International Journal of Engineering & Technology, 7 (2.8) (2018) 68-71

International Journal of Engineering & Technology

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

Research Paper

Information extraction in current Indian web documents

Kolla Bhanu Prakash *

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur, India
**Corresponding author E-mail: drkbp@kluniversity.in*

Abstract

Communication and Internet are two major resources in today's technical, social and scientific disciplines offering a wide range of possibilities in bringing in new approaches and variations in current ones. Web documents are increasingly growing in size, volume and time, bringing in the need to access and process them off and online over the Internet with a PC or a smart phone. When viewed in Indian context, web documents pose different kinds of challenge and the present study addresses some of them taking into account the vagaries in the Indian languages. This has become very relevant in Indian education scenario, where bilingual and multi-lingual communication and web documents through on-line courses, are being generated. When regional native dialect comes into picture, another dimension of complexity is added. After presenting the different kinds of web pages in the Indian perspective, the case for the development of a generic approach id highlighted so that it can blend with current tools of data mining and at the same time cater to vagaries in Indian texts. The approach based on a pixel level addressing of data-which is of large size-, is later modified and reduced to numerical equivalents using matrix manipulations so that they form inputs to some classification approaches, like statistical, pattern matching and neural models. Some typical case studies on text letters and words are presented to highlight the generality of approach and its flexibility to fit into different tools.

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

1. Introduction

Current social, technical and scientific approaches to different problems and issues revolve around communication and Internet and web-based and mobile communication are becoming the two main sources 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 feature, in the Indian context is becoming a significant tool particularly in education, business and social order, where on-line and off-line data 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 communication resorts to using regional dialect with English words and this results in the development of websites and web documents, which for people in different regions do not provide clarity on content. Conventional DOM parser may not be helpful for data mining or content extraction. Data on the web now-a-days has structured and unstructured form of documents, homogeneous, heterogeneous and hybrid forms of media data and modern websites present more challenges and complexities than conventional ones. It is preferable to look at different vagaries in the Indian web pages.

2. Vagaries in Indian web pages

Considering web pages in the Indian scenario, beginning with news and education, some typical web pages are shown in Fig.1.News being displayed in India differs in different regions and the content may not be consistent and same as shown in Fig.1(a) and Fig.1 (b) gives an idea of variations in content even-for the same discipline 'education'. It is clear that Indian web pages need a different approach to enable the user to get at the content. At the first level, variation in text in different Indian languages is a starting point to present the complexity and Fig. 2 shows the word 'physics' given in four different languages in translated form at top and another word 'Internet' written in different languages as 'Internet', since, many times it is convenient to use English words as they are either in communication or while teaching. These aspects lead to development of web sites which are unstructured, non-homogeneous and multi-lingual as shown in Fig.3 for a University in South India.

A) News in Dailies on the Same Day

B) Vagaries in Web Page for 'Education'

Fig. 1: Types of Indian Web Pages.

physics भौतिक विज्ञान طبيعيات భౌతశాస్త్ర

Internet इंटरनेट انٹرنیٹ ಇಂಟರ್ನೆಟ್

Fig. 2: Text and Script Variations in Indian Regional Languages.

Fig.3 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. So it is necessary to look at different levels in Indian web pages beginning with letter to text to words and documents.

Fig. 3: An Example Bilingual Web Document 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. That is the reason; a generic approach is needed here to give a basic approach to develop inputs for data mining, so that translation and transliteration do not play that much difference. The main purpose and need of this study 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] and another approach of relation extraction for Arabic languages [7] is also available. 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. But, most of these methods do not address the complexities, mentioned earlier to extract information from Indian regional web documents. So, a generic and fundamental pixel level approach is developed [11], [12] and content is extracted using conventional methods, can be easily merged into them.

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. But pixel-maps are large in size and representation, reducing the 3D matrix through different means to get proper and efficient attribute representation is the first phase of the approach.

