Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 89 (2016) 514 – 520

Procedia Computer Science

Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016)

Mining of Bilingual Indian Web Documents

Kolla Bhanu Prakash[a,*] and Arun Rajaraman[b]

[a] *Chirala Engineering College, Chirala, India*
[b] *IITM, Chennai, India*

Abstract

Web and mobile communication are growing in popularity globally and regionally catering to different ways of information dissemination, rendering complex web documents having script, language and media content embedded into them. Thus information extraction from different web documents in the modern day scenario is becoming a real challenge, as one has to cater to format and script variations in documented form and media variations in soft-web form. This has become very relevant in Indian education scenario, where bilingual and multi-lingual communication and web documents through on-line courses, are considered. When regional native dialect comes into picture, another dimension of complexity is added. The present paper focuses on content extraction of such documents through a generic approach using pixel-based approach and mining through classification. Indian bilingual web documents are considered and attribute generation is done through reducing the pixel matrix. Five different attributes were identified and studied. A clear state of art comparison between trained dataset and test dataset is given. The results give reasonable content extraction with good accuracy of the datasets studied.

Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016

Keywords: Attribute; Bilingual; Classification; Content Extraction; Mining; Pixel-based Approach; Voxel.

1. Introduction

Web and mobile communication are becoming the two main aspects of present day social and cultural information exchange and dissemination. While web and internet are major sources data and information generation, cellular communication through oral, SMS and other forms of media is opening a new dimension as language, dialect and regional flavor are the main forms used, leading to complex web/mobile data generation. This aspect in the Indian context is becoming a significant tool particularly in education, where on-line courses and distance education are gaining popularity. In this scenario, Indian web documents are quite complex and varied and pose a very interesting problem for mining and content extraction. Bilingual and in some cases multilingual communication plays a major role as present day teachers resort to using regional dialect with English words and this results in development of websites and web documents, where a DOM parser may not be helpful for data mining or content extraction. The concept of content extraction has its origin and key role in NLP, where its main use is on recognizing entities like person names and company information in news magazines and websites. Data on the web now-a-days has structured and unstructured form of documents, homogenous, heterogeneous and hybrid forms of media data and modern websites

*Corresponding author. Tel.: +91 -9840519623.
E-mail address: bhanu_prakash231@rediff.com

Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016
doi:10.1016/j.procs.2016.06.103

physics भौतिक विज्ञान طبيعيات ಭೌತಶಾಸ್ತ್ರ

Fig. 1. Structure Variations in Indian Regional Languages.

Fig. 2. An Example Bilingual Web Document in Indian Context.

present more challenges and complexities than conventional ones. At the first level, variation in text in different Indian languages is a starting point to present the complexity and Fig. 1, shows the word ‘physics’ given in four different languages in translated form.

If one looks at web pages it is even more involved and Fig. 2 shows the web page for an educational institution in Tamil Nadu, which has multilingual texts and different images integrated onto it. While English dominates there are regional dialects in Tamil language either in translated or transliterated form like ‘ANNAMALAI’, Tamil word written in English script. The present paper focuses on such web pages having bilingual web documents in Indian context.

It is observed that even among Indian languages, scripts have similarities like in Telugu and Kannada; but, a general Indian webpage may have lot of variation, as many scripts are derived from Arabic, Urdu, Hindi and other Indian regional languages. Arabic and Urdu are the languages where text is written from right to left. In all other Indian regional languages text is written left to right. In Chinese language, text is written top to bottom. In the presence of so many variations in text, complexities arise when only natural language processing tools are used for content extraction and hidden knowledge discovery. That is the reason; a generic approach is needed here to give better results. In media mining translation and transliteration do not play that much difference as is observed in NLP. Since, in media mining input is treated in terms of pixel-map variations.

2. Related Work

The main purpose and need of this research was influenced by Automatic Content Extraction (ACE)[1] and manually coded rules using rule generation[2]. Since, manual coding is becoming difficult, automatic rule learning algorithms are developed to solve the present problem[3]. Later, Hidden Markov models and conditional models based on entropy were deployed[4]. Statistical interpretation is also applied depending on the nature of application[5]. Some hybrid techniques were also applied for information extraction[6].

