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.

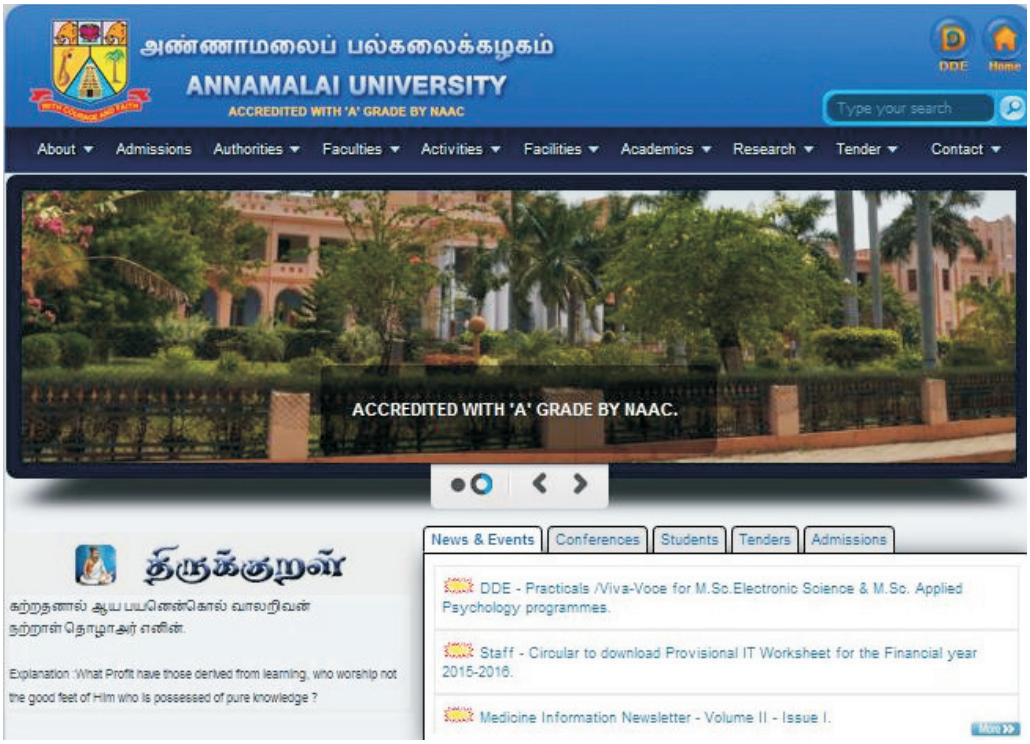

Fig. 2. An Example Bilingual Web Document in Indian Context.

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.