Sind Univ. Res. Jour. (Sci. Ser.) Vol: 37 (2) 63 -68 (2005)

Determination of the Frequency of Occurrences in Sindhi Language Alphabet

A. Shah and R. Shaikh

Department of Computer Science, Isra University, Hyderabad

Abstract

Sindhi language is based upon a set of fifty-two characters. Some of these characters occur more frequently than others. The data compression algorithm exploit this behavior of the language character set to save bits allocated per characters. This in turn saves the storage space and the transmission over load on communication facilities such as internet channels. In this paper a detailed study of such statistics is presented.

Keywords: Frequency, Occurrence, Sindhi alphabet

Introduction

There has been recently a trend among researchers to develop computer applications in regional languages. These applications can also provide information over the Internet to other users. Some of these applications may require huge bandwidth and storage facilities, such as transmission of a newspaper-related material from one user to another user. This causes overload on the communication channel and other computer facilities. In order to reduce the load and improve the efficiency of the system, compressing the real data of the source can help in this regard.

All the languages are based upon certain number of alphabetic characters. When languages are spoken or written each of these characters do come with certain frequency of occurrence (Abramson, 1963). For example, in English language character E comes with highest frequency of occurrence http://Data-Commpression.com], so is the case of regional, national languages.

Sindhi language is based upon 52 characters set (Sindhi Text book for class I, 2000). It has been observed that the character "Yaah(ي)" comes with highest frequency (Shah and Shaikh, 2001). The next is "Alaf(ا)" and the least frequency of occurrence is observed for "Phay(ڦ) " The theory of data compression indicate that the characters having higher frequency of occurrence can be coded for lower number of bits to save storage space and transmission bit rate. For those characters which occur with least frequency, the number of coding bits allocated is high (Nelson and Gailly 1995). In succeeding sections of this paper the determination of statistical measures, and other work is detailed, including results, discussion and conclusion.

Determination of Probabilities

In order to determine the probabilities of each character of Sindhi language, a set of 50 passages has been considered. The Average Probabilities for each of the character are determined as given below:

$$Av_{(prpb)1} = \frac{\text{Passage 1 + Passage 2}}{2} \quad \text{eq.1}$$

$$Av_{(prpb)2} = \frac{\text{Passage 1+Passage 2 Passage 3}}{3} \quad \text{eq. 2}$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$Av_{(prpb)49} = \frac{\text{Passage 1+Passage2+...Passage 50}}{50} \quad \text{eq.49}$$

Graphical representation of these $Av_{(prpb)49}$ is indicated in Figure 1-4 for characters "Yaah (ي)" Raay (ر)""Giay(گ)" "Bhay (پ)" respectively.

Discussions

The rigorous study indicates that some of the characters, especially those with higher frequency of occurrences get stable after certain number of sample such as character "Yaah (ي)" and character "Raay (ر)", group–1 and group–11 characters respectively. However, the characters with lower frequency of occurrences do not get stable with even fifty samples of our experimentation. To determine their AV (Prob) more samples can be included in sample space.

Conclusion

The results indicate that the character "Yaah(ي)" has an Av (prob) of 0.13 that is 13% of the time.The character "Raay(ر)", has $Av_{(prob)}$ of 0.06 that is 6 % and the character "Bhay(پ)" has not yet an accurate result, but it is approximated that it is 0.3%. However the character "Giay(ﮔ)" occurs with least frequency and it is 0.1 %.