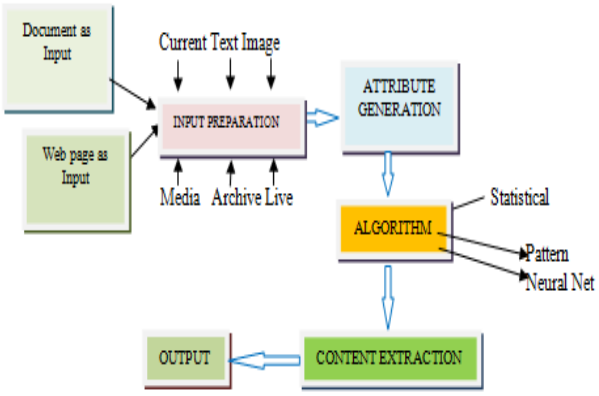

Fig. 4: Overall Organization of the Proposed Method.

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. 5. 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. Attribute generation

The first phase of the proposed method is to develop attributes based on pixel maps, as they are large in size. Since Indian languages have segments above and below the main character, the pixel-map is divided into three regions – top, middle and bottom segments.

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 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.

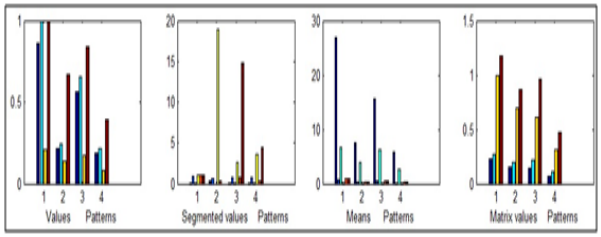

Fig. 5: Attributes for Letter-Words for Equivalent to 'I'.

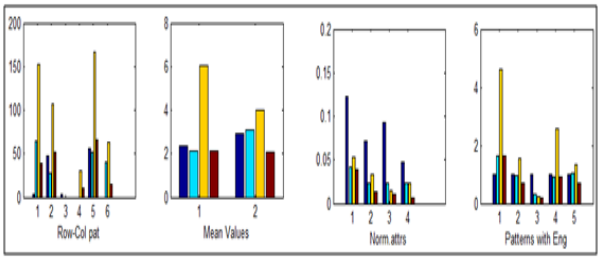

Fig. 6: Attribute Patterns for Four Different Sets.

In order to give a detailed case study on attribute generation, [11] the differences in choice of attributes for letters, scripts and equivalent words like 'I', a letter needs a word '*நான்*' in Tamil when translated and similarly a script like '*कौ*' in Hindi needs a word 'kau' in English. Taking 'I' the pixel map when converted to jpg file is a voxel, 3D matrix with rows, columns and intensity value of pixels. Feature extraction studies for pixel maps was mentioned earlier[11] and here a comparison of attributes for letter-words and different sets are shown as typical ones to show complexity. Fig. 5 shows for letter 'I' in English with word translation in three Indian languages, Hindi, Tamil and Telugu. One can see similarities in trend to enable classification.

Now choosing four different sets and a comparison of different attributes like scalars, matrices and means, Fig.6 shows a comparison to bring out pattern similarity. Just to get an idea of the variations in input data for some words, histogram representations of typical ones are given. Fig. 7 gives attribute variations for 30 pixel maps w.r.t 3x3 matrix which can be compared to any new data set for pattern matching studies.

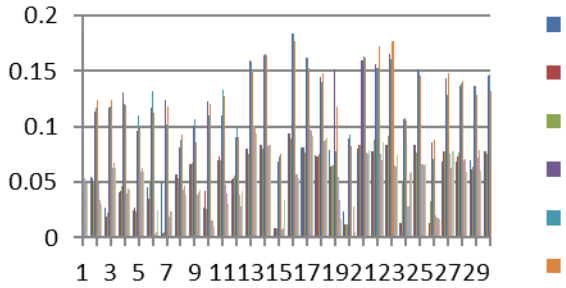

Fig. 7: 3x3 Attribute Variations of Base Dataset.

Fig. 8 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.

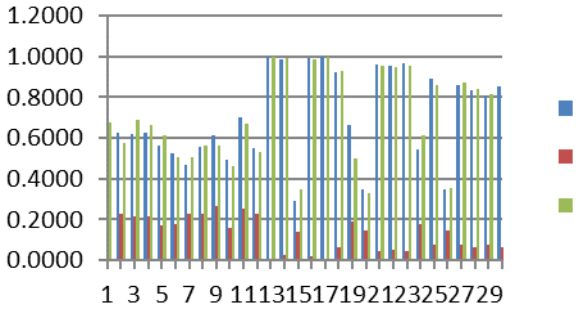

Fig. 8: Three Valued Attribute Variations of Base Dataset.

