information extraction from web documents, both online and offline[4,5,6]. The focus of the present study is to develop a generic content extraction approach which is based on the unstructured, non-homogeneous and text and/or non-text based data, as that of the web page shown in Figure 2. This is a major difference to be looked into when one considers Asian web pages, which contain language and information, which are older than those used in European web pages and this aspect gets much more complex in Indian context, where dialect and text differ widely even in small regions[7]. The present study is an attempt to develop a pixel-based approach-which gives flexibility in dealing with any language or media- and start from generic text level to a hybrid unstructured level.



a) Regional language web page

b) English language web page

Figure 2. Web documents dealing with same discipline.

2. Content Extraction Issues

As content extraction is different from text or data mining, where a set of keywords form the basis, and most of the earlier studies[2,5,11,15] use HTML tags, xml and DOM parsing procedures to separate the main content from other entities or add-ons. Each Non-English language has an equivalent ASCII and Unicode conversion[8,9,10]. Even techniques like VISP[15] can only give details regarding webpage layout structure, but this technique cannot judge the content of the webpage, when a multilingual webpage is given as input.

A collage of data in the form of text in different languages and sizes, numerals, images and blocks, forms the web page with the intent that content is reached to the web-surfer, who may be from different country with different languages and dialect and culture. This is typically an unstructured, heterogeneous and hyper media web page. So, a generic model without using wrapper and employing basic features of data is needed and the proposed model is from basic pixel level making it applicable to any kind of data or text or image or even media to assess the content in quick time.

3. Nature and Features in Web Documents

Typically, a modern web page for commercial intent looks like the one shown in Figure 3 and this is taken from web for the same day 20th Jan, 2015. The contents are different as can be seen from the images and content of the page is essential for one to pursue either in Hindi or Arabic. Here one can notice use of English words as they are like 'cricket' or in regional language text without translation. As seen earlier web pages are unstructured-not conforming to any document form-, non-homogeneous with information and data presented in different forms from text to images to video, and multi-lingual depending on the audience and their location. This gets more complex and involved when Asian or Indian regional web pages display information.

Figure 2 shows variation in content in web page on the same day which occurs due to region and language.

Figure 3 shows a typical web page displaying news on the same day and here web pages in Tamil and Hindi are shown. Even if one looks at script or character level, or even word level, complexities are many-fold, as the web pages try to present information in easily understandable form using words freely from different languages. As an example, a word 'magnet' in English translated in other languages like Hindi, Tamil and Telugu is shown in Figure 4(a). But many times, popular words in one language are used as they are like word 'magnet' in English is written in local scripts as in Figure 4(b).

Figure 4. Complexities in Indian and Foreign languages with English.

Figure 3. Modern day news web page in two languages on the same day. Modern day news web page in two languages on the same day.

So, sometimes it is necessary to assess the content irrespective of the language used and the manner in which the text is produced[13].

4. Text and Character Issues

One of the basic steps in any content extraction or mining approach is in processing the data as it is. So, a pixel map of any dataset can form the basis for any form or format of data as computer processes at this level[14]. But in unstructured and non-homogeneous documents, complexities begin at character level and later extend to word or document or web page form.

Figure 5 shows a Physics web page in two different languages English and Tamil used in schools. Here one can see free mixing of words in Tamil and English in both forms of documents. The present work aims at developing a generic tool based on pixel map data, to extract content in a web page and later, using reduced attributes and features of pixel maps, a pattern matching approach is used to assess the content.

5. Development of Pixel Map Attributes

A web document may contain texts, images, audio/video files; and in some regional documents, scanned copies of hand-written texts or images are found. So, it is necessary to look at the generic level of data which is used by computer for processing[12]. Here an overall organization of the proposed model is presented as flowchart in Figure 6; where input preparation followed by attribute generation; algorithm usage and content extraction are the major segments.