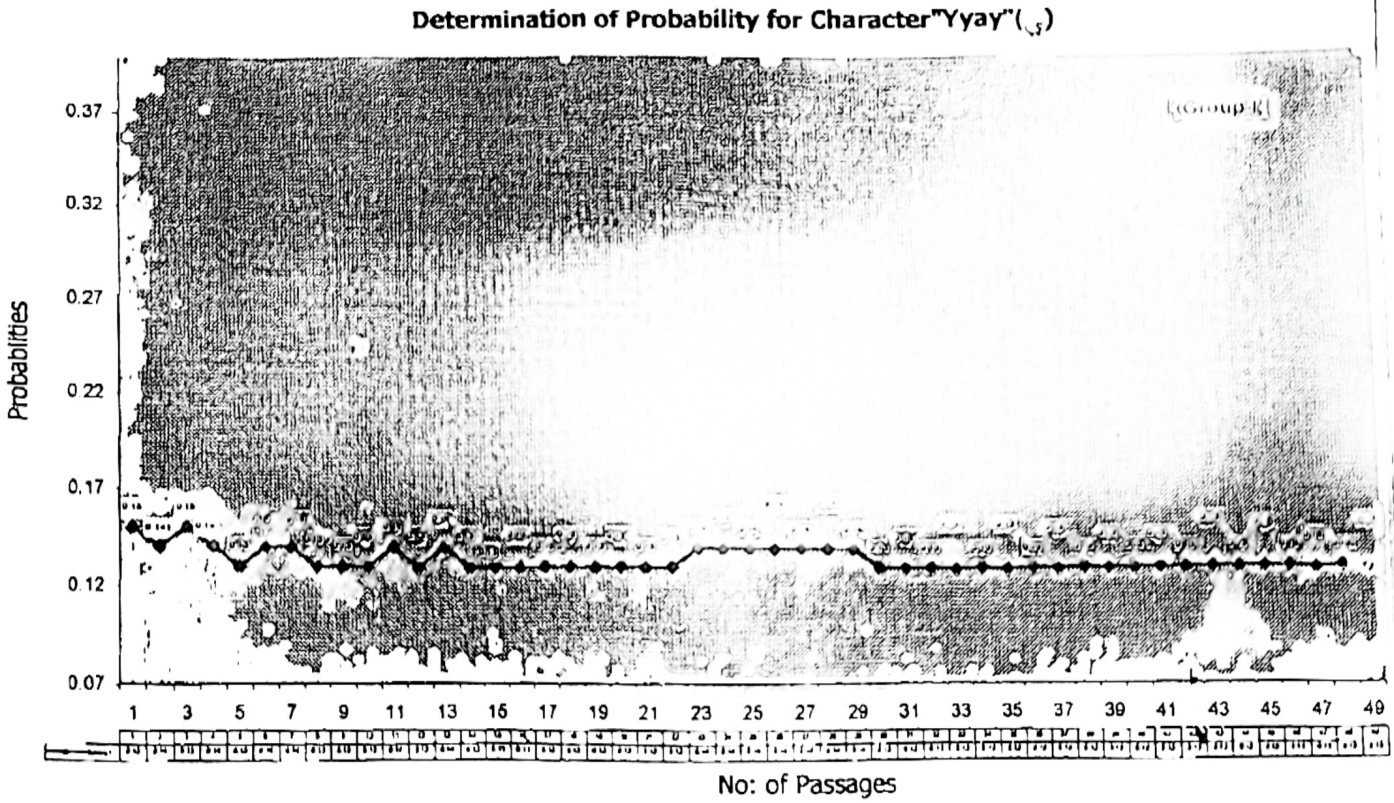

Figure 1:- Illustrates the determination of the stability of Av(prob) for the character "Yyay"(ي)