5. Content extraction

Now that different kinds of attributes are available for different datasets, content extraction through statistical matching and neural network are some of the methods used to demonstrate the prediction. Fig. 9 gives statistical interpretation of five attributes generated of 30 words which is considered as trained dataset. This is compared with fig. 11 which is a set of 18words considered as target dataset. Similarly, comparing fig. 4 and fig. 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.

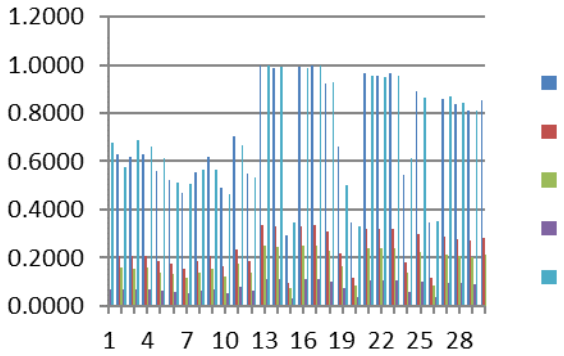

Fig. 9: Statistical Attribute Variations of Base Dataset.

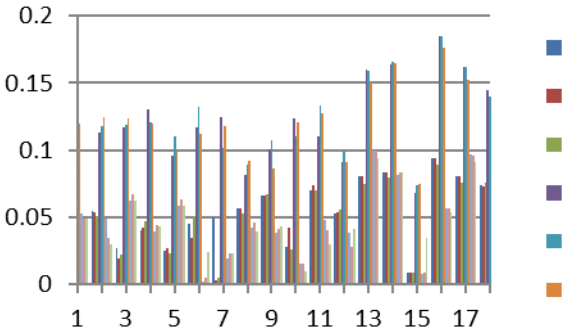

Fig. 10: 3x3 Attribute Variations of Test Dataset.

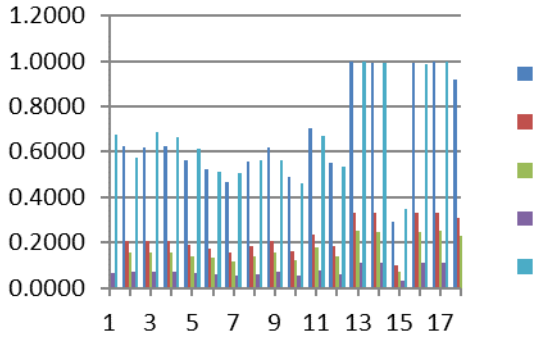

Fig. 11: Statistical Attribute Variations of Test Dataset.

Similarly, a neural network approach was also done and one typical result is shown in Fig.12.

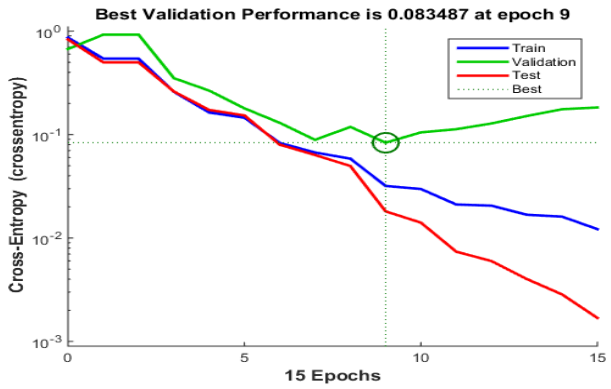

Fig. 12: Neural Method for Classification.

6. Conclusions

Present day online and offline web pages and files, in the Indian context need content extraction and this necessitates development of generic and fundamental strategy to handle documents like structured, semi-structured, unstructured, hybrid, heterogeneous and having multi-tasking and multi-lingual features. 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 comparison of statistical, neural, pattern matching algorithms will also give a comparative classification for better prediction with this generic approach.

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 15th 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. https://doi.org/10.3115/1118693.1118695.

[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 40th 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, LuluhAldhubayi, 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), 2016, pp 56-73. https://doi.org/10.22452/mjcs.vol29no1.5.

[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):Jun. 2011. pp 63-72.

[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. https://doi.org/10.1007/s11042-013-1380-5.

