EXTRACTING CONTENT IN REGIONAL WEB DOCUMENTS WITH TEXT VARIATIONS

Kolla Bhanu Prakash[1], M.A.Dorai Rangaswamy[2]

[1]Research Scholar, Sathyabama University, Chennai, India
[2]Professor and Head, AVIT, Chennai, India
Email: 1bhanu_prakash231@rediffmail.com

Abstract

The growth of the World Wide Web has led to a dramatic increase in accessible information. Today, people use Web for a large variety of activities including travel planning, comparison shopping, entertainment, and research. However, the tools available for collecting, organizing, and sharing Web content have not kept pace with the rapid growth in information. Today people continue to use bookmarks, email, and printers for managing Web content. Use of mobile phones has transformed the culture of communication with even villagers using sophisticated computer-related words like SMS and MBBS. But the major complexity arises when web documents in regional languages are displayed. Understanding the content of the document and later communication through oral or text means becomes difficult and this is the area the current paper addresses and in the process tries a generic concept-based mining model is proposed, for how the knowledge is created in the minds of illiterate user. The paper first presents how letters and words which form the basis of text-based communication can be used for content. The objective of this task is to achieve a concept-based term analysis on sentence and document levels rather than a single-term analysis in the document set only.

***Key words**:* Media Mining; Features; Multilingual; Web Communication; Statistical Interpretation; Content Extraction;

I. INTRODUCTION

Natural Language Processing (NLP) is both a modern computational technology and a method of investigating and evaluating claims about human language itself. Text mining attempts to discover new, previously unknown information by applying techniques from natural language processing and data mining. Usually, in text mining techniques, the frequency of a term (word or phrase) is computed to explore the importance of the term in the document. However, two terms can have the same frequency in a document, but one term might be contributing more to the meaning of its sentence than the other term (1, 2, 3).

As the amount of content delivered over the World Wide Web grows, so does the consumption of information. Today people collect and organize information in ways different from before: in order to plan a vacation, they no longer go to a travel agent; when making purchases, they no longer go to a neighborhood specialty store; and when learning about a new topic, they no longer go to the library. Instead, they browse the Web. One study of Web usage in 2005 shows that users visit thousands of web pages per week and in any thirty minute browsing session may visit hundreds of individual pages. Although advancements in search technologies make it much easier to find information, users often browse the Web with a particular task in mind and are concerned not only with finding but also with collecting, organizing, and sharing content. These types of browsing sessions, which this works calls exploratory web research sessions, typically last a long time, span several sessions, involve gathering large amounts of heterogeneous content, and can be difficult to organize ahead of time, as the categories emerge through the tasks themselves. Research in current practices for collecting and organizing Web content shows that users save bookmarks or tabs, collect content in documents, store pages locally, and print them out. These methods require a great deal of overhead as pages must be saved manually and organized into folders, which distracts from the real task of analyzing the content and making decisions (4,5,6).

The focus of the current study is to extract content of such web documents without going through translation or conventional data mining approaches as most of the web documents have media-related data like images or

handwritten texts. This led to the basic premise on communication where the mind translates a text or media and later interacts or just interacts with content based understanding. The latter approach is taken and studies related to a particular discipline are presented to focus attention on how statistical interpretation supports content extraction faster.

II. CONTENT EXTRACTION TECHNIQUES

The term "Content Extraction" was first coined by Rahman et al. [7] along with first and very basic algorithm. Finn et al. [8] discusses method (BTE) to identify useful information within the "Single Article" document, where the content is presumed to be in a single body. The method depends upon the tokenization of contents of the document into tags or words. The document is divided into three contiguous sections and boundaries are estimated in such a way to maximize the number of words within the sections. The method is useful for the documents having simpler layouts but do not produce the desired results with the dynamic content sites and complex layouts designed to make most of space available on a web page.

McKeown et al finds the body of a tag with the highest weight in terms of text present within the body. The detected body is considered to be the one containing useful information and the rest is ignored. Gottron applied Document Slope Curves (DSC), an extension to BTE algorithm to create Advanced DSC which employs windowing technique to identify the regions in the document containing content. Mantazir's used the amount of text within the anchor tags to identify the navigational lists. The idea was that to remove the navigational links from the document and the approach is named as Link Quota Filter (LQF). Largest Size Increase (LSI) algorithm works out the nodes of DOM tree which contributes most to the visible content in the rendered document.