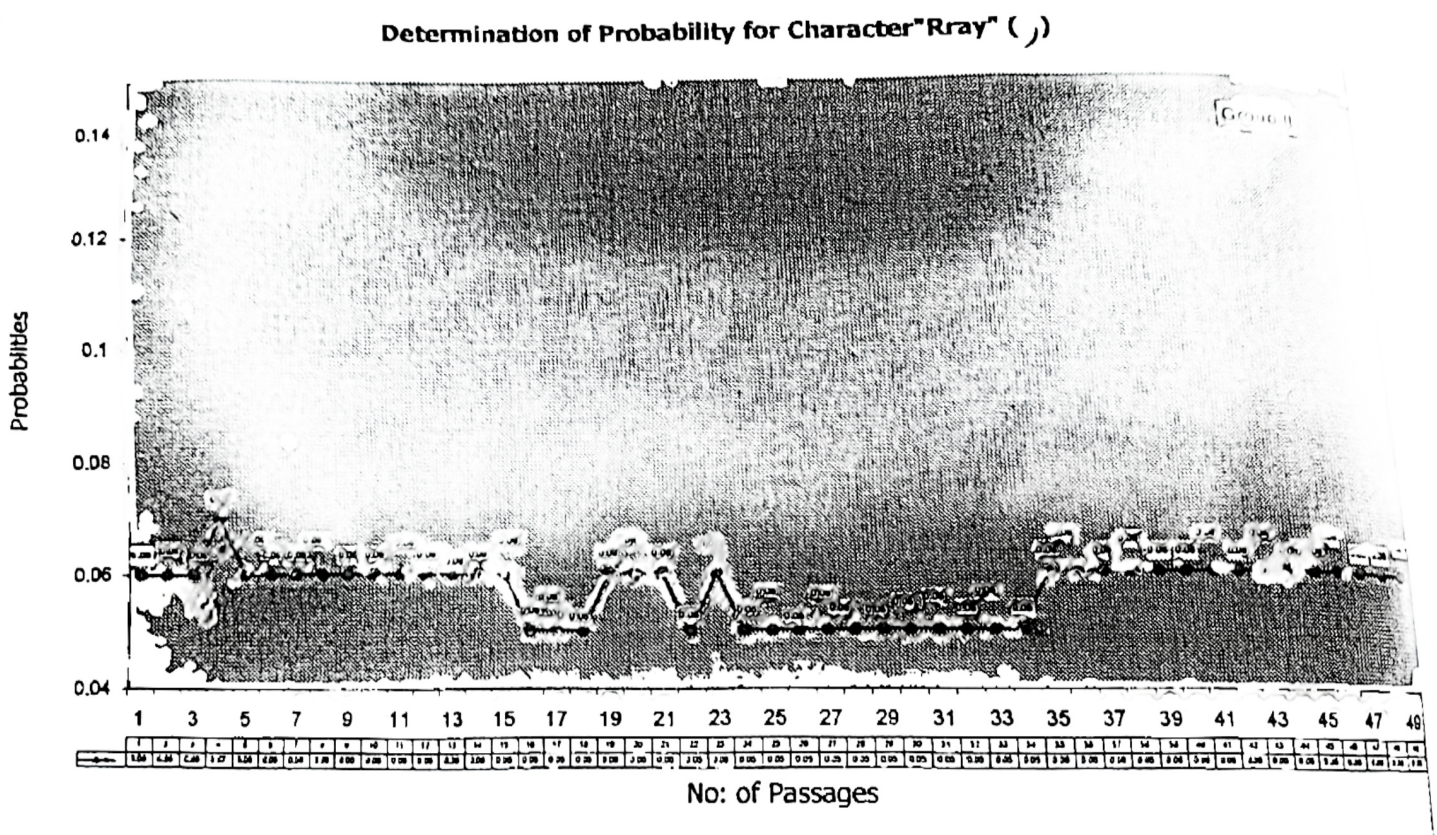

Figure 2:- Illustrates the determination of the stability of Av(prob) of the character "Rray" (ر).