Another approach of relation extraction for Arabic languages[7] is studied recently. A good amount of literature is available on relation extraction in English and European languages but nothing as such is given for Indian regional languages. Pattern based methods[8], supervised methods[9] and Bootstrapping methods[10] are other existing popular methods in relation extraction. A robust approach of OCR for identifying script in Indian native languages has been studied recently for Historical Indian document images[11]. Earlier approaches generally focused on plain background images and used connected component analysis for recognition of scripts[12]. New Gradient-Spatial-Structural-Features for Video Script Identification is studied recently[13]. But, none of these methods were helpful to extract information from Indian regional web documents. So, a naïve algorithm is developed and content is extracted using attribute generation. Statistical Interpretation using media mining was presented in our earlier research. In the next sections, the methodology proposed and detailed case study of attribute generation is given.

3. Proposed Methodology

Since pixel-maps are the core of computer data representation and are completely independent of type of data, the pixel-map is used as the basis in the proposed method and since the pixel-maps are large in size and representation, reducing the 3D matrix through different means to get proper and efficient attribute representation.

There can be many ways data reduction is possible and here after converting the image to grey-scale –resulting in 2D matrix– different attribute types are generated and used in content extraction. A typical block diagram for content extraction of Indian web documents is given in Fig. 3. Proposed approach is in three segments, where the first segment is data preparation where the pixel-map is used as input matrix for the second stage where feature extraction is done to form the input for the numerical approach which may be statistical or neural or attribute generation and pattern matching.

4. Results and Discussion

A pixel-map is divided into three regions – top, middle and bottom halves. It is observed that Telugu, Kannada and Hindi occupy more part in top and bottom halves compared to English. In English major part lies in middle half and

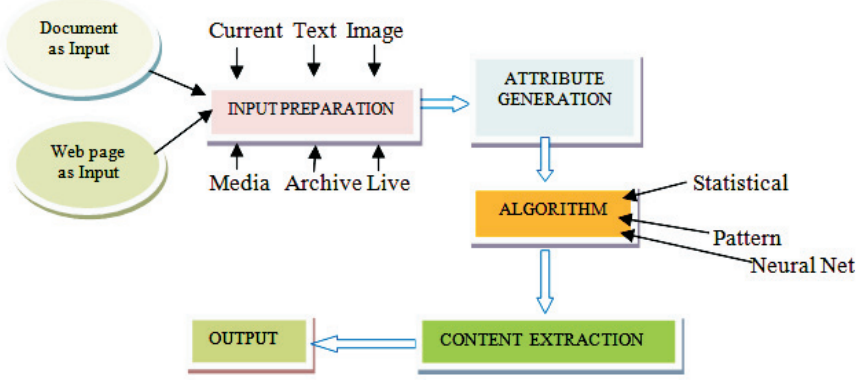

Fig. 3. Overall Organization of the Proposed Method *[Courtesy of Fig. 9 – Reference[14]].*

very less portion is seen on the top and bottom half. Considering all these complexities it is difficult to identify the content using xml DOM parsing, semantic tagging and many other conventional techniques. So, a generic approach of media mining using attribute generation, statistical interpretation and neural network techniques is better suited for content extraction of multi-lingual web documents. In order to give a detailed case study on attribute generation,[14] a preliminary approach of feature extraction of pixel maps in different ways was mentioned in our previous published work[14] as given below:

a) Unique feature giving the sum of all non-zero voxel's-volume pixel in a non-dimensional form
b) Equivalent mean and deviation of the distribution of non-zero values
c) Converting the image into pre-assigned three row vector of non-zero values
d) Converting the actual matrix into equivalent 2×2 matrix of non-zero values and
e) Converting to 3×3 reduced matrix nine non-zero values[14].

