International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol.1, Issue1, pp-146-150 ISSN: 2249-6645

Performance of Content Based Mining Approach for Multi-lingual Textual Data

Kolla Bhanu Prakash
Research Scholar
Sathyabama University
Chennai, INDIA

M.A.Dorai Rangaswamy
Dept. of Computer Science and Eng.
AVIT
INDIA
Member IEEE

Arun Raja Raman
Retd. Professor, IITM
Chennai, INDIA
Member IEEE

Abstract— **Data mining has become a necessary and powerful tool in the present era of web and internet communications. It has also evolved into media mining wherein heterogeneous data inputs like figures, videos and audios are gradually getting embedded into the web and this makes it quite complex and different. These and other aspects like currency and 'liveliness' of the web bring in more interesting features making a shift from translation to especially content extraction. Content extraction in web pages with Indian regional languages or English as the parent language have many aspects like free use of one language in another like 'computer' being used as it is with regional text and inclusion of other forms of data like hand written texts or sketches or drawings. This is common in education, news and entertainment and the focus of the current paper is in extracting content in a hybrid document with hand-written texts embedded. Work has been carried out initially with web documents in the form of computer generated text since they are more crisp in nature. Extending the idea, the present paper discusses on the results of hand-written text format and a comparative study with computer generated text format, which are less crisp in nature and more fuzzy depending on the writer. Beginning with letters having common content to words with common content, results of features on pixel maps are presented first. Later extraction using normalisation studies and classification means are presented.**

I. Introduction

Web communication is becoming increasingly a powerful medium in variety of disciplines and with the spread of mobile and ad-hoc networks, it is an essential component in many areas of application. But with English as the main language used in the development in many of these conceptual and innovative applications, its adoption in regional and multi-lingual level needs more and more extensive work. One of the main problems here is in assessing the content of a web document and NOT the translated version of it, which may take more time when one searches content related information on-line perspective. So a content mining approach based on the file format of the web document is needed and this is developed so that the user or the node can react immediately for getting an in-depth view of a particular aspect in the document. In an earlier study, this idea was given for web documents having computer-generated texts. But many times the web documents may contain hand written texts in a different language or same language. It is the focus of the present study to look into aspects dealing with web documents either in English or in a regional language like Tamizh or Telugu or Hindi, prepared in different modes.[1,2,3,4,5] The study and results are presented for letter content in a text with three variations.

II. Character of Present Day Web Documents

Web documents are prepared in different ways with HTML occupying a standard form for developing web pages. But if one looks at the documents generated by a browser for presenting various aspects, the contents might differ. Fig.1 shows a typical multi-lingual web pages with varying characteristics. In Fig.1 (a) the web document shows image, icon and description in two languages in Tamizh and on the right hand side in English and it may be noted that left text is a translation of the right text in English. Fig.1 (b) shows another web page where the texts in regional language are literal replication of what is written in English like the word 'computer', being used in all language texts.

a) Multi-lingual web page with translation

b) Multi-lingual web page with replication. in regional text

Figure.1 Variations in multi-lingual Web pages

Even confining to one language like English, web pages in different regions show different content, depending on which is current in that region. This divergence is shown in "Fig.2".The web page in World on top is completely different from the one for Asia, which again is different from what is in USA as shown in "Fig.2".

Fig.2 Web pages on the same day in different regions.

So content extraction is more needed than literal translation of the document.[6,7,8] Many times the format of the web page is such that video, audio and text are in built in such a way that content is very apparent from the audio and video so that the user or node can decide immediately which is his need for further browsing. With this in view, a method based on pixel maps alone which any computer can 'understand' and can use to extract content, was developed and detailed elsewhere [17]. In this study the performance of content extraction with reference to text and character variations of different languages are discussed to form the basis for classification and training.

III. Content Based Approach For Textual Document

Pixel maps of any text or figure or audio form the basis of storage transfer and interaction in any computer and though many formats are there beginning with .bmp to .png or .jpg, the jpg format has become universal for transfer, download and interaction between browsers and other application softwares. Here these are considered as the bases for content extraction. Beginning with letters and then on to words, content similarity is in different levels of usage and communication. So the study classifies text of different languages into three cases viz.,a) letters having same content in most of the languages –CER(Content same in English and Regional language) , (b) letters unique to English and not present in others-CE and c) letters peculiar to other languages but not present in English. –CR. Later the study uses the pixel maps to convert into three parameter vectors defining attribute of that pixel map and later normalized with reference to parent one to get an index for the study and training[12,13,14].

