Indian Journal of Science and Technology, *Vol 8(16), DOI: 10.17485/ijst/2015/v8i16/54252, July 2015*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Information Extraction in Unstructured Multilingual Web Documents

Kolla Bhanu Prakash[1,3*], M. A. Dorai Rangaswamy[2], T. V. Ananthan[3] and V. N. Rajavarman[3]

[1]Faculty of Computer Science Engineering, Sathyabama University, Chennai - 600 119, Tamil Nadu, India; bhanu_prakash231@rediff.com
[2]C.S.E & IT, AVIT, Chennai - 603104, Tamil Nadu, India; drdorairs@yahoo.co.in
[3]Faculty of C.S.E, Dr. M.G.R. Educational & Research Institute, Chennai - 600 095, Tamil Nadu, India; tvananthan@rediffmail.com, nrajavarman2003@gmail.com

Abstract

Objectives: The objective is to develop a generic pixel-map based method to extract content in a short period of time for web documents. **Method of Analysis:** The method for extraction of content is in three levels, first level is in developing data inputs as attributes, second level in using the attributes to formulate a model and third level in interpretation of results. All three have variations so that validation comparison is possible for different parameters. Input data had all variations like language, script and usage and modeling is done using statistical, pattern recognition and ANN approaches. **Findings**: The method has demonstrated how quality and size of input data in the form of scalars, vectors and matrices affects the model and the result and this has been done for unstructured word sets chosen from web pages. The models chosen also give an idea of input/output variations in the outcome of the results. The uniqueness of the method is demonstrated for mono lingual, multi-lingual and transliterated datasets so that the applicability is universal. **Novelty/Improvement:** The method is generic in using pixel-maps, analytically stable in that the matrix input is used and versatility is demonstrated for adoption to different models.

Keywords: Data Mining Extraction, Image Processing Multilingual, Pre-processing, Segmentation, Unstructured

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 is the process of identifying main content of a web page which may consist of different forms of data in an unstructured and non-homogeneous manner. 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 Bilingual web page in present day context is shown in Figure 1.

*Author for correspondence

Figure 1. Variations in form, text and language levels in Bilingual webpage.

This web page 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. Content extraction for web pages can be considered as a pre-processing step for text mining and Web information retrieval. 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 1. 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. 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.

2. Related Work

Debnath[1] in his paper mentioned that Yi and Liu[2,3] have proposed an algorithm for identifying non-content blocks of Web pages. Their technique is intuitively very close to the concept of "information content" of a block. In order to identify the presentation styles of elements of Web pages, Yi and Liu's algorithm constructs a "Style Tree." A "Style Tree" is a variation of the DOM substructure of Web page elements. We have seen that our work even in the presence of advertisement images that vary from page to page can identify them as unrelated content by making use of the text in the blocks that are almost the same.

Another work that is closely related is the work by Lin and Ho[4]. The algorithm they proposed also tries to partition a Web page into blocks and identify content blocks. They used the entropy of the keywords used in a block to determine whether the block is redundant.

Cai et al.[5] have introduced a Vision-based Page Segmentation (VIPS) algorithm. This algorithm segments a Web page based on its visual characteristics, identifying horizontal spaces and vertical spaces delimiting blocks much as a human being would visually identify semantic blocks in a Web page. They use this algorithm to show that better page segmentation and a search algorithm based on semantic content blocks improves the performance of Web searches. Song et al.[6] have used VIPS to find blocks in Web pages. Then, they use Support Vector Machines (SVM) and Neural Networks to identify important Web pages.

Ramaswamy et al.[7,8] propose a Shingling algorithm to identify fragments of Web pages and use it to show that the storage requirements of Web caching are significantly reduced. Bar-Yossef and Rajagopalan[9] have proposed a method to identify frequent templates of Web pages and pagelets.

Kushmerick[10,11] has proposed a feature-based method that identifies Internet advertisements in a Web page. It is solely geared toward removing advertisements and does not remove other non-content blocks.

There has been substantial research on the general problem of extracting information from Web pages. Information extraction or Web mining systems try to extract useful information from either structured or semi-structured documents. Since a large percentage of dynamically generated Web documents have some form of underlying templates, Wrapper[10,11], Roadrunner[12], Softmealy[13] and other systems try to extract information by identifying and exploiting the templates. Systems like Tsimmis[14] and Araneus[15] depend on manually provided grammar rules. In Information Manifold[16,17], Whirl[18] or

Ariadne[19], the systems tried to extract information using a query system that is similar to database systems.