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[14]. Just to get an idea of the variations in input data for some words, histogram representations of typical ones are given. Figure 4 gives 3×3 attribute variations of base dataset consisting of 36 words related to education. The value on y-axis gives the corresponding attribute value and on x-axis all 30 words – 9 non-zero values are represented. This is considered as basic trained dataset and any new input or target data can be compared with this to match the pattern and identify the content.

Figure 5 gives three row vectors of non-zero values of the 30 words related to education considered. Although histogram representation for results is older technique, it is best suited to our generic approach of attribute generation and pattern matching.

Figure 6 and 7 gives Zero and Non-zero attribute variations of base dataset of the 30 words related to education considered and 2×2 attribute variations of trained dataset.

Figure 8 gives statistical interpretation of five attributes generated of 30 words which is considered as trained dataset. This is compared with Fig. 10 which is a set of 18 words considered as target dataset. Similarly, comparing Fig. 4 and 9 using pattern matching studies the measure of closeness is calculated and the content is identified. After careful observation, 90–90% accuracy is obtained using pattern matching in measuring the similarity of the trained and target datasets. Any new inputs further considered can be compared with the base dataset results and pattern matching studies gives content extraction to a marginal extent considerable.

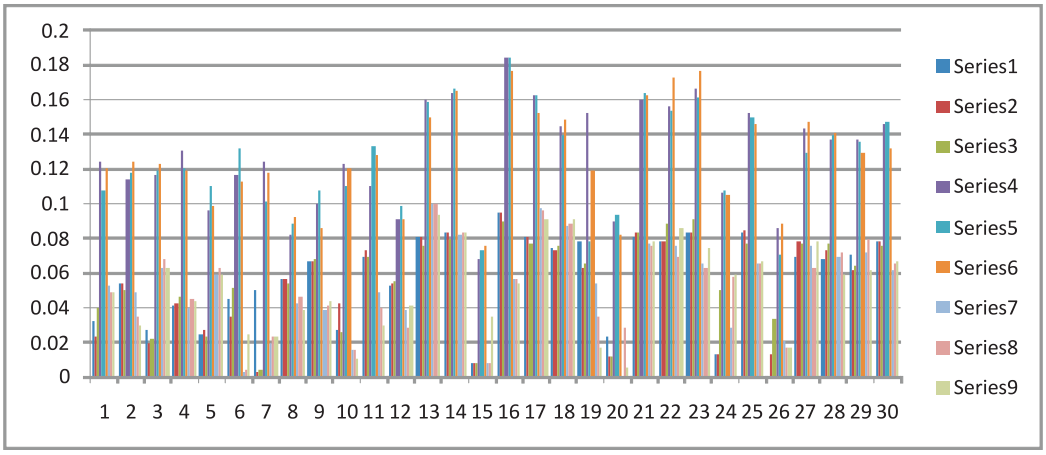

Fig. 4. 3×3 Attribute Variations of Base Dataset.

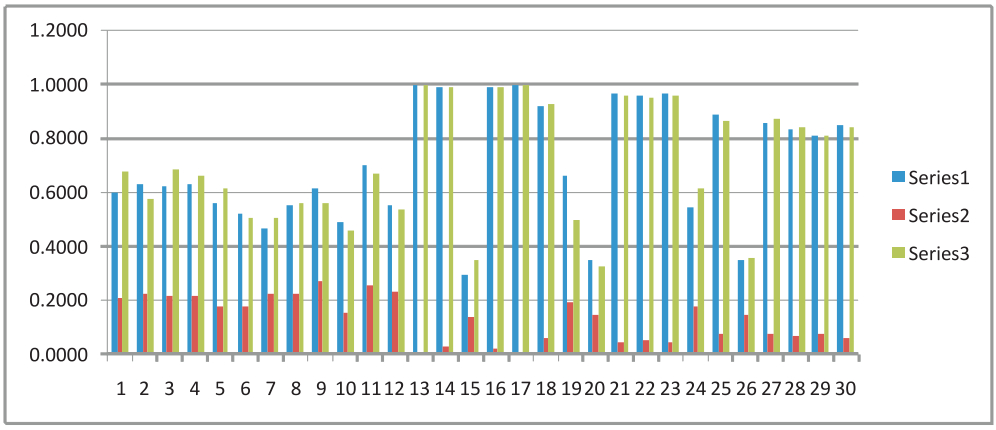

Fig. 5. Three Valued Attribute Variations of Base Dataset.

