1. Introduction

The last decade witnessed a complete shift in communications with emphasis on cellular and web technology dominating. Availability of cost-effective cell-phones to suit different segments of society has created a new shift in many human activities at societal and organizational level with SMS, Internet and web becoming common communication language. With the emergence of Internet and easier access to it anytime globally, large amount of data availability, interaction on the web and dependence on data for day-to-day issues are becoming necessary. It is said that Internet and web are probably the largest source of information availability and access at any time anywhere and more than 1.5 billion people use the Internet on a daily basis as described in several news items. People have different reasons and purposes for using Internet. Some use search engines, such as Google, Yahoo, etc., to get to know and access information for a topic of interest. Some advanced users will be extracting information from pages for later processing. This can be achieved by the so-called information extraction systems. Pre-processing, data mining and post-processing are normally terms that are associated with information extraction. All these aspects relate predominantly to English-based web development and even here terms like web noise, static or dynamic or temporal nature of data and methods of web generation have come into focus in web usage. Web noise is defined as information present on web pages and not relevant to the main content of the page. For example, in an article with a specific title, all the copyright, banners, advertisements and other redundant segments within that website are considered noise; because, they do not add any value to the article. But in Indian context, web and Internet communication need to cater to different languages, dialects and traditions as diversions and variations in customs, languages and dialects are quite significant geographically. Since communication can be in any language, web also should be able to cater to this and the present study is on issues related to extraction of information in typical Indian web pages.
2. Brief Review on Related Literature

Extracting information and content from conventional web pages, done in English, have received considerable attention in the research community, and selected references for survey like Gupta et al.\textsuperscript{12} and books like Han et al.\textsuperscript{13} are also available. Works related to structured and unstructured web pages deal with using software\textsuperscript{4,14}; block importance\textsuperscript{15,16} and cleaning\textsuperscript{9}. Recently, Akbar et al.\textsuperscript{19} proposed an algorithm to extract the main blocks in blog posts. Another study using predictive approach by Weiss et al.\textsuperscript{20} and a different approach based on pattern matching by Wu et al.\textsuperscript{21} significantly deals with text documents predominantly in English. Most of the work relates to extracting content from structured and well-developed web pages. But web pages in Indian context and in particular using traditional data are quite complex.

2.1 Web Documents in Indian Context

Indian web documents are quite varied and if one looks at the web pages developed in different regions on the same day, show considerable variation in content. Figure 1 gives two typical web pages taken on same day-25\textsuperscript{th} December, 2014, in two languages, Hindi and English. One can see very few common items and even images and text for the same content are different. Extracting content from these web pages calls for an approach which needs to be generic and at the same time different from earlier data mining and text mining techniques for handling language, script and dialect related issues.

So, current Indian web pages are multi-lingual, unstructured and non-homogeneous; but, if one adds tradition to this through digitized data, complexities increase like the two documents shown in Figure 2. Here an image of a ‘vimana’ used in Ramayana is shown in a sketch form and definitions in the form of sutras-verses, whose translations are also shown therein. A varied form

![Figure 1](image1.png)

Figure 1. Web pages on the same day-Indian context. (a) Hindi news web page. (b) English news web page.
is the text written on stone or metal and palm leaf as shown in Figure 3.

With this background, Indian web pages -current or traditional- have specific features which can be grouped into unstructured, non-homogeneous, and multi-lingual and strongly dialect related. A higher level of information extraction more general than conventional data mining and which deals with web pages as they are is needed and towards this, media mining approaches have come into prominence.

2.2 Media Mining Approach for Indian Texts

Media mining is progressing very fast in the last decade and has become a powerful tool with number of software backups for translating documents in different languages and in different forms. These approaches which rely on the structural features of the documents may not be directly applicable to multi-lingual documents in the Indian context, mainly due to varied forms of regional language texts and dialects besides the syntax formats in some cases being different. A letter in Indian regional language is quite different from English as the text may be a combination of two or three different characters like ‘सौ’ in Hindi meaning hundred. This is one aspect of Indian regional language text making it difficult to apply directly data mining approaches. Further a letter in one language can be a part of a word or a word by itself like the letter ‘a’ or ‘I’ in English and ‘म’ in Hindi.

Figure 4 gives text complexities at letter level in Indian context. Figure 4(a) gives a letter in English ‘a’ with its equivalent words in translated form in four different Indian regional languages. This shows the special feature of a single character required in English to form a word; whereas it requires more than one character to form a word in other languages. Figure 4(b) gives three examples of ‘a’ used as a character in other words to give a different
meaning. Figure 5(a) gives a word in Hindi ‘मा’ which is a single character in Hindi, but it requires more than one character to form its equivalent in other languages like ‘mother’ in English and ‘உள்ளே’ in Tamil. Figure 5(b) gives another variation of Hindi ‘माँ’ used as part of other words to give a meaningful word but of different content.