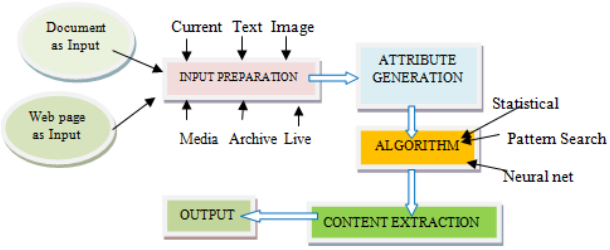
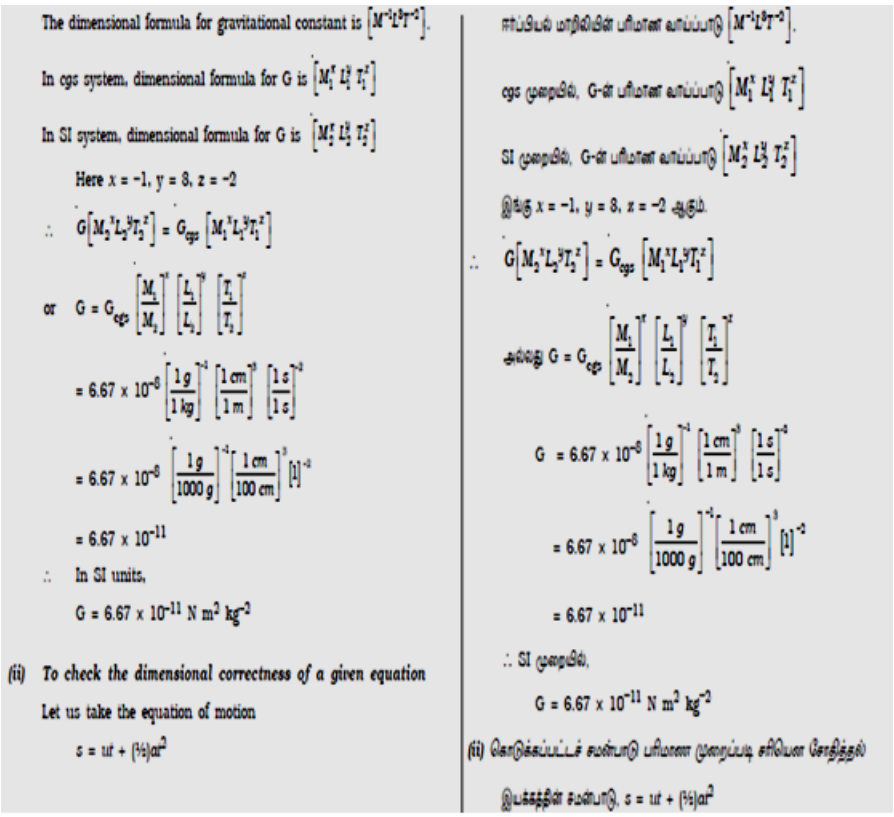


Figure 6. Overall organization of the proposed method.


The dimensional formula for gravitational constant is $[M^{-1}L^3T^{-2}]$.

In cgs system, dimensional formula for G is $[M_1^x L_1^y T_1^z]$

In SI system, dimensional formula for G is $[M_2^x L_2^y T_2^z]$

Here $x = -1$, $y = 3$, $z = -2$

$\therefore G[M_2^x L_2^y T_2^z] = G_{cgs}[M_1^x L_1^y T_1^z]$

or $G = G_{cgs}\left[\frac{M_1}{M_2}\right]^x \left[\frac{L_1}{L_2}\right]^y \left[\frac{T_1}{T_2}\right]^z$

$= 6.67 \times 10^{-8}\left[\frac{1\,g}{1\,kg}\right]^{-1}\left[\frac{1\,cm}{1\,m}\right]^3\left[\frac{1\,s}{1\,s}\right]^{-2}$

$= 6.67 \times 10^{-8}\left[\frac{1\,g}{1000\,g}\right]^{-1}\left[\frac{1\,cm}{100\,cm}\right]^3[1]^{-2}$

$= 6.67 \times 10^{-11}$

$\therefore$ In SI units,

$G = 6.67 \times 10^{-11}\ N\ m^2\ kg^{-2}$

(ii) To check the dimensional correctness of a given equation

Let us take the equation of motion

$s = ut + (1/2)at^2$

ஈர்ப்பியல் மாறிலியின் பரிமான வாய்ப்பாடு $[M^{-1}L^3T^{-2}]$.

cgs முறையில், G-ன் பரிமான வாய்ப்பாடு $[M_1^x L_1^y T_1^z]$

SI முறையில், G-ன் பரிமான வாய்ப்பாடு $[M_2^x L_2^y T_2^z]$

இங்கு $x = -1$, $y = 3$, $z = -2$ ஆகும்.