Every pixel map after generating the code returns a vector consisting of three values. The first value indicates the occupancy ratio of pixels at the top portion, second value indicates the occupancy ratio for the center portion and the third value indicates the occupancy ratio of pixels at the bottom portion respectively. The sum of three values will be equivalent to 1.

In the present example pixel maps in four languages, English,Tamizh,Telugu and Hindi, are taken to get four vectors resulting in a 3x4 matrix. Later these vectors are normalised with parent language say English and a matrix of order 4x4, is generated and this done by differential and ratio-based normalisation. In this matrix, the variations among the elements of the diagonal are of great use for predicting similarity in content and this is given in the present study.

IV. Performance of Content Based Approach

Letters and words in different languages have their own unique and distinctive features; but with English dominating the web in the last two decades, a tendency to use words mutually in English and regional languages has become popular. For example the word 'computer' is used as it is in many languages and communication. So content extraction calls for similar and dissimilar features in letters and words for better assessment of pixel map attributes. As mentioned before, the classification CER,CE and CR type interpretation is used for content study. Examples of CER are shown in Fig. 3 where letters have same content. Letters unique in English-CE- are shown in "Fig.4" and letters unique in other languages-CR- are shown in "Fig.5"

Fig.3. Pixel-maps of characters in four languages of same –CER-content

Figure 4. Letters unique to English-CE

Figure 5. Letters unique to regional language-CR

Normally text in any languages consists of words formed in a certain structured way and each of these words consists of characters native to the language in which the text is prepared. So it is preferable to look at extraction of features in characters[14,15,16,17] and here one can see how it is quite complicated between English and any regional language like Tamil or Telugu or Hindi.

Fig.6 gives a comparison of individual features of four letters in four languages having the same content in computer generated text format.

This gives us a clear idea of feature extraction in four different languages taken into consideration same content.

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.

Later on we compare this with the handwritten text format which gave interesting features due its different crisp nature.

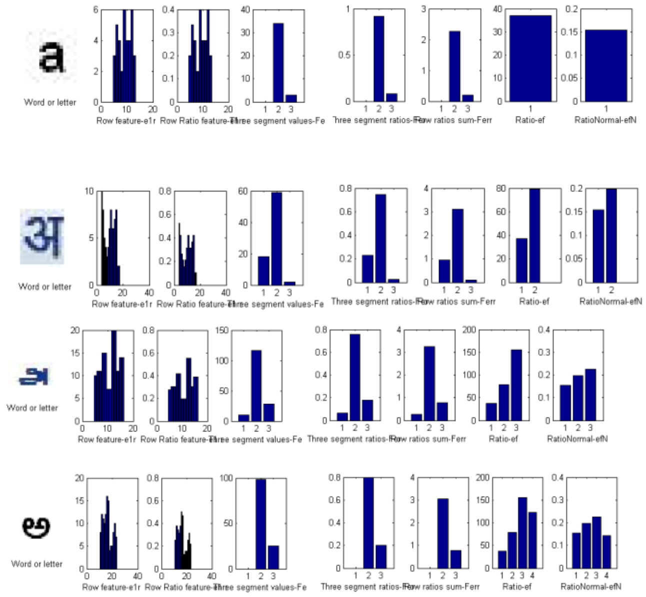


Figure 6. Feature extraction for CER-letter 'a' in CG format

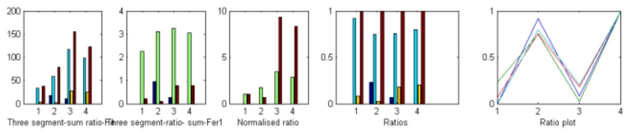
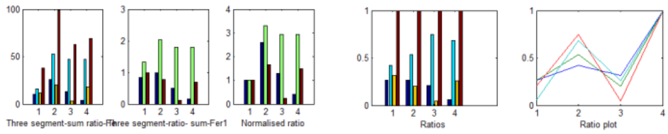
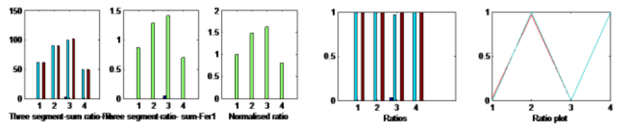


Fig.7 Feature comparison in letters in CER

Fig.7 clearly shows the variation in features in three segments and this can be used for classification and training. Similar figures could be obtained for CE and CR texts and these are shown in Fig.8 given below.



a) Feature comparison for CE letter 'x'



b) Feature comparison for CR-letter 'ke'

Fig.8 Feature comparison for CE, CR.

Now the performance of this approach is assessed through a new pixel map and the possibility of that belonging to CER, CE or CR is presented in terms on variations.