Debnath et al. developed feature Extractor (FE) and KFeature Extractor (KFE) based on the block segmentation of HTML document. Another approach is to extract content from HTML web page is Content Code Blurring (CCB) [9]. It works on identifying the maximum number of formatted source code characters to identify the content blocks.

Another popular algorithm is Visual Page Segmentation (VIPS) that heuristically segments the document into visually grouped blocks. However, the weakness of the technique is that it does not classify those blocks into the content and non-content blocks. Kaasinen et al., Buyukkokten et al. and Gupta et al. proposed models based on DOM and conversion of DOM to WML for detection of useful information, but these all focus towards improved visibility of the page. These all models have some sort of markup in the output, so all of them fail to filter markup text from the HTML web document. Especially, the model proposed by Gupta et al works fine to a certain extent in reduction of unwanted material within a HTML document.

There are also some attempts to detect the template of the web pages [10] for content extraction. The approach employed supervised classification models and needed a collection of training documents in order to train the classification models. One of the examples of such strategy is Bar-Yossef et al., which automatically detects the template of the web page from the Largest Pagelet (LP). The weakness of the approach is that it depends upon the template of the web page and does not perform adequately on the data sets with the web pages of varying templates. This weakness points towards the necessity of having a generalized solution which can perform well on varying templates or web page designs.

Content Extraction via Tag Ratios (CETR) is a better and generalized solution, which can work with variety of web page designs and templates. The basic principle of the approach is to calculate the tag ratios in different parts of HTML document. There are two more variations of CETR approach i.e. CETR-TM which is based on thresholds, CETRKM based on K-Mean clustering to extract the content from HTML web page. CETR is found to have best performance and the approach does not depend upon the web page designs or templates. However, the only weakness of the approach is that it fails to extract content when the quantity of content present at any part of document is lower. For example, the user comments embedded in many web pages contains comparatively less quantity of text than the remaining content of the web page. CETR is found to have reduced accuracy in such web pages.

Our proposed models objective is to develop a generic model which can apply for complexities to check whether it is possible to assess the content in a short period of time.

III. FEATURES OF FOREIGN AND INDIAN REGIONAL LANGUAGES

Indian languages are very much different from European –German or Russian-or other Asian languages- like Japanese or Persian, 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 communication and forms the basis in higher education, some complexities in migrating from English to regional language or vice versa exist like the ones shown in Fig.1. Here a word 'computer' in English, if written in other languages Arabic, Hindi, kannada, Russian, tamizh and Telugu as shown in Fig 1, shows clearly the variations in structure of text in different forms.

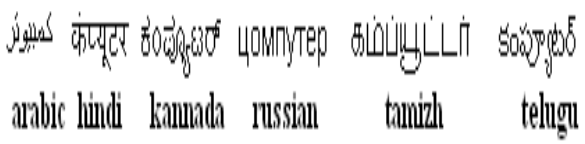

Fig. 1. Complexities in Indian and Foreign languages with English

Web documents developed in Indian regions vary differently in content as well in form and Fig.2 gives one such a typical example of IIT Madras university home page wherein we find both English as well as Hindi, content in translated form.

Fig. 2. Web document variations in form and text -local level

It may be seen clearly the content is different and similar complexity with regional news exists in India and Fig.3 shows this with webpage in English, Bengali and Tamizh and we can see that even the news content varies as the region is changed.