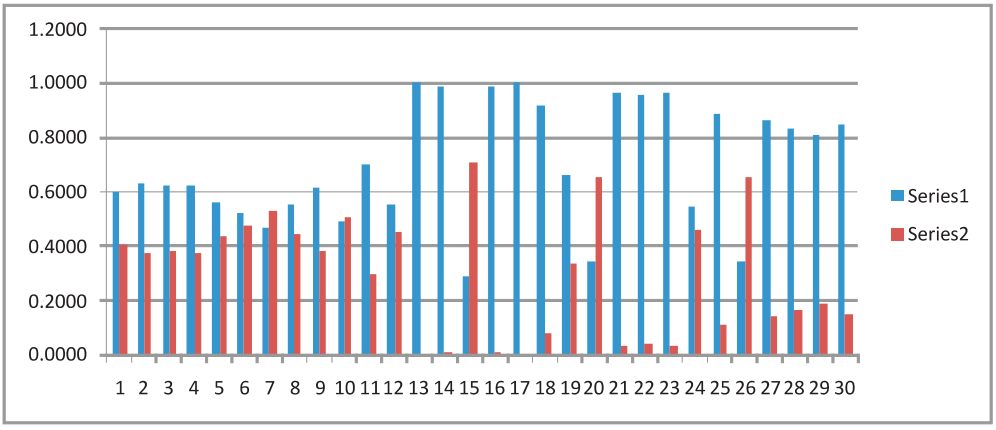

Fig. 6. Zero and Non-zero Attribute Variations of Base Dataset.

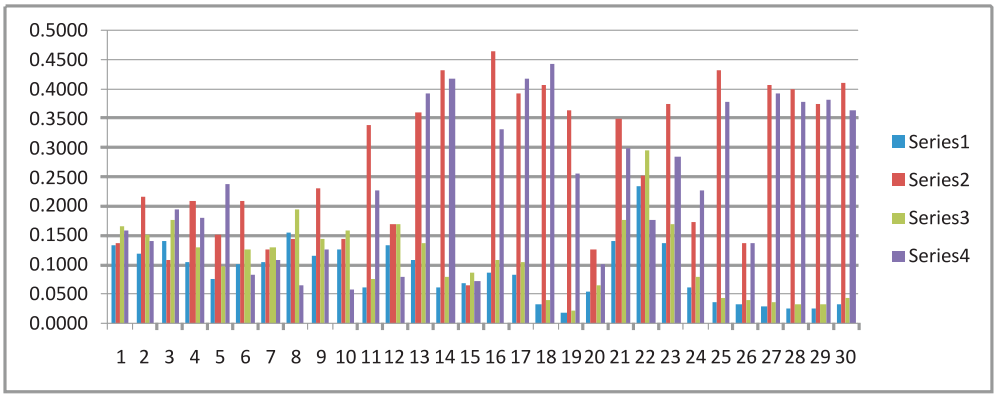

Fig. 7. 2×2 Attribute Variations of Base Dataset.