$\therefore G[M_2^x L_2^y T_2^z] = G_{cgs}[M_1^x L_1^y T_1^z]$

அல்லது $G = G_{cgs}\left[\frac{M_1}{M_2}\right]^x \left[\frac{L_1}{L_2}\right]^y \left[\frac{T_1}{T_2}\right]^z$

$G = 6.67 \times 10^{-8}\left[\frac{1\,g}{1\,kg}\right]^{-1}\left[\frac{1\,cm}{1\,m}\right]^3\left[\frac{1\,s}{1\,s}\right]^{-2}$

$= 6.67 \times 10^{-8}\left[\frac{1\,g}{1000\,g}\right]^{-1}\left[\frac{1\,cm}{100\,cm}\right]^3[1]^{-2}$

$= 6.67 \times 10^{-11}$

$\therefore$ SI முறையில்,

$G = 6.67 \times 10^{-11}\ N\ m^2\ kg^{-2}$

(ii) கொடுக்கப்பட்டச் சமன்பாடு பரிமான முறைப்படி சரியென சோதித்தல்

இயக்கத்தின் சமன்பாடு, $s = ut + (1/2)at^2$

Figure 5. Text book page in two languages – English and Tamil- for class 12.

Any pixel map can be seen as a matrix of columns and rows with each element giving the colour scheme for the pixel. Mathematically a grey-scale image is a mapping of a subset D_f of the domain Z^n into a finite matrix of non-negative integers as in eq.1 below:

$$f : D_f \text{ in } Z^n \rightarrow \{0,1,2,..t_{max}\} \quad (1)$$

with $t_{max} = 2^n - 1$ for pixels as n bits

Image transformations could be done using a matrix of threshold operators T between t_i and t_j as

$$[\,T\,]\,(f)\,]\,(\mathbf{x}) = 1 \quad (2)$$

for $t_i <= T_{i,j} <= t_j$ or 0

Attributes can be generated with the transformed matrix by identifying pixel values and their pattern. So, the characteristic and attribute of any pixel map can be deduced from these three values and most of image processing and data mining techniques depend on this basic matrix[12]. The matrix size being large, it is preferable to reduce it by converting into greyscale or binary form giving 0-7 or 0-1 values in the matrix. Typically a letter 'a' in English has [18 x 16 x 3] matrix and this is reduced to [10 x 11] with 0 and 1 value and even then there are 288 values to reflect the matrix fully. Figure 7(a) gives pixel map attribute variations for letter 'a' in English in the form of bar charts. Attributes normally relate to the pattern and arrangement of pixels which form the image or text and here only the non-zero pixels give us the character.

These could be the number or arrangements in the matrix and so a reduced representation could be in many ways and here five different types as

- Mean and standard deviation for all rows.
- Only number of pixels having non-zero values.
- Number of pixels in three segmented rows.
- Equivalent 2 x 2 matrix and
- Reduced 3 x 3 matrix.

For a typical text character 'a' in English these attributes with absolute values – all less than 288- are shown as bar charts in Figure 7(a). The attribute variation gives us clearly the picture of focus of our study, how importance is pixel map manipulation. Mean and standard deviation gives us the prominent details for statistical interpretation studies. The same could be done for words and Figure 7(b) gives the attributes for a word 'magnet'. Here the attributes are normalised with area of the pixel map so that all are less than 1. After careful analysis study is made, we observed 2x2 and 3x3 attributes gave more information to interpret the core findings of the study.

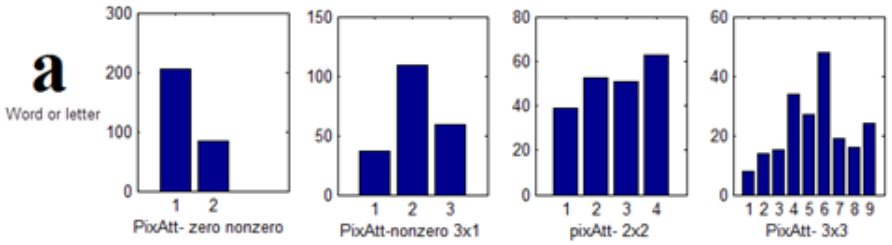


Figure 7. (a). Variation of pixel map attributes for letter 'a'.