For other semi-structured wrapper generators like Stalker[20], a hierarchical information- extraction technique converts the complexity of mining into a series of simpler extraction tasks. It is claimed that Stalker can wrap information sources that cannot be learned by existing inductive learning techniques.

Most of these approaches are geared toward learning the regular expressions or grammar induction[21] of the inherent structure or the semi-structure and, so, computational complexities are quite high.

The efforts mentioned above are involved in extracting information that originally came from databases. This underlying data stored in databases is very structured in nature. Our work concentrates on Web pages where the underlying information can be structured, semi-structured and unstructured text[22]. The techniques used for information extraction are applied on entire Web pages, whereas they actually seek information only from the primary content of the Web pages.

As content extraction is different from text or data mining, where a set of keywords form the basis, most of the previous approaches use HTML tags to separate the main content from the extraneous items. This implies the need to employ a parser for the entire Web page. Consequently, the computation costs of these main content extraction approaches include the overhead for the parser. In the early stage of the Internet, the contents of Web pages were written only in English language. Now, especially in the last decade, a large part of information is being published in regional or native languages, like for example Spanish, Chinese, Tamizh, Hindi with native tongue and usage reflected in the text[23].

A simple Chinese web page looks like the one shown in Figure 2. 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. But, the content in terms of a person and his kid's photos and their appearances reaches a common man. This is typically an unstructured, heterogeneous and hyper media web page. Extracting content requires language, text and image processing. Extracting main content from web page is pre-processing of web information system. The content extraction approach based on wrapper is limited to one specific information source and greatly depends on web page structure. It is seldom employed in practice. So, a generic model 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 a short period of time [24].

Figure 2. Typical Chinese web page.

3. Nature and Features in Web Documents

As seen earlier web pages are unstructured and data presented in different forms from text to images to video, and multi-lingual depending on the audience and their location[25,26]. This gets more complex and involved when Asian or Indian regional web pages display information. Indian languages are very much different from other languages - like Japanese or Chinese - in that regional customs and practices bring in certain commonalities like the scripts of Tamizh or Telugu or Kannada have similarities of different kinds as compared to the northern Hindi or Punjabi scripts. But, English being the link language both in oral and written communication and forms the basis in higher education some complexities in migrating from English to regional language or vice-versa exist. Figure 3 shows a typical Chinese multi-lingual web page displaying news on the same day.

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 'Computer' in English translated in other languages like Hindi, Arabic, Tamizh and Telugu is shown in Figure 4 (a). But many times, popular words in one language are used as they are like word 'computer' in English is written in local scripts as in Figure 4 (b). Also one can see clearly the variations in structure of text in different forms and these do not form part of the local language dialect.

Figure 3. Find Cut Dimension – Pseudo Code.

Figure 4. (a) and (b) Complexities in Indian and Foreign languages with English (a) Word 'computer' translated in other languages, (b) Word 'computer' written in other languages.

4. Content Extraction - Results and Discussion

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. Any pixel map can be seen as a matrix of columns and rows with each element giving the colour scheme for the pixel. 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.

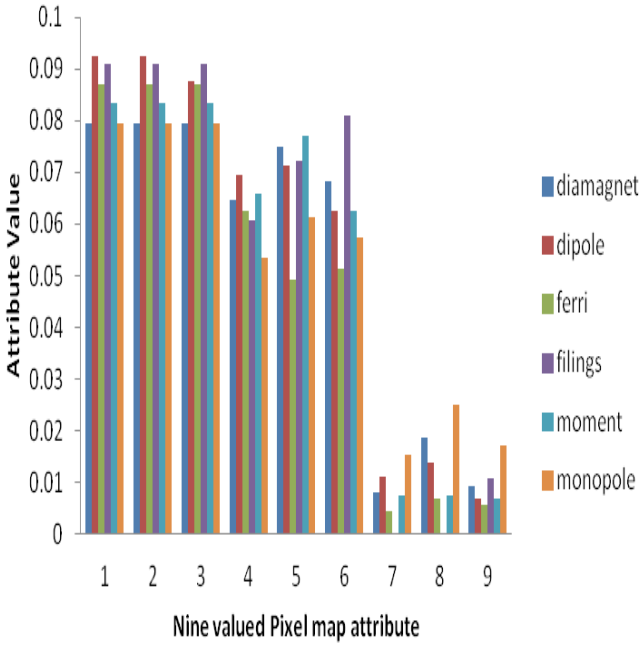

Figure 5. Pixel map attributes for six pixel maps considered.

