ABSTRACT

This thesis investigates various issues related to Multilingual Data Representation in terms of efficiency and storage, transfer through the net, and processing.

In the early eighties, the Dept. of Electronics of the Govt. of India set up an expert committee to set up standards for information processing of Indic languages. The Indian Script Code for Information Interchange (ISCII) first launched in 1984 is the outcome of this exercise. This code, namely, ISCII is an 8-bit umbrella standard, defined in such a way that all Indian languages can be treated using one single character encoding scheme. ISCII is a bilingual character encoding (no glyphs!) scheme. Roman characters and punctuation marks as defined in the standard lower-ASCII take up the first half of the character set (first 128 slots). Characters for Indic languages are allocated in the upper slots (128-255). The Indian Standard ISCII-84 was subsequently revised in 1991 & 1997 (ISCII-91,ISCII-97). The research and development wing of the DOE, Govt. of India (called Center for Development of Advanced Computing (CDAC) based in Pune, India) has developed software packages based on these Indian standards.

Unicode is a rapidly evolving international standard for multi-lingual word-processing. Unicode is a more ambitious 16-bit character encoding scheme with defining of over 65000 slots for 50+ world languages. Along with other Indic languages, Tamil has been assigned specific slots U+0B80 -> U+0BFF (which, in decimal, is 2944 -> 3071; 128 locations) in this multilingual standard. For obvious reasons, the choice of characters in UNICODE for Indic languages is based on ISCII. Microsoft has already implemented Unicode in its Windows 95/NT OS and even distributes a Unicode font free for multilingual word processing.

This thesis compares different character encoding schemes used to encode the characters in different languages. A new character encoding protocol called Protocol for ApplicatioNs Development In THAmizh and Multilingual Computing (PANDITHAM) has been developed to deal with the characters in different languages. The languages English and Tamil are taken for a case study and its performance under networking environment is compared with regard to PANDITHAM, Unicode and UTF-8 encodings. This study has proved that PANDITHAM is optimal for all other languages as it reduces the network congestion.

New techniques have also been evolved so as to encrypt and/or decrypt Tamil based Text using a 16-bit key. This scheme is called Crypto Index Scheme. Tamil based text does not have digrams and trigrams compared to English based text. Moreover, Tamil language has a large character set (247+66). Hence it is all the more difficult for brute force attackers to decrypt the Tamil based ciphered text.

To investigate the performance of Unicode, UTF-8, and PANDITHAM, NS-2 based simulations have been carried out by sending a text file from source to destination using different topologies, Queue sizes, and Bandwidth etc. For each TCP agent, a new FTP application has been defined. A text file containing N lakh Tamil & English characters are transmitted using PANDITHAM, Unicode, UTF-8 encoding with different Topologies, Bandwidth and Queue size. The results show that PANDITHAM has demanded less number of packets for Tamil characters compared to Unicode and UTF-8 for the same information. But, for English characters PANDITHAM and UTF-8 have required less but same number of packets compared to Unicode. This clearly shows that the network congestion and the number of retransmissions get reduced. Moreover, it is demonstrated that PANDITHAM based text needs less time to render Tamil characters in contrast to UTF-8 to Unicode based transformation .

From the traffic analysis done, it is found that the rapidly accelerating trend of globalization of businesses and the success of e-Governance solutions require data to be stored and manipulated in many different natural / local languages. When the data is encoded using PANDITHAM as character encoding scheme for different languages more information can be transmitted in less amount of time. The network congestion can be significantly reduced by reducing the number of retransmissions. It can be used for all other languages as well where the characters in the language have to be stored in a PANDITHAM table, and the language codes have to be standardized world wide.

This thesis concludes that PANDITHAM scheme sits on par with Unicode and is a viable alternative to Unicode, as it is character oriented, consumes optimal space for storage and hence more throughput in terms of number of information that can be transferred (per unit time) through the network.