So, content extraction in multi-lingual web pages, calls for a quick and efficient overall assessment of the content of a web page irrespective of the language used for its development and it is proposed to cover this aspect by giving examples at letter and character level and later at word level. Figure 6 gives a typical bilingual web page with multi-tasking features like text in two different languages, images embedded with text, advertisements, and add-ons all on the same webpage.

Indian languages are very much different from English. English being the link language both in communication and forms the basis in higher education, some complexities in migrating from English to regional language or vice-versa exist like the one shown in Figure 7. Here a word ‘magnet’ in English, translated into Hindi, Telugu, Tamil as shown in Figure 7, shows clearly the variations in script and structure of text in different forms. This is true in education where text books written by authors in regional languages are used in web pages. Typically in Figure 8 a page of computer science text book is shown in two languages English and Tamil. Here, one can see computer terms in English are used as they are - in Tamil, scripts with letters in English and image does not change. Translating everything into Tamil may not work and is not possible and even if done may not reach the student easily in terms of content.
A general method is proposed which operates at pixel level so that text and language variations can be taken care of. This is becoming very important in education where mixing letters and words mutually with and without translation is becoming quite common while teaching. This is detailed in the following sections.

3. Methodology Proposed and Typical Results

Here an overall organization of the proposed model is presented as flowchart in Figure 9; where input preparation followed by attribute generation; algorithm usage and content extraction are the major segments.

3.1 Input Preparation

For any data processing or numerical modeling, data and its form play an important role. Data in the Internet and web context could be of different kinds; like text or script or image or even video and as such for computer modeling and processing these needs to be presented in the basic computer-recognizable form which is a pixel map. Pixel maps are matrix representations of data—whether text or image—giving x and y locations and the intensity or color.

Figure 8. Computer Science class-12 Textbook page in English and Tamil.

Figure 9. Overall organization of the proposed method.
code of each pixel. Typically the letter ‘a’ in text form can be represented as \([nR \times nC \times nI]\) matrix where, ‘nR’ is the number of rows, ‘nC’ number of columns and ‘nI’ is the intensity in grayscale form of 0-7 or 0-255. This matrix also depends on the file format like whether it is .bmp or compressed .jpg form. Using all the elements of matrix – total elements = \(nR \times nC \times nI\) – could be quite complex and so representing through features or attributes, which are less in size for computer modeling and processing is done through data-preparation and data cleaning. In data preparation, the matrix of pixel map is converted to a scalar or vector or matrix values. In data cleaning, a binary representation is chosen so that only 0 and 1 values are present. The matrix after cleaning is used to get its major features so that a unique representation in numerical modeling is possible. These attributes form the input phase of the algorithm.

### 3.2 Attribute Generation

Feature extraction of pixel maps in different ways could be:

- One number giving the sum of all non-zero voxels-volume pixels.
- Equivalent mean and deviation of the distribution of non-zero values.
- Converting the image into pre-assigned three row vector of non-zero values.
- Converting the actual matrix into equivalent 2 x 2 matrix of non-zero values, and
- Converting to 3 x 3 reduced matrix nine non-zero values.

Voxel stands for volume pixel which opens another dimension in pixel map manipulation, since presence of pixel gives zero and non-zero values as attributes, the other third dimension is given by voxel.

Figure 10 gives the histograms for attribute variations of ‘a’ treated as word in comparison with Hindi and Tamil equivalents for that word ‘a’. In comparing these attribute variations; the first chart gives values less than 1 as they are normalized, while others are actual values. It is clear that for voxel values ‘a’ in Hindi comes more closer to ‘a’ in Tamil when compared to ‘a’ in English. In matching the pattern with the values given, the content is easily predicted. The next case is when ‘a’ is used as part of a word like ‘apple’ or ‘mango’ or ‘orange’ and Figure 11 gives similar bar charts for comparison.

In comparing Figures 10 and 11 using pattern matching, it shows clearly that both of them are of different content. This proves our assumption of letter ‘a’ which is having unique meaning when used as a single letter, and when used as part of other words like ‘apple’ has totally different meaning, bringing the context of totally different content.

Now the observations made for English need translations to other Indian languages like Hindi where letter ‘मा’, means mother and equivalents in other two languages are taken and Figure 12 gives a comparison.