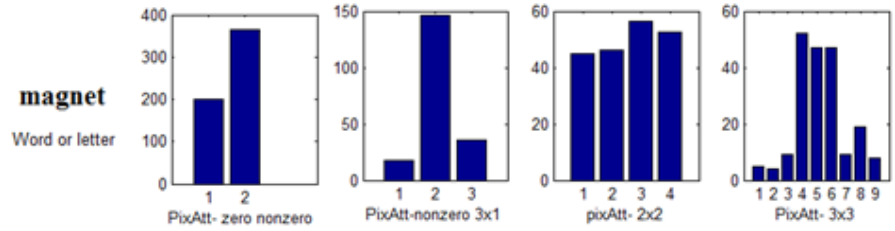


Figure 7. (b). Variation of normalised pixel map attributes for word 'magnet'.

In comparing Figures 7(a) and 7(b) we can discuss clearly how attributes vary for a character like 'a' and a word like 'magnet'. Even here it was observed 2x2 and 3x3 attributes give good sufficient information for analysis. Now these variations could be used to compare letters or words to check for similar content irrespective of type of font or script or language. Figure 8 gives a comparison of features of pixel map attributes for letter 'a' in English, Hindi and Tamil, all normalised with area of pixel map to get consistency.

This figure when we compare with standard base normalised figure, we can clearly predict whether it belongs to the same subject of interest or not. The advantage of using normalised results is we can remove certain inconsistencies if we find in intensity or zero and non-zero nature of pixel-map. In the next section the attributes are used in a neural model so that content extraction for any web document relating to a discipline can be developed.

6. Results and Discussion

As explained in overall flowchart Figure 6, different algorithms can be used for assessing the content in a web document once the attributes are generated. One method is to look at the attributes and use a pattern search to relate the content. Here once a certain pattern is identified, depending on a pre-assigned threshold, any new pattern can be identified to belong to the basic set or not. Now for the letter 'a' base patterns could be generated in different languages using different scripts. The proposed technique is purely data driven and does not make use of domain dependent background information, nor does it rely on predefined document categories or a given list of topics. Character 'a' which is unique in content, similar in many languages – Hindi, Telugu, Tamizh and English. Uniqueness of letter 'a' is that, it has same meaning or content in all the above mentioned languages.

Figure 9 gives the histograms for attribute variations of six chosen words 'magnet', 'diamagnet', 'dipole', 'filings', 'moment' and 'monopole' in English. In comparing these attribute variations; the first chart gives values for non-zero values only after normalising with pixel area. Similarly other charts show values of normalised non-zero values for the other types. In matching the pattern with the values given, the content is easily predicted. After observing carefully it is found that in all the attributes, first two and last words fall into a unique similar pattern and third and fourth into another similar pattern. So, if a threshold is fixed like any of those two patterns observed, whenever a new word or dataset is considered, its results are obtained with the similar procedure and in comparing with this pattern we can conclude that the new dataset considered falls into the same content or not.

Figure 10 gives a comparison of all the attributes normalized with 'diamagnet' in English for 9 words [diamagnet, filings, dipole in original, translated and transliterated form]. Since all the features are normalized, this comparison gives better indication of the content. Any discrepancies observed in general attributes can be nullified when we consider discussing with normalised attributes.

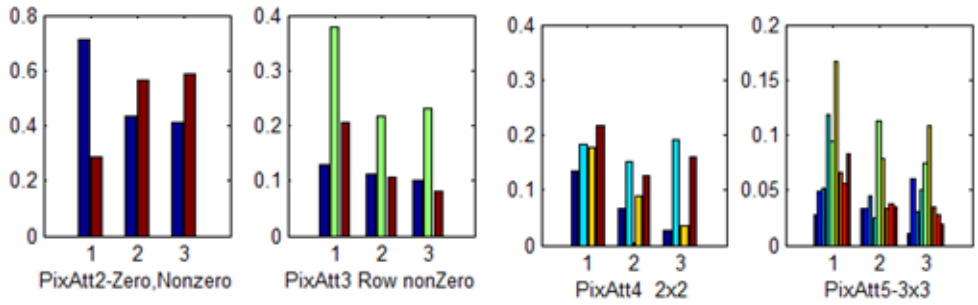
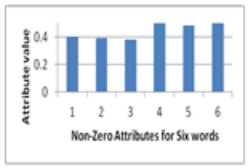
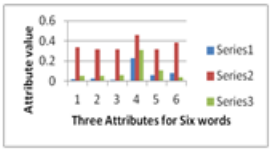


Figure 8. Variation of Normalised pixel map attributes for 'a'.