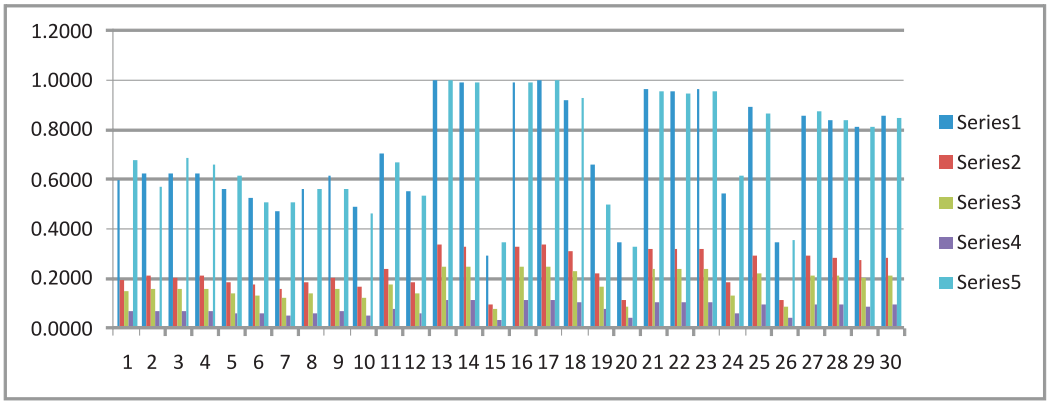

Fig. 8. Statistical Attribute Variations of Base Dataset.

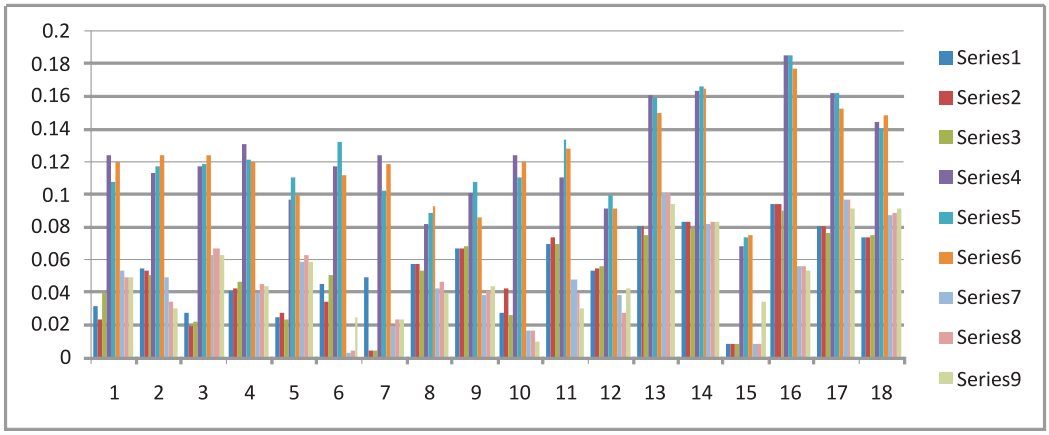

Fig. 9. 3×3 Attribute Variations of Test Dataset.

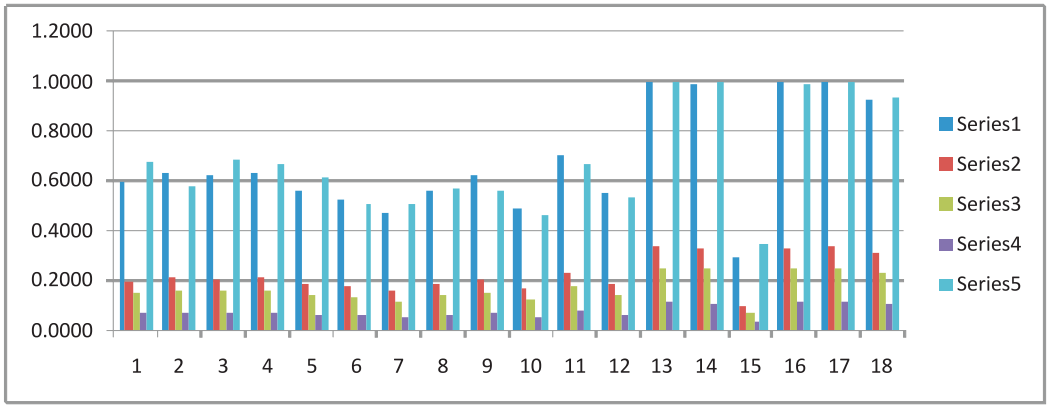

Fig. 10. Statistical Attribute Variations of Test Dataset.