V. Results and Discussion

Pixel map of letter 'y' is taken for assessing the performance as it is clear that it does not belong to any of the three categories CER, CE and CR mentioned earlier. The vector representing the letter in terms the basic three parameters is taken and normalization with pixel maps used as parent ones is done and after converting to 3x3 matrix variations are presented for both difference and ratio approach. The variations are shown in Fig. 9 with CER CE and CR values.

Type	Range(Dmatrix)	Range(Rmatrix)
CER	-1.33% to 4.91%	41.43%
CE	0% to 18.07%	41.27%
CR	-1.78% to 8.45%	24.91%
'y'(CER)	-3.1% to 5.36%	0.2513 to 2.3328%

Fig. 9 Performance of 'Y' pixel map

Once a new pixel map is later taken in four languages and the minimum and maximum variation is observed in the diagonal values of the resultant matrix, say for eg., letter 'x' for differences the minimum and maximum variation is found to be 0.0009 and 0.1679 respectively; whereas for letter 'a' the minimum and maximum variation is found to be 0 and 0.1769 respectively.

So, We have two ways of identification one by difference which is algebraic and another which is more rational. So the conclusion for 'x' is a) if we use diffMatrix attribute probability of 'x' belonging to CER is to the extent of (.0009+.1679)/(0+.1769) which is 90% and this means content is similar to 'a'.b) if we use ratiMatrix the value is 7.1/.63 >1 so content is not 'a'. Similar approach can be applied to any new pixel map and we can find out whether the given pixel map belongs to CER or not.

In order to assess the performance of CER,CE and CR in the present paper we have taken a new pixel map 'y' and did the same procedure and made a comparison with all the three cases in both difference and ratio. The detailed study is given below.

In the case of differences for CER the variation is found to be 63%, for CE the variation is 52%, for CE the variation is 19%. So, in this category we can give a conclusion that the considered pixel map may belong to the category CER.

In the case of ratios for CER and CE the variation is found to be 41% and for CR the variation is 25%. So, the probability that the considered pixel map may belong to CER is more compared to the other two categories. The variations are shown in the form of histograms in Figures.10(a) and 10(b).

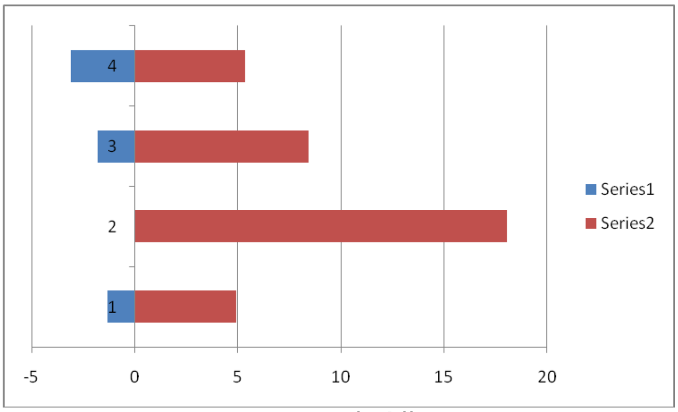


Fig.10(a) Feature comparison for differences

In observing both the categories differences and ratios we can conclude strongly that the pixel map 'y' may belong to CER category, which says about the content unique to English and other regional languages.

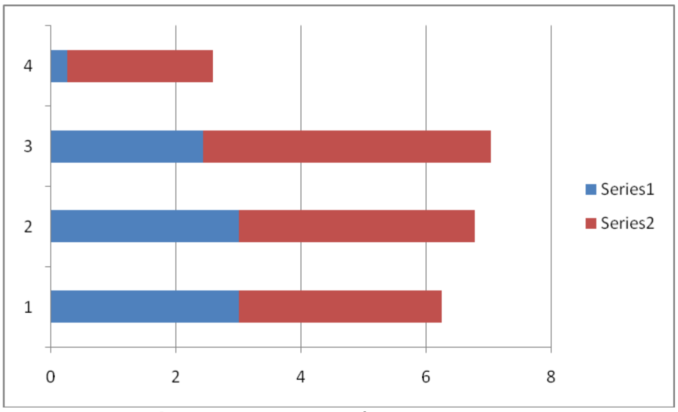


Fig.10(b) Feature comparison for ratios

Similar approach can be applied to any new pixel map and we can find out whether the given pixel map belongs to CER, CE or CR.

In the next case we combined all CER,CE and CR into one category which tells us about single concept that is all of them talk about same content. Same pixel map 'y' is taken for analysis of performance. And it is observed that for differences the probability is 40% and for ratios the probability that 'y' belongs to the above category is 37%. Since, the probability is very less we can say that 'y' may not belong to the above category.