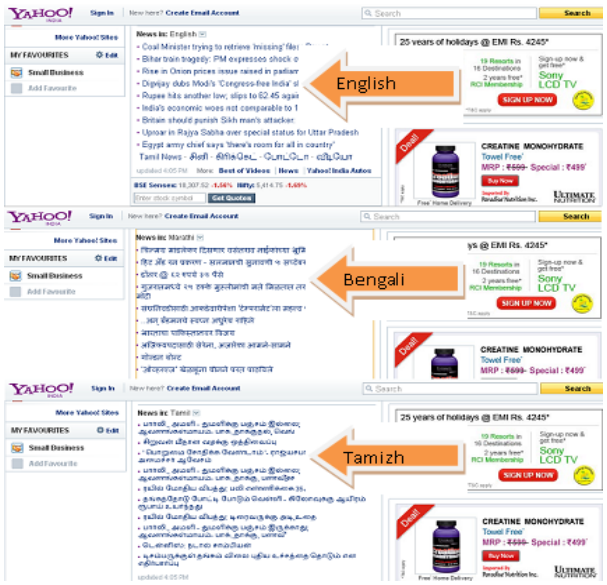

Fig.3 Web documents in regional India –different languages, different contents

For example, if we compare Bengali and Tamizh, there is variation in news content; we find news in their region rather national in both the cases. So, if one wants to continue surfing and later interact, content of the web is the only way to go about. It may be seen clearly that regional web documents pose different problems in terms of comprehension, understanding and interaction in another language. Hence, it is preferable to assess the content even before looking at the document fully. Images and figures do help, but many times texts and sketches with words pose problems as they reflect local dialect and flavor. So, it is necessary to assess the content irrespective of the language and the way text is produced.

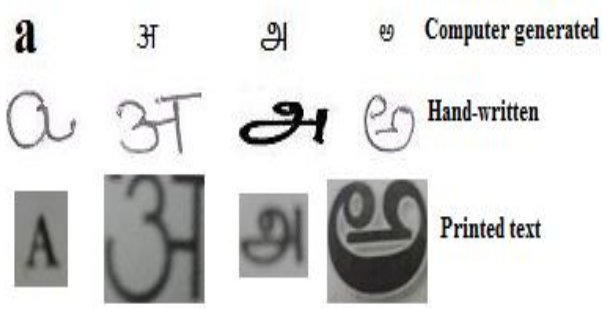

Fig.4 Character variations in CG, HW and PT formats

Hence, the objective is to develop a generic model and later apply for complexities to check whether it is possible to assess the content in a short period of time.

IV. RESULTS AND DISCUSSION

One of the basic steps in mining is processing the data as it is, may be in different forms like pictures, texts or media or in different formats either in full form or compressed form. So, the pixel-map of any data can be seen as a matrix of columns and rows, with each element giving the color scheme for the pixel. Typically, the matrix could be [Nx, Ny, Ip] where, Nx indicates the rows, Ny the columns and Ip the value of each element. The matrix is rectangular with rows being less than columns, mostly in the case of texts and letters and the values Ip may be either 0 or 1 for black and white, and 0 to 255 for color depending on the resolution. So, the characteristic and attribute of any pixel map can be deduced from these three values and most of image processing and data mining depend on this basic matrix.

This is true in education, where text books written by authors in regional languages are used in web pages. Typically, in Fig.5 a page of Physics text book is shown in two languages English and Tamizh. Here, again one can see equations in English are used as they are in Tamizh scripts. Keeping in view all the above mentioned issues, it is preferable an approach based on extracting the content of the document rather than translation or data-mining seemed better and here, statistical approach is presented.

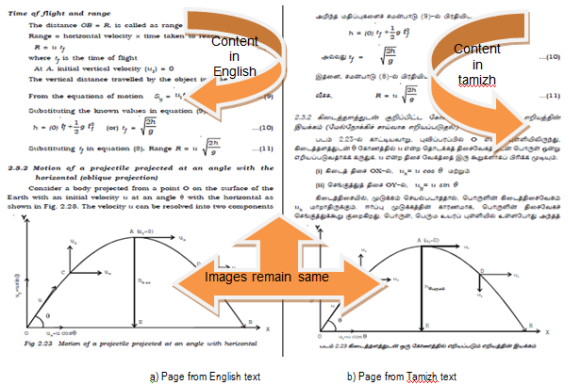

Fig.5 Text book page in two languages

A web document may contain texts, images, audio/video files and in regional documents hand-generated ones. Many a time, the document may have been developed and a scanned version is used in the web. Figure.6 gives a comparison of features of “a” in three different formats – handwritten, printed text and computer generated texts in English. This gives us a clear essence of how characters vary with structure and crispness, when we go for handwritten instead of computer generated texts format. This gives us a clear idea of feature extraction. Since, regional language letters have characters surrounding the main body; the pixel-map is divided into three segments like 25% top, 50%middle and 25% bottom. Letters ‘g’ and ‘y’ in English have bottom 25% for example.