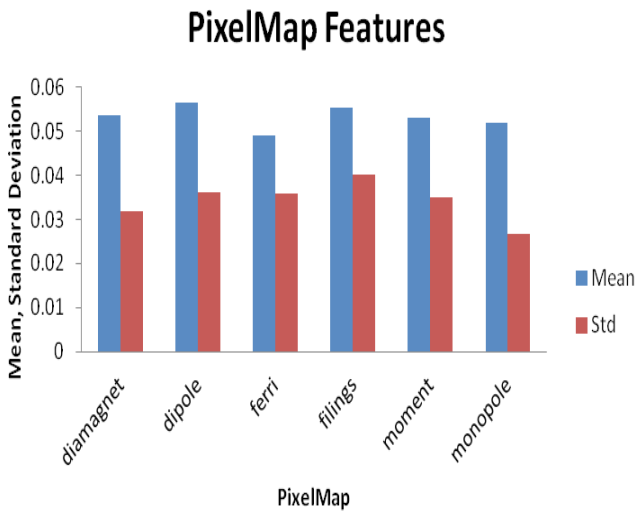

Figure 6. Mean and standard deviation variations for six pixel maps considered.

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 [10 x 11 x 3] matrix and this is reduced to [10 x 11] with 0 and 1 value and even then there are 110 values to reflect the matrix fully. Figure 5 gives clear idea of pixel map attributes for six pixel maps "diamagnet, dipole, ferri, filings, moment and monopole".

Figure 6 gives a comparison of mean and standard deviation for six pixel maps "diamagnet, dipole, ferri, filings, moment and monopole".

But, contents of the matrices are different and if processed in terms of either non-zero values – which gives the pattern or vector matrix values with content being same. This gives a clear idea of feature extraction. Since,

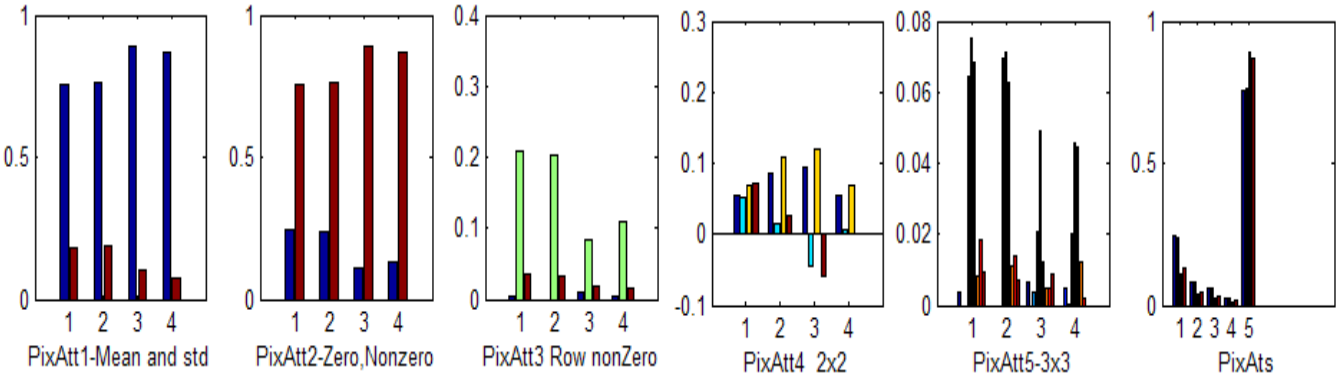

Figure 7. Variation of pixel map attributes for diamagnet, dipole, flower-Arabic, magnet-Arabic.

Asian language letters have characters surrounding the main body; the pixel map may be divided into three segments like 25% top, 50% middle and 25% bottom. Letters 'g' and 'y' in English have bottom 25% for example. And in the case of Arabic fonts, most of them have occupancy in top and bottom halves also. So, processing of text and documents ultimately has to be considered as a problem related to the content and context and natural language understanding. Figure 7 gives the variation of pixel map attributes for diamagnet, dipole, flower-Arabic and magnet in Arabic. This shows the clear variation of a word not related to magnetism when given as input, what happens to the attribute features. Magnet in Arabic gives the concept of multilinguality to show the attribute variations of different language but of same content.

The method described earlier is used with pattern recognition to compare whether any new input in the form of letter or word or image can relate to the content of base patterns. 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. The attribute variations of 'diamagnet' in comparison with 'magnet-Arabic'and 'dipole' are given in Figure 8.