Since the web document can contain hand written, computer generated or both the formats, the model that is discussed here should support any of the formats. Identifying the features obtained in both formats and comparing the both it is observed that 10% variation at the top portion, less than 5% variation in the middle portion and 20% variation in the bottom portion between both the formats for CER letter 'a'.

So, in the present discussion we are converting hand written text features into computer generated format. This can be done in two different ways. First ratios between the both formats can be calculated. Second differences between both the formats are calculated. Since the variations may not be completely satisfactory, these can be further normalized to give better results.

Till now all the discussion given above is with computer generated texts. But, often we find hand written text also available as web documents. So, as a final study we compared the pixel map with the handwritten text category also and the conclusion for this is given below.

It is observed that for differences the probability is 43% and for ratios the probability that 'y' belongs to the above category is 40%. Since, the probability is very less we can say that 'y' may not belong to the above category.

VI. CONCLUSIONS

Extraction of content in multi-lingual web documents is essential for education and other activities on the net so that the user can surf on interested areas immediately. A method based on feature extraction for words in multi-lingual documents is developed and the complexities and numerical aspects are discussed for typical examples. The examples are from letters to words bringing out the need to include character variations in developing the mining approach.

REFERENCES

[1] Rafael C. Gonzalez, Richard E. Woods, Steven L. Eddins "Digital image processing using matlab", 2002.

[2] Renu dhir "Feature extraction and classification for bilingual script (Gurumukhi and Roman)", April 2007.

[3] Bing Zhao, Stephen Vogel "Adaptive parallel sentences mining from web bilingual news collection", 2002.

[4] S.-C. Chen, S. H. Rubin, M.-L. Shyu and C. Zhang, A dynamic user concept pattern learning framework for content-based image retrieval, IEEETransactions on Systems, Man, and Cybernetics: Part C 36(6) 2006) 772–783.

[5] Z.-N. Li and M. S. Drew, Fundamentals of Multimedia (Prentice Hall, NJ, 2004).

[6] L. Pan and C. N. Zhang, A criterion-based role-based multilayer access control model for multimedia applications, in Proc. Eighth IEEE Int. Symposium on Multimedia (San Diego, CA, USA, 2006), pp. 145–152.

[7] G. Lu, Multimedia Database Management Systems (Artech House Publishers, Boston/London, 1999).

[8] Y. Li, C.-C. J. Kuo and X. Wan, Introduction to content-based image retrieval —Overview of key techniques, in Image Databases: Searchand Retrieval of Digital Imagery, eds. V. Castelli and L. D. Bergman (John Wiley, New York, 2002), pp. 261–284.

[9] Q. Iqbal and J. K. Aggarwal, CIRES: A system for content-based retrieval in digital image libraries, in Proc. Int. Conf. Control, Automation, Robotics and Vision (ICARCV) (Singapore, 2002), pp. 205–210.

[10] A. Kuchinsky, C. Pering, M. Creech, D. Freeze, B. Serra J. Gwizdka, Fotofile: A consumer multimedia organization and Retrieval system, in Proc. ACM CHI Conference(New York, NY,USA, 1999), pp. 496–503.

[11] A. Gupta and R. Jain, Visual information retrieval, Communications the ACM 40(5) (1997) 71–79.

[12] A. Pentland, R. W. Picard and A. Sclaroff, Photobook: Content based manipulation of image databases, Int. J. Computer Vision 18(3) (1996) 233–254.

[13] Kolla Bhanu Prakash, M.A.Dorai Ranga Swamy, Arun Raja Raman "Mining approach for documents containing multilingual Indian texts"(NCRTAC-09), Bharath University, Chennai.

[14] Kolla Bhanu Prakash, M.A.Dorai Ranga Swamy, Arun Raja Raman "A Two-Input Neuron model for documents containing multingual Indiantexts" (EPPCSIT-09), Guru Nanak Dev Engineering College, Ludhiana.

[15] Kolla Bhanu Prakash, M.A.Dorai Ranga Swamy, Arun Raja Raman "Feature extraction for content mining in multi-lingual documents" (NCICN 2010), Sathyabama University, Chennai.

[16] Kolla Bhanu Prakash, M.A.Dorai Ranga Swamy, Arun Raja Raman "Text Studies Towards Multi-lingual Content Mining for Web Communication" (TISC2010), Sathyabama University, Chennai.

[17] Kolla Bhanu Prakash, M.A.Dorai Ranga Swamy, Arun Raja Raman "Content Extraction for Multi-lingual Web documents" CIT Journal of Research, Volume 1, Issue 3, NOV 2010, pp.93-101.