In comparing all the attribute variations, it is found that for non-zero attribute values Hindi and Tamil have closure similarities when compared to ‘maa’ in English. For 3-value attribute, English and Tamil patterns match

![Figure 10. Feature variations of word ‘a’ in comparison with other languages.](image-url)
much closure when compared to Hindi ‘maa’. Similarly, for 2 x 2 value attributes, Hindi and Tamil match closely in pattern. This is shown in Figure 12 and Figure 13 respectively. Another distinct observation came out of this study is ‘माँ’ if compared carefully with ‘mother’ and ‘உலகம்’, a dot is seen on the top half in Hindi, where the other characters presence is not seen. This is proved by presence of zeroes in values observed in Figure 14. Here again the first chart gives values less than 1 as ratio with total pixels is taken whereas others are actual values.

But values given in these figures need generality and normalization is done with number of pixel maps so that all values will be less than 1. These normalized values for all four types of attributes are shown in Figure 15 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. With these five types of attributes, algorithm can be started to assess the content.

Figure 11. Feature variations of ‘a’ as part of other words.

Figure 12. Feature variations of word ‘माँ’ in comparison with other languages.
Figure 13. Feature variations of ‘maa’ as part of other words.

Figure 14. Normalized attributes comparison.
In comparing all the attribute variations, it is found that for non-zero attribute values Hindi and Tamil have closure similarities when compared to ‘a’ in English. For 3-value attribute, English and Tamil patterns match much closure when compared to Hindi ‘a’. For 3 x 3 value attributes, Hindi and Tamil match closely when compared to ‘a’ in English. At last when we compared all the 5 value attributes, it is observed that ‘a’ in Hindi is close to ‘a’ in Tamil. As the base pixel map is that of English for ‘a’ and for Hindi ‘a’, it better to normalize all with respect to base values. This is shown in Figure 15 for ‘a’ and Figure 16 for ‘मा’ and this normalization helps one to assess for any new letter, whether it is a word in both languages.

For 3 x 3 value attributes, Hindi and Tamil match closely when compared to ‘मा’ in English. At last when we compared all the 5 value attributes, it is observed that ‘मा’ in Hindi is close to ‘maa’ in Tamil. This is shown clearly in Figure 12 and Figure 13 respectively. Figure 16 gives a comparison of all the attributes normalized with ‘मा’ in Hindi. Since all the features are normalized, this comparison gives better observations, like if a line is drawn at the center, many values are in the same range for each attribute indicating content ‘मा’.

Our next interesting observation is letter ‘x’ which is unique to English and doesn’t have any meaning when used as single character. Even very rarely any proper meaningful word starts with ‘x’. For example “xeon-lamp” or “x-mas” is not unique English word. So, we don’t have an exact translation for ‘x’ in other Indian regional languages. We can write ‘x’ in other languages like Hindi and Tamil as a combination of two or more characters, as shown in Figure 17.

Figure 15. Comparison of all the attributes normalized with ‘a’ in English.

Figure 16. Comparison of all the attributes normalized with ‘maa’ in Hindi.

Figure 17. Character complexities for ‘x’ in English compared with Hindi and Tamil.
In examining closely in all the above cases it is observed that, a minimum of 2 x 2 attributes is required to get an idea of content. Good conclusions came from 2 x 2, 3 x 3 and 5 attributes. Such as 'a' when treated as single character as a word has a different meaning, compared to ‘a’ when used as part of another words. The same results were observed with a regional language word. So, to understand more better about these attributes, it is better to train a neural network with any of these three different attributes.

4. Algorithm for Content Assessment

Since it is possible to generate attributes for any letter or word or image using the approach mentioned earlier, algorithms like statistical or pattern matching or neural network model can be used to assess the content. Already some results for statistical methods were presented by the authors, here ANN model is presented. To train any neural network model, we need two inputs and a target value with which the attribute values were normalized. For which we considered here ‘a’ in English as target value and 'a' in Hindi and Tamil as Inputs. We trained the network using LM algorithm. Epoch stands for iteration. The performance of the network can be measured with different inputs keeping target constant. The best performance is achieved at some epoch value. If we are not satisfied with the performance of the network, we can retrain the network. Still if we are not satisfied we can also adjust the network, by altering the hidden neurons in the network.
Figure 19 gives the neural network plot for comparing 3 x 3 attribute variations. It is observed that for the first iteration the best validation performance is attained.

Figure 20 gives attribute variations for 3 x 3 compared with 2 x 2 using neural network fitting plot, the function fit for output element 1 shows the input and output targets along with errors observed. The results are satisfactory and indicate that the net has been trained to assess the content for any similar or dissimilar words. More work is under way for creating a depository for content assessment.