One can see clearly that even though pixel map variations are significant, matching patterns can help in identifying the content. As an extension of this work, it is proposed to give a clear comparative analysis of Statistical Interpretation approach with Artificial Neural Network and Pattern matching studies, which will be the scope of our next publication.

5. Conclusion

A generic model for Content Extraction for regional web documents is developed based on the basic data system in computers, namely pixel maps. 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. This preliminary study is focused to bring out the complexities in regional web documents and how a generic tool based on pixel maps – which do not have language or form of data as inputs - can be used for either text mining or content extraction. Further enhancements and techniques

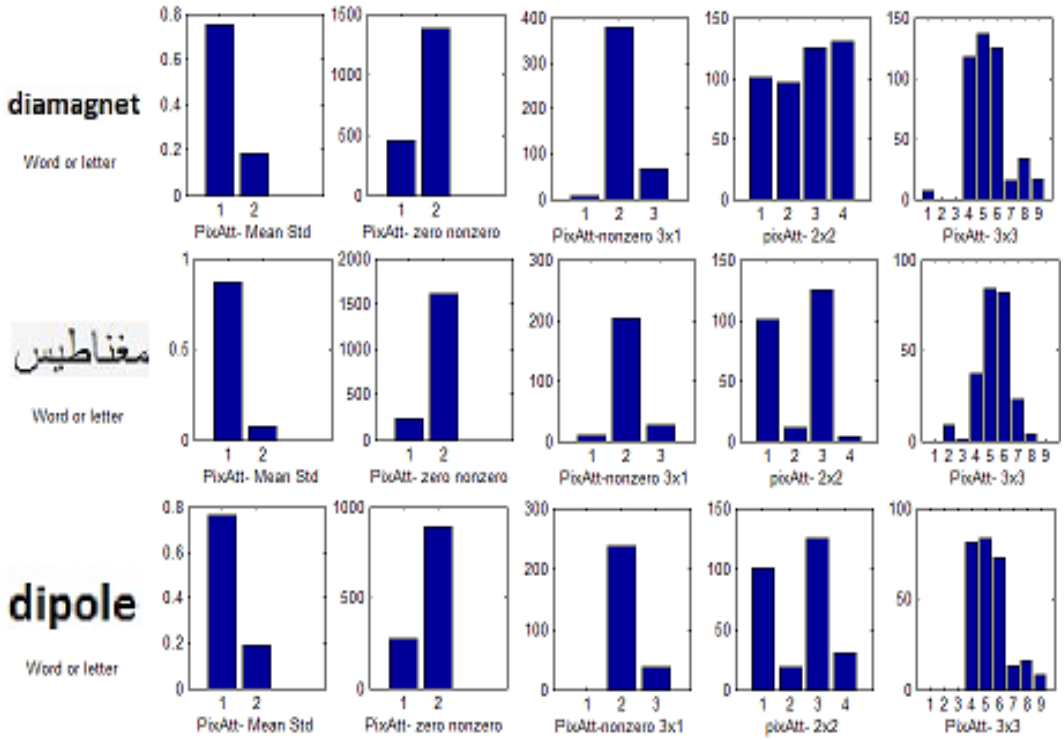

Figure 8. Five different attribute features comparison.

are to be suitably generated to account for the vagaries, so that, web content is extractable in any region.

6. Acknowledgment

The author likes to thank the management of Dr. M.G.R. Educational and Research Institute for their kind support during the preparation of this paper.