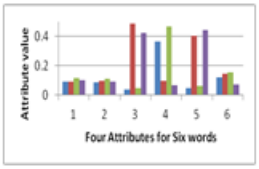
0.4028
0.3908
0.3808
0.9977
0.4813
0.5



0.0165	0.3335	0.0528
0.0268	0.3142	0.0498
0.0077	0.313	0.06
0.2285	0.4615	0.3077
0.059	0.3187	0.1036
0.0777	0.3841	0.0381



0.0915	0.0918	0.1176	0.102
0.0881	0.0958	0.1133	0.0936
0.0365	0.4898	0.047	0.4267
0.3643	0.095	0.4683	0.0701
0.0489	0.4054	0.0629	0.4449
0.1227	0.1471	0.1578	0.0724



0.0034	0.0142	0.0885	0.0747	0.0705	0.0198
0.0063	0.0055	0.0885	0.0747	0.0729	0.029
0.0068	0.0071	0.0862	0.0792	0.0705	0.029
0.1136	0.1013	0.177	0.1493	0.1585	0.122
0.1125	0.1067	0.177	0.1493	0.164	0.1334
0.1074	0.1062	0.1724	0.1629	0.1585	0.1288
0.0239	0.0115	0.0708	0.0995	0.0881	0.0076
0.0176	0.0235	0.0708	0.0995	0.0911	0.0168
0.0114	0.0148	0.069	0.1086	0.0881	0.0137

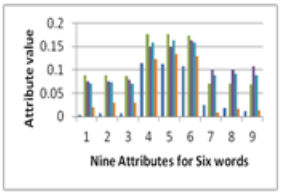


Figure 9. Normalised Attribute variations for base dataset.

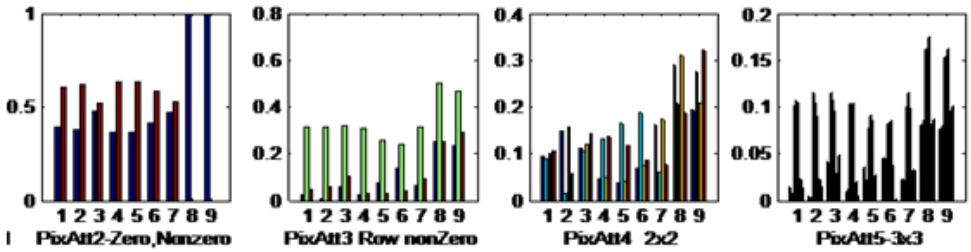


Figure 10. Normalised Attribute variations with base pattern for 'diamagnet'.

These normalized values for all four types of attributes are shown in Figure 10 and one can see clearly that as the size of attribute increases from scalar to vector and then onto matrix, proximity to the actual matrix increases. Similarly statistical interpretation using distribution functions can also be used and already results for the two approaches have been presented[13,14] and here results for ANN are highlighted. Neural modelling emulates capabilities of biological neural systems and consists of simple processors 'neurons' and a connection network, which adjusts itself suitably through the strengths of connections and can adapt to new environments.

An MLP-multi layer perceptron - is used here and studies for different attributes as input patterns are carried out so that the net gets trained properly. The input vector or matrix could be the different attributes developed earlier and here results using input attribute matrix [3x3] are presented. Figure 11 shows comparison of performance for base and trained datasets. [6words related to magnetism and science respectively].

Figure 11 gives the neural network plot for comparing 3x3 attribute variations. It is observed that for 'Base dataset' in the fifth iteration the best validation performance is attained and for the 'Trained dataset', in the sixth iteration the best validation performance is attained.