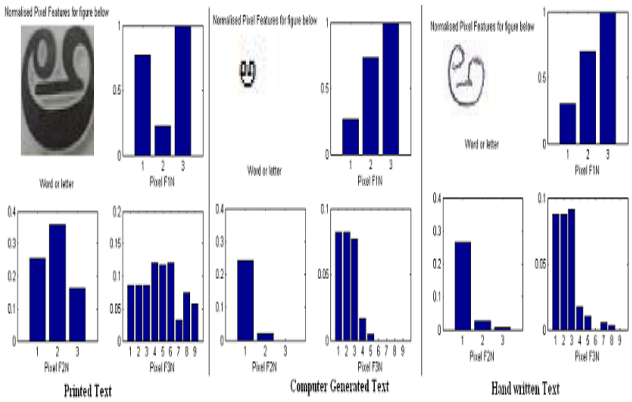

Fig.6 Feature variations for letter ‘a’ in Telugu in three different formats- PT, CG, HW

And if we see Arabic fonts most of them have occupancy in top and bottom halves also. This paper presents a statistical approach, learning of language models for context dependent, word occurrences and discusses the applicability of this model for interactive information access. 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. Character ‘a’ which is actually unique in content similar in four languages – English, Hindi, Telugu and Tamizh. Uniqueness of letter ‘a’ is, it has same meaning in all the four languages. Fig.7 gives the mean and standard deviation values for a character ‘a’ in all the three formats- computer generated, hand-written and printed text formats, for all the four languages considered. A comparison of these three forms with base values gives us an idea of how much the variation is.

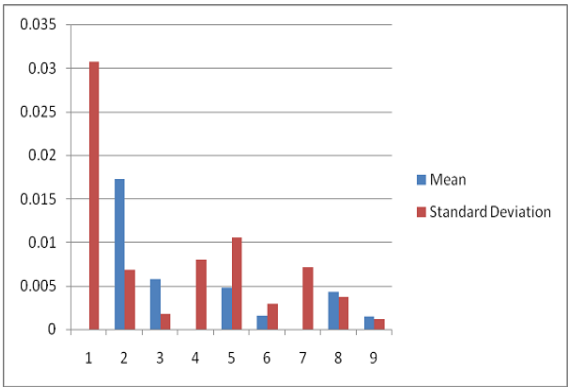

Fig.7 Mean and Standard deviation variations for character 'a'

The purpose of choosing 'ke' in English is that, it is unique to regional languages, but, not unique to English. In Hindi, it requires only one character to write 'ke' but, we need two characters in English; which is another interesting feature. Fig.8 gives the mean and standard deviation values for a character 'ke' in all the three formats- computer generated, hand-written and printed text formats in four languages considered.

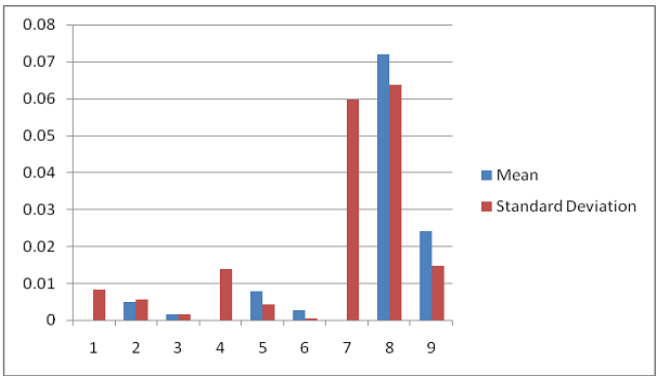

Fig.8 Mean and Standard deviation variations for character 'ke'

A comparison of these three forms, with base values gives us an idea of how much the variation is. The importance of choosing 'I' is that, it is a single character and word in English, but in other languages it requires more than one character to write that word. This is another interesting feature of importance. Fig.9 gives the mean and standard deviation values for word 'I' in all the three formats- computer generated, hand-written and printed text formats in four languages considered. A comparison of these three forms with base values gives us an idea of how much the variation is.