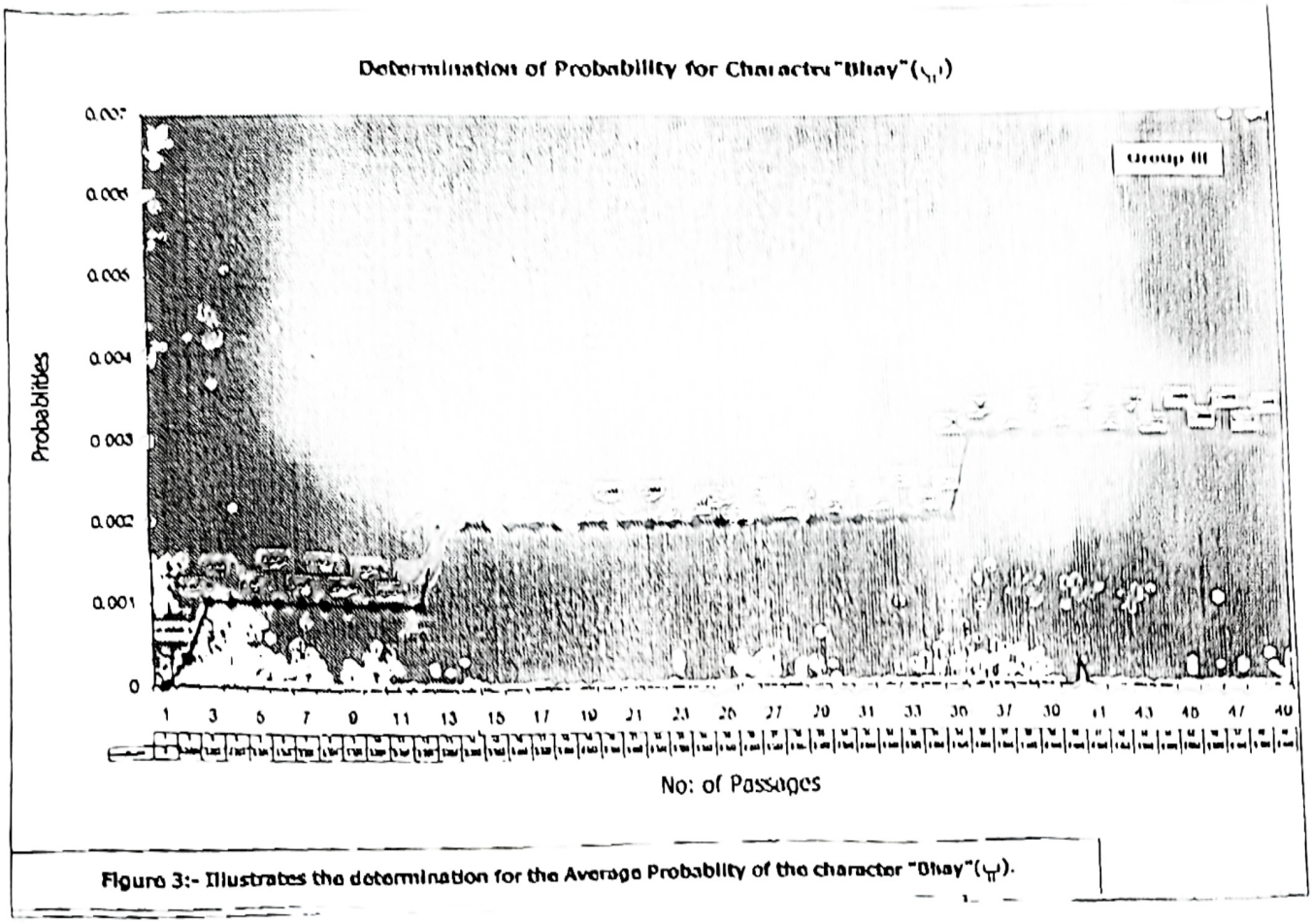

Figure 3:- Illustrates the determination for the Average Probability of the character "Bhay"(γ_{ij}).