[10] P. Pantel and M. Pennacchiotti, Espresso: leveraging generic patterns for automatically harvesting semantic relations. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA, 2006, pp. 113–120. https://doi.org/10.3115/1220175.1220190.

[11] Bhanu Prakash, K, Mining Issues in Traditional Indian Web Documents. Indian Journal of Science and Technology, 8(32), 2015. https://doi.org/10.17485/ijst/2015/v8i1/77056.

[12] Bhanu Prakash K, DoraiRangaSwamy MA, Raja Raman A. Feature Extraction studies in a heterogenous web world. International Journal of Applied Engineering Research, 9(22), pp. 16571-16579, 2014.

[13] Dr. Seetaiah Kilaru, Hari Kishore K, Sravani T, Anvesh Chowdary L, Balaji T "Review and Analysis of Promising Technologies with Respect to fifth Generation Networks", 2014 First International Conference on Networks & Soft Computing, ISSN:978-1-4799-3486-7/14,pp.270-273,August2014.

[14] Meka Bharadwaj, Hari Kishore "Enhanced Launch-Off-Capture Testing Using BIST Designs" Journal of Engineering and Applied Sciences, ISSN No: 1816-949X, Vol No.12, Issue No.3, page: 636-643, April 2017.

[15] P Bala Gopal, K Hari Kishore, B.Praveen Kittu "An FPGA Implementation of On Chip UART Testing with BIST Techniques", International Journal of Applied Engineering Research, ISSN 0973-4562, Volume 10, Number 14 , pp. 34047-34051, August 2015

[16] A Murali, K Hari Kishore, D Venkat Reddy "Integrating FPGAs with Trigger Circuitry Core System Insertions for Observability in Debugging Process" Journal of Engineering and Applied Sciences, ISSN No: 1816-949X, Vol No.11, Issue No.12, page: 2643-2650, December 2016.

[17] Mahesh Mudavath, K Hari Kishore, D Venkat Reddy "Design of CMOS RF Front-End of Low Noise Amplifier for LTE System Applications Integrating FPGAs" Asian Journal of Information Technology, ISSN No: 1682-3915, Vol No.15, Issue No.20, page: 4040-4047, December 2016.

[18] N Bala Dastagiri, K Hari Kishore "Novel Design of Low Power Latch Comparator in 45nm for Cardiac Signal Monitoring", International Journal of Control Theory and Applications, ISSN No: 0974-5572, Vol No.9, Issue No.49, page: 117-123, May 2016.

[19] N Bala Gopal, Kakarla Hari Kishore "Reduction of Kickback Noise in Latched Comparators for Cardiac IMDs" Indian Journal of Science and Technology, ISSN No: 0974-6846, Vol No.9, Issue No.43, Page: 1-6, November 2016.

[20] S Nazeer Hussain, K Hari Kishore "Computational Optimization of Placement and Routing using Genetic Algorithm" Indian Journal of Science and Technology, ISSN No: 0974-6846, Vol No.9, Issue No.47, page: 1-4, December 2016.

[21] N.Prathima, K.Hari Kishore, "Design of a Low Power and High Performance Digital Multiplier Using a Novel 8T Adder", International Journal of Engineering Research and Applications, ISSN: 2248-9622, Vol. 3, Issue.1, Jan-Feb., 2013.

[22] Harikishore Kakarla, Madhavi Latha M and Habibulla Khan, "Transition Optimization in Fault Free Memory Application Using Bus-Align Mode", European Journal of Scientific Research, Vol.112, No.2, pp.237-245, ISSN: 1450-216x135/1450-202x, October 2013.

[23] T.RajeshKumar & G.R.Suresh, "Examination of Militants utilizing NAM Microphone and Wireless Handset for Murmured Speech in view of Concealed Markov Model". International Innovative Research Journal of Engineering and Technology. 112-119.

[24] T. Padmapriya and V. Saminadan, "Inter-cell Load Balancing technique for multi-class traffic in MIMO-LTE-A Networks", International Journal of Electrical, Electronics and Data Communication (IJEEDC), ISSN: 2320- 2084, vol.3, no.8, pp. 22-26, Aug 2015.

[25] S.V.Manikanthan and V.Rama, "Optimal Performance of Key Predistribution Protocol In Wireless Sensor Networks" International Innovative Research Journal of Engineering and Technology, ISSN NO: 2456-1983, Vol-2, Issue –Special –March 2017.