5. Conclusions

In observing carefully modern online and offline web pages and files, it gave rise to an urgent requirement of generic and naïve strategy to handle documents like structured, semi-structured, unstructured, hybrid, heterogeneous and having multi-tasking and multi-lingual features. So, a method using pixel-map manipulation to extract content from Indian regional web documents is developed. This method is tested with other Indian and foreign native language words to form a more elaborate base set. To assess the similarity between trained and tested datasets, number of new datasets with new words was identified and tested using our present algorithm. More analysis on new strategies and algorithms is under progress. A detailed state-of-art analysis can be done with neural network[15] and cluster analysis. A comparison of statistical, neural, pattern matching algorithms will give better analysis of this generic approach.

Acknowledgement

The author sincerely thanks the Chirala Engineering College Management for their kind support in providing resources for doing this research work.

References

[1] ACE. Annotation Guidelines for Entity Detection and Tracking, (2004).
[2] J. Aitken, Learning Information Extraction Rules: An Inductive Logic Programming Approach, *Proceedings of the* 15^{th} *European Conference on Artificial Intelligence*, pp. 355–359, (2002).
[3] M. E. Califf and R. J. Mooney, Relational Learning of Pattern-match Rules for Information Extraction, *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, pp. 328–334, July (1999).
[4] D. Klein and C. D. Manning, Conditional Structure Versus Conditional Estimation in NLP Models, *Workshop on Empirical Methods in Natural Language Processing (EMNLP)*, (2002).
[5] H. Cunningham, D. Maynard, K. Bontcheva and V. Tablan, Gate: A Framework and Graphical Development Environment for Robust NLP Tools and Applications, *Proceedings of the* 40^{th} *Anniversary Meeting of the Association for Computational Linguistics*, (2002).
[6] G. Ramakrishnan, Using ILP to Construct Features for Information Extraction from Semi-structured Text, *ILP*, (2007).
[7] Maha Al-Yahya, Sawsan Al-Malak and Luluh Aldhubayi, Ontological Lexicon Enrichment: The Badea System for Semi-Automated Extraction Of Antonymy Relations from Arabic Language Corpora., *Malaysian Journal of Computer Science*, vol. 29(1), pp. 56–73, (2016).
[8] R. G. Raj and S. Abdul-Kareem, A Pattern Based Approach for the Derivation of Base Forms of Verbs from Participles and Tenses for Flexible NLP., *Malaysian Journal of Computer Science*, vol. 24(2), pp. 63–72, Jun. (2011).
[9] S.-P. Choi, S. Lee, H. Jung and S.-K. Song, An Intensive Case Study on Kernel-based Relation Extraction, *Multimedia Tools Appl.*, vol. 71, no. 2, pp. 741–767, Jul. (2014).
[10] P. Pantel and M. Pennacchiotti, Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, *Proceedings of the* 21^{st} *International Conference on Computational Linguistics and the* 44^{th} *Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, pp. 113–120, (2006).
[11] S. Kavitha, P. Shivakumara, G. Hemantha Kumar and C. L. Tan, A Robust Script Identification System For Historical Indian Document Images, *Malaysian Journal of Computer Science*, vol. 28(4), pp. 283–300, (2015).
[12] P. Krishnan, N. Sankaran, A. K. Singh and C. V. Jawahar, Towards a Robust OCR System for Indic Scripts, *Document Analysis Systems, IEEE*, pp. 141–145, April (2014).
[13] P. Shivakumara, Z. Yuan, D. Zhao, T. Lu and C. L. Tan, New Gradient-Spatial-Structural-Features for Video Script Identification, *Computer Vision and Image Understanding*, Elsevier, vol. 130, pp. 35–53, January (2015).
[14] K. Bhanu Prakash, Mining Issues in Traditional Indian Web Documents, *Indian Journal of Science and Technology*, vol. 8(32), (2015).
[15] K, Bhanu Prakash, M. A. Dorai Ranga Swamy and A. Raja Raman, ANN for Multi-lingual Regional Web Documents, *ICONIP, LNCS*, pp. 473–8, (2012).