7. References

1. Debnath S, Mitra P, Giles GL. Automatic extraction of informative blocks from webpages. Proceedings of Special Track on Web Technologies and Applications in the ACM Symposium on Applied Computing; 2005 Mar 13-17; Santa Fe, New Mexico. p. 1722–6.
2. Yi L, Liu B, Li X. Visualizing web site comparisons. Proceedings of 11th International Conference World Wide Web; 2002 May 7-11; Honolulu, Hawaii. p. 693–703.
3. Liu B, Zhao K, Yi L. Eliminating noisy information in web pages for data mining. Proceedings of Ninth ACM SIGKDD International Conference Knowledge Discovery and Data Mining; 2003. p. 296–305.
4. Lin SH, Ho JM. Discovering informative content blocks from web documents. Proceedings of Eighth ACM SIGKDD International Conference Knowledge Discovery and Data Mining; 2002 Jul. p. 588–93.
5. Cai D, Yu S, Wen JR, Ma WY. Block based web search. Proceedings of 27th Annual International ACM SIGIR Conference; 2004 Jul. p. 456–63.
6. Song R, Liu H, Wen JR, Ma WY. Learning block importance models for web pages. Proceedings of 13th World Wide Web Conference; 2004 May. p. 203–11.
7. Ramaswamy L, Iyengar A, Liu L, Douglis F. Automatic detection of fragments in dynamically generated web pages. Proceedings of 13th World Wide Web Conference; 2004. p. 443–54.
8. Ramaswamy L, Iyengar A, Liu L, Douglis F. Automatic fragment detection in dynamical web pages and its impact on caching. IEEE Trans Knowledge and Data Eng. 2005 May; 17(5):1–16.
9. Bar-Yossef Z, Rajagopalan S. Template detection via data mining and its applications. Proceedings of WWW 2002; 2002 May 7-11; Honolulu, Hawai, USA. p. 580–91.
10. Kushmerick N. Wrapper induction: efficiency and expressiveness. Artificial Intelligence. 2000 Apr; 118(1-2):15–68.
11. Kushmerick N, Weld DS, Doorenbos RB. Wrapper induction for information extraction. Proceedings of International Joint Conference Artificial Intelligence (IJCAI); 1997. p. 729–37.
12. Crescenzi V, Mecca G, Merialdo P. Roadrunner: towards automatic data extraction from large web sites. Proceedings of 27th International Conference Very Large Data Bases; 2001; Rome Italy. p. 109–18.
13. Hsu C. Initial results on wrapping semi structured web pages with finite-state transducers and contextual rules. Proceedings of AAAI-98 Workshop AI and Information Integration; 1998 Aug. p. 66–73.
14. Chawathe S, Garcia-Molina H, Hammer J, Ireland K, Papakonstantinon Y, Ullman J, Widom J. The Tsimmis project: integration of heterogeneous information sources. Proceedings of 10th Meeting Information Processing Society of Japan; 1994 Oct; Tokyo, Japan. p. 7–18.
15. Atzeni P, Mecca G, Merialdo P. Semistructured and structured data in the web: going back and forth. Proceedings of Workshop Management of Semi structured Data; 1997 May.
16. Kirk T, Levy AY, Sagiv Y, Srivastava D. The information manifold. Proceedings of AAAI Spring Symposium Information Gathering from Heterogeneous Distributed Environments; 1995. p. 85–91.
17. Levy AY, Srivastava D, Kirk T. Data model and query evaluation in global information systems. J Intelligent Information Systems, special issue on networked information discovery and retrieval. 1995; 5(2):121–43.
18. Cohen WW. A web-based information system that reasons with structured collections of text. In: Sycara KP, Wooldridge M, editors. Proceedings of Second International Conference Autonomous Agents (Agents '98); 1998; New York, USA. p. 400–07.
19. Ambite JL, Ashish N, Barish G, Knoblock CA, Minton S, Modi PJ, Muslea I, Philpot A, Tejada S. Ariadne: a system for constructing mediators for internet sources. Proceedings of SIGMOD; 1998; New York, USA. p. 561–3.
20. Muslea I, Minton S, Knoblock CA. Hierarchical wrapper induction for semi structured information sources. Autonomous Agents and Multi-Agent Systems. 2001 Mar-Jun; 4(1–2):93–114.
21. Chidlovskii B, Ragetli J, de Rijke M. Wrapper generation via grammar induction. Proceedings of 11th European Conference on Machine Learning (Machine Learning: ECML 2000); 2000. p. 96–108.
22. Kolla BP, Dorai Ranga Swamy MA, Arun RR. ANN for Multi-lingual Regional Web Communication Part V. ICONIP LNCS. 2012; 7667:473–8.
23. Kolla BP, Dorai Ranga Swamy MA, Arun RR. Statistical interpretation for mining hybrid regional web documents. ICIP, CCIS. 2012; 292:503–12.

24. Kolla BP, Dorai Ranga Swamy MA, Arun RR. Performance of content based mining approach for multi-lingual textual data. International Journal of Modern Engineering Research. 2011; 1(1):146–50.
25. Oh Y. Sequential layered approach for optimized context integration. Indian Journal of Science and Technology. 2015 Apr; 8(S8):284–9.
26. Sathick KJ, Jaya A. Natural language to SQL generation for semantic knowledge extraction in social web sources. Indian Journal of Science and Technology. 2015 Jan; 8(1):01–10.