Figure 12 gives attribute variations for 'Base dataset' compared with 'Trained dataset' using neural network fitting plot, the function fit for output element 1 shows the input and output targets along with errors observed. To overcome from errors network was trained using perceptron learning algorithm. The objective of this training is to reduce the error, which is the difference between perceptron output and target vector. This is done by adjusting weights and bias values accordingly. The perceptron learning rule for our problem is applied repeatedly to adjust weights and bias values due to the error. The process is repeated for more iteration until the errors are removed and the desired result is obtained. The results are satisfactory and indicate that the net has been trained to assess the content for any similar or dissimilar words.

The ANN model is tested for different datasets in comparison with base word 'magnetism'. The base dataset consisting of 6words when trained with 'magnetism' gave performance index value =0.0139. Another different dataset consisting of 7words with 'pole' as 7th word gave performance index value = 0.00315, whereas other dataset consisting of 8words with 'science' as 8th word gave performance index value = 0.0121. The results show clearly that 'science' is closer to 'magnetism' than compared to 'pole' as the performance index value says. This methodology can be applied to any datasets and documents to judge whether it is related to content or not.

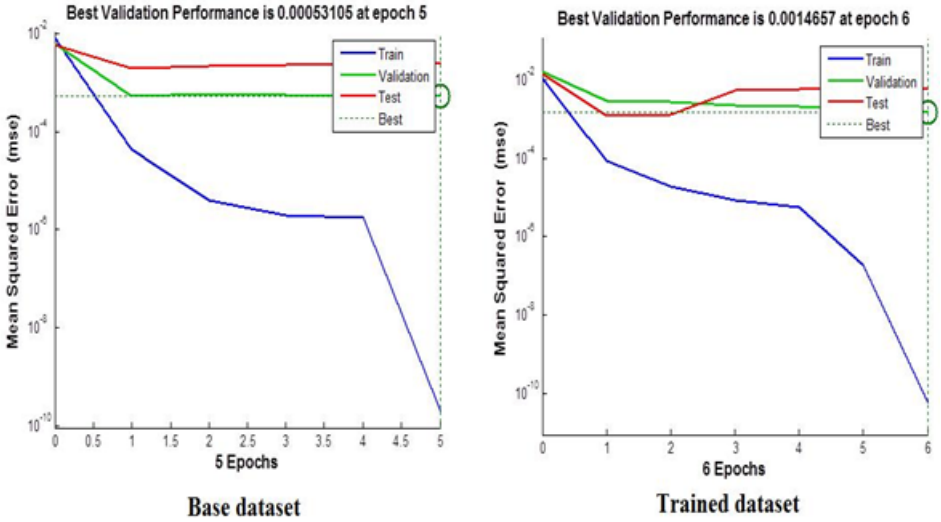


Figure 11. Comparison of best performance index value.

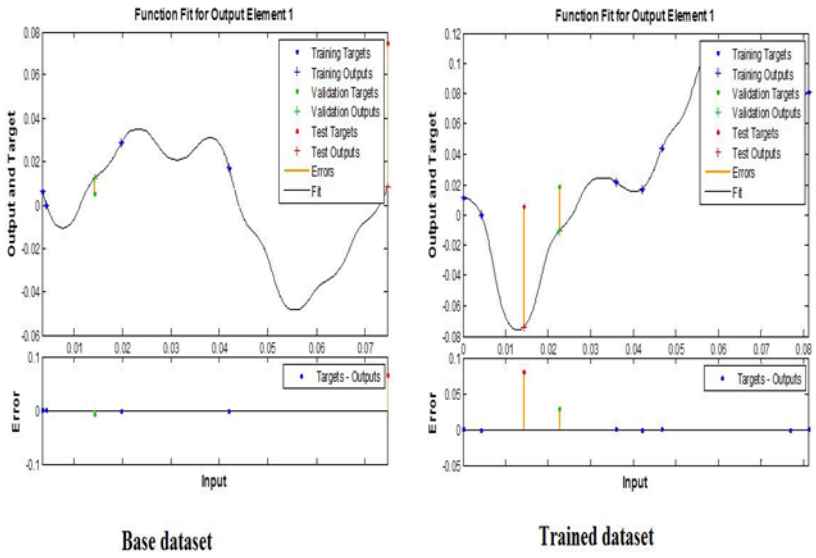


Figure 12. Comparison of Function plots.