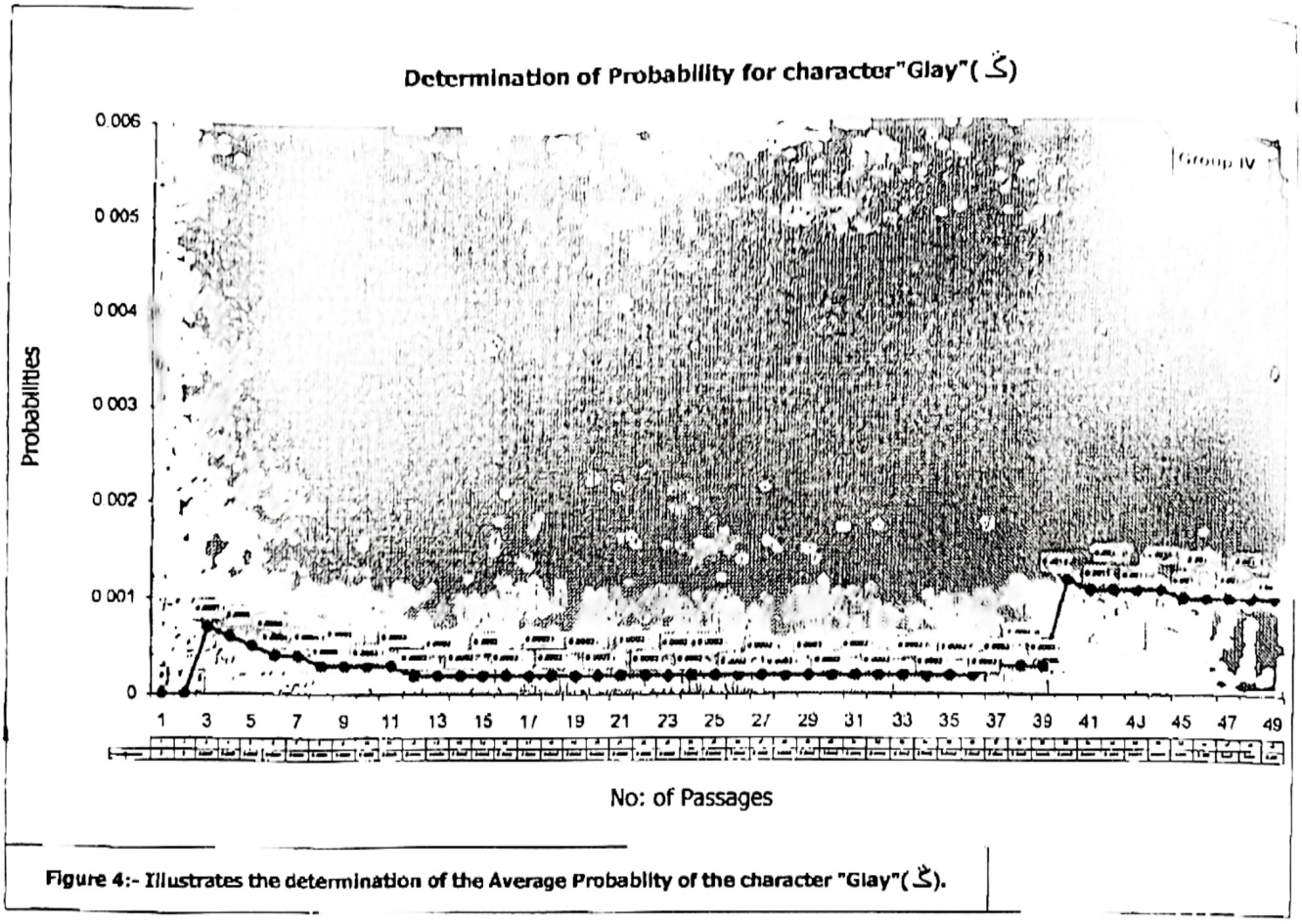

Figure 4:- Illustrates the determination of the Average Probability of the character "Glay"(ڳ).

References

Abramson, N. (1963) Information Theory and Coding. McGraw-Hill, New York.

A. Shah, R. Shaikh, (2001) Statistical Measures of Sindhi Language Alphabet For Text Compression. Expo Asia (2001) Hyderabad.

A. Karim, M., B. Memon A. Bahio W. M. Ansari; A. Kadri; G. Q. Somro; U. Bukhari; M. I. Shaikh; R. A. Memon; (2001) Sindhi Text Book (Sindhi Pehriun kitab) (2000) Sindh Text Book Board. Jamshoro.

http://www.Data-Compression.com

Nelson, M. J. and L. Gailly, (1995) Data Compression Book. Second Edition. M and T Books, USA.