5. Conclusion

In this paper we presented a detailed study of attribute variations at character and word level. The complexities and structure variations in Indian regional languages are identified. Pattern matching techniques at attribute level yielded good results to identify the content. It is observed that, a minimum of 2 x 2 attributes is required to get an idea of content. Good conclusions came from 2 x 2, 3 x 3 and 5 attributes. Such as ‘a’ when treated as single character as a word has a different meaning, compared to ‘a’ when used as part of another words. The same results were observed with a regional language word. Neural network fitting helped to judge best valid performance of the attribute values. The work will be extended further with in-depth study of neural network processing at more technical word and sentence level.

6. References

1. Liu B, Chen-Chuan-Chang K. Editorial: special issue on Web content mining. ACM SIGKDD Explorations Newsletter. 2004. p. 1–4.
2. Available from: http://www.internetworldstats.com/stats.htm
3. Gibson D, Punera K, Tomkinds A. The volume and evolution of Web page template. Chiba, Japan: International World Wide Web Conference, ACM. 2005. p. 830–9.
4. Song R, Liu H, Wen JR, Ma W-Y. Learning block importance models for Web pages. New York, NY, USA: Proceedings of the 13th International Conference on World Wide Web, ACM. 2004. p. 203–11.
5. Chen J, Zhou B, Shi J, Zhang H, Fenqwq Q. Function-based object model towards website adaptation. Proceedings of the 10th International Conference on World Wide Web; Hong Kong. 2001.
6. Liu B. New York: Web data mining: Exploring hyperlinks contents and usage data. Springer. 2006.
7. Liu B, Hsu W, Ma Y. Integrating classification and association rule mining, ACM SIGKDD. AAAI Press; 1998. p. 80–6.
8. van Gils B, Proper HAE, van Bommel P, van der Weide ThP. On the quality of resources on the Web: An information retrieval perspective. Information Sciences. 2007; 177(21):4566–97.
9. Li J, Ezeife C. Cleaning Web pages for effective Web content mining. Krakow, Poland: 17th International Conference Database and Expert Systems Applications. 2006. p. 560–71.
10. Yi L, Liu B, Li X. Eliminating noisy information in Web pages for data mining. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM; Washington. 2003. p. 296–305.
11. Ahmadi-Abkenari F, Selamat A. An architecture for a focused trend parallel Web crawler with the application of clickstream analysis. Information Sciences. 2012; 184(1):266–81.
12. Gupta V, et al. A survey of text mining techniques and applications. Journal of Emerging Technologies in Web Intelligence. 2009 Aug; 1(1):60–76.
13. Han J, et al. Data mining: Concepts and techniques. MK Publishers. 2006.
14. Gupta S, Kaiser G, Neistadt D, Grimm P. Budapest, Hungary: DOM-based content extraction of HTML documents, ACM. 2003.
15. Lo L, To-Yee Ng V, Ng P, Chan SC. Automatic template detection for structured Web pages. Nanjing, China: Proceedings of the 10th International Conference on Computer Supported Cooperative Work in Design. 2006. p. 1–6.
16. Lin S-H, Ho J-M. Discovering informative content blocks from Web documents. New York, NY, USA: The 8th ACM, SIGKDD. International Conference on Knowledge Discovery and Data Mining. 2002. p. 588–93.
17. Cai D, Yu S, Wen J-R, Ma W-Y. VIPS: A vision-based page segmentation algorithm. Redmond: Microsoft Corporation; 2003.
18. Cai D, Yu S, Wen J-R, Ma W-Y. Sheffield, South Yorkshire, UK: Block-based Web search, SIGIR’04, ACM. 2004.
19. Akbar S, Slaughter L, Nytro. Extracting main content-blocks from blog posts. Proceedings of the International Conference on Knowledge Discovery and Information Retrieval. 2010. p. 438–43.
20. Weiss, et al. Text mining: Predictive methods for analyzing unstructured information. Springer Science Media. 2005.
21. Wu, et al. Deploying approaches for pattern refinement in text mining , data mining, ICDM ’06. 2006.
22. Bhanu Prakash K, Dorai Ranga Swamy MA, Raja Raman A. Statistical interpretation for mining hybrid regional Web documents. ICIP, CCIS. 2012. p. 503–12.
23. Bhanu Prakash K, Dorai Ranga Swamy MA, Raja Raman A. ANN for multi-lingual regional web documents. ICONIP, LNCS. 2012. p. 473–8.
24. Bhanu Prakash K, Dorai Ranga Swamy MA, Ananthan TV, Rajavarman VN. Information extraction in unstructured multilingual web documents. Indian Journal of Science and Technology. 2015 July; 8(16).
25. Bhanu Prakash K, Dorai Ranga Swamy MA, Ananthan TV. Feature extraction studies in a heterogeneous web world. Int J Appl Eng Res. 2014; 9(22):16571–9.