7. Conclusions

After discussing the complexities in different kind of present day web documents, need for a more basic approach to handle the unstructured and multi-lingual features is emphasised. Later a pixel-based generic model for Content Extraction for regional web documents is developed. Beginning with complexities in letters, different methods of generating attributes are presented which form the basis for pattern matching and later for neural modelling. Some preliminary test results are given for pattern matching of features, for letter and word level relating to the same content. Later neural model with different attribute inputs is developed and using the 3 x 3 matrix pattern of input, training for set of words dealing with magnetism is done. This is later tested with other language words to form a more elaborate base set. It was tested with new words to assess the closeness of the content. Further enhancements and techniques are under development to account for the vagaries of font and text dissimilarities and hybrid usage, so that, ant regional multi-lingual web content is extractable.

8. References

1. Gottron T. Content code blurring: A new approach to content extraction. DEXA '08: 19th International Workshop on Database and Expert Systems Applications. IEEE Computer Society. 2008; p. 29–33.
2. Gupta S, Kaiser G, Neistadt D, Grimm G. DOM based content extraction of HTML documents. New York, NY, USA: ACM Press: WWW '03: Proceedings of the 12th International Conference on World Wide Web. p. 207–14.
3. Moreno J, Deschacht K, Moens M. Language independent content extraction from web pages. Proceeding of the 9th Dutch-Belgian Information Retrieval Workshop. 2009; p. 50–55.
4. Kolla Bhanu Prakash. Mining Issues in Traditional Indian Web Documents. Indian Journal of Science and Technology. 2015 November; 8(32), ISSN: 0974-5645. Doi: 10.17485/ijst/ 2015/v8i1/77056.
5. Mantratzis C, Orgun M, Cassidy S. Separating XHTML content from navigation clutter using DOM-structure block analysis. New York, NY, USA: ACM Press: HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia. 2005; p. 145–47.
6. Debnath S, Mitra P, Lee Giles C. Identifying content blocks from web documents. Foundations of Intelligent Systems, ser. Lecture Notes in Computer Science. 2005; p. 285–93.
7. Kolla Bhanu Prakash, Dorai Rangaswamy MA, Ananthan TV, Rajavarman VN. Information Extraction in Unstructured Multilingual Web Documents. Indian Journal of Science and Technology. 2015 July; 8(16). Doi: 10.17485/ijst/2015/v8i16/54252.
8. Hawkey Kirstie, Inkpen Kori. Web browsing today: the impact of changing contexts on user activity. New York, NY, USA, ACM Press: CHI '05: CHI '05 extended abstracts on Human factors in computing systems, ages. 2005; p. 1443-46.
9. William Jones, Harry Bruce, Susan Dumais. Once found, what then? A study of keeping behaviours in the personal use of web information, Proc. of ASIST.
10. Abigail J Sellen, Murphy Rachel, Kate L Shaw. How knowledge workers use the web. New York, NY, USA, ACM: CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems. 2002; p. 227-34.
11. Rahman AFR, Alam H, Hartono R. Content extraction from html documents. In WDA. 2001; p. 7–10.
12. Kolla Bhanu Prakash, Dorai Rangaswamy MA, Raja Raman Arun. ANN for Multi-lingual Regional Web Communication, ICONIP 2012, Part V, LNCS 7667. 2012; p. 473-78.
13. Kolla Bhanu Prakash, Dorai Rangaswamy MA, Raja Raman Arun. Statistical Interpretation for Mining Hybrid Regional Web Documents, ICIP 2012, CCIS 292. 2012; p. 503–12.
14. Kolla Bhanu Prakash, Dorai Rangaswamy MA, Raja Raman Arun. Performance of Content Based Mining Approach for Multi-lingual Textual Data. International Journal of Modern Engineering Research. 2011; 1(1) p. 146-50.
15. Deng Cai, Shipeng Yu, Ji-Rong Wen,Wei-Ying Ma. VIPS: a Vision-based Page Segmentation Algorithm, Technical Report, MSR-TR-2003-79, Microsoft Research, Microsoft Corporation, One Microsoft Way, Redmond, WA 98052. 2013.