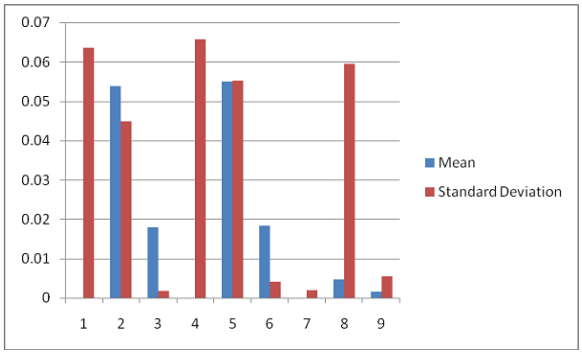

Fig.9 Mean and Standard deviation variations for word 'I'

Here, features are classified into one value, three values and nine value attributes and mean and standard deviations are calculated in all the three formats - computer generated, hand-written and printed text formats, in four languages considered and compared with the base value. The difference in comparison is calculated and the variation is shown clearly in fig.9.

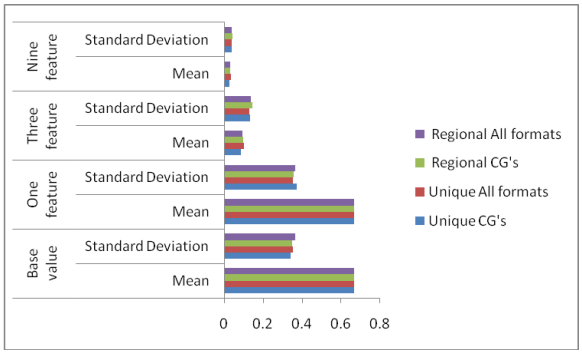

Fig.10 Table showing comparison values for 'a' and 'ke'

And at last fig.10 gives the table showing comparison values for 'a' and 'ke', mentioning about base value, one-valued attribute, three-valued attribute and none-valued attribute which were discussed before.

V. CONCLUSION

This work bridges the gap between natural language processing and text mining disciplines. A new content based mining model is proposed to improve the text mining to a precision level higher than achieved with other conventional data mining or natural language processing methods.

REFERENCES

[1] Y. Li, C.-C. J. Kuo and X. Wan, Introduction to content-based image retrieval — Over view of key techniques, in Image Databases: Search and Retrieval of Digital Imagery, eds. V. Castelli and L. D.

Bergman (John Wiley, New York, 2002), pp. 261–284.

[2] Honkela, T., Kaski, S., Lagus, K., and Kohonen, T. (1996). Exploration of full-text databases with self-organizing maps. Submitted to ICNN-96, Washington D.C.

[3] Honkela, T., Pulkki, V., and Kohonen, T. (1995). Contextual relations of words in Grimm tales analyzed by self-organizing map. In Fogelman-Soulie, F. and Gallinari, P., editors, Proc. ICANN-95, Int. Conf. on Articial Neural Networks, volume II.

[4] Kirstie Hawkey and Kori Inkpen. Web browsing today: the impact of changing contexts on user activity. In CHI '05: CHI '05 extended abstracts on Human factors in computing systems, pages 1443–1446, New York, NY, USA, 2005. ACM Press.

[5] William Jones, Harry Bruce, and Susan Dumais. Once found, what then? A study of "keeping" behaviors in the personal use of web information. In Proc. of ASIST, 2002.

[6] Abigail J. Sellen, Rachel Murphy, and Kate L. Shaw. How knowledge workers use the web. In CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems, pages 227–234, New York, NY, USA, 2002. ACM.

[7] F. R. Rahman, H. Alam, and R. Hartono. Content extraction from html documents. In WDA, pages 7–10, 2001.

[8] Finn, N. Kushmerick, and B. Smyth. Fact or fiction: Content classification for digital libraries. In DELOS Workshop: Personalization and Recommender Systems in Digital Libraries, 2001.

[9] T. Gottron. Content code blurring: A new approach to content extraction. In DEXA Workshops [1], pages 29–33.

[10] R. Cathey, L. Ma, N. Goharian, and D. A. Grossman. Misuse detection for information retrieval systems. In CIKM, pages 183–190. ACM, 2003.