

ices and content to distant areas

>> Online Entertainment sources like Digital TV, Video on demand etc.

Approach: What's required?

Collaboration is the key

The need of the hour is to focus on comprehensive efforts across building awareness, education on the benefits of Internet adoption, enabling affordable access through affordable devices and accelerating deployment and roll out of wireless to connect the last mile. Given India's expanse, and lack of copper infrastructure, the most cost effective and yet the fastest way, is to drive the adoption of wireless broadband infrastructure.

We thus need a co-coordinated and concentrated attempt to accelerate broadband Internet access through the availability of a slew of affordable devices and with Government policy and industry efforts to build a sustainable wireless broadband infrastructure. As per research provided by IMRB, PCs and notebook prices have progressively fallen by 7-10% every year for the last ten years and becoming affordable by the day.

There are new processor being invented and offered through Internet-centric netbooks and nettops for the masses. The focus thus shifts entirely to enabling Wireless broadband for mass market deployment in India and south Asia, as it offers the only reasonable alternative for a price conscious popula-

tion.

SO Can a Billion Indians get connected?

There are tens of efforts that could be seen in the market to increase the mobile reach, geometric progression in connectivity. Thus, recently, even we have also got into the grind: Intel along with its over twenty partners has initiated Internet awareness across 37 towns through the Netyatra , a bus equipped with Internet connectivity, latest category of Internet centric devices - netbooks and nettops, & demos to communicate the benefits of Internet to improve lives. 75,000 citizens, school children, college students have visited the Netyatra to learn and experience the power of the Internet.

It is overwhelming to observe that how a little effort can prove the point that if the services and infrastructure could be initiated then citizens would positively respond. Across the country, over 200,00 Indians have pledged their support by joining the "Connected India" movement through the website - www.connectedindians.com. The onus is now on us and partners to make each and every Indians to get connected. I urge you all to join the connected Indian movement at www.connectedindians.com and pledge to put the power of the Internet in the hands of every Indian.

R. Sivakumar is Managing Director, Sales & Marketing Group, Intel South Asia. He can be contacted at ramamurthy.sivakumar@intel.com

Digital Content in Local Languages: Technology Challenges

VASUDEVA VARMA

World Wide Web holds the key to shape our societies and cultures in the future. The cultures and languages who dominate the Internet presence will have more chances of survival than those with less Internet penetration.

It is a well known fact that information on the Internet is skewed and all languages are not represented equally well on the Internet. English appeared to be the universal language of the Internet for sometime but now there are other languages establishing themselves. Some of the dominant languages of Internet are English, Mandarin (Chinese), Japanese, Korean, German, French and Spanish. In fact, recently Mandarin language has overtaken English as the most popular language on the Internet. There are reports that Chinese government is making a conscious effort to make Mandarin as a number one language in the world and promoting the Mandarin language teaching globally. The success of this campaign is evident from the news that starting from mid March, Vatican will be making its website available in

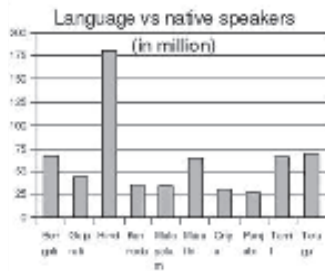
Chinese. From a simple Internet search, one can easily figure out that there are a large number of tools available for learning, analyzing and processing Chinese languages. My intention here is to emphasize the need to do similar things for Indian languages.

Indian languages are among most widely spoken in the world in terms of population. India is a multi language, multi script country with 22 official languages and 11 written script forms. About a billion people in India use these languages as their first language. About 30% of the Indian population speaks Hindi but it is concentrated only in the northern and central region. There are many areas within India where Hindi is not known.

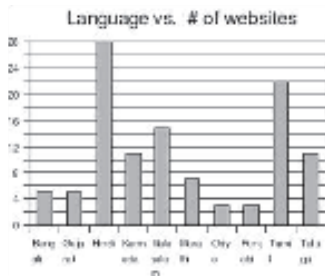
In fact, nine out of top 30 most widely spoken languages in the world are Indian languages with Hindi ranked at number 3 after Mandarin and English. Other languages are Bengali, Urdu, Punjabi, Tamil, Telugu, Marathi, Gujarati, and Kannada in that order. Please note that Bengali, Urdu,

Punjabi and Tamil are also spoken in neighborhood countries.

The following chart gives the number of native speakers in popular Indian languages in India:



According to the study conducted by search and information extraction lab of Language Technologies Research center at IIIT Hyderabad in 2006 (published in WWW conference in 2006), the Internet penetration of the Indian languages is very low. That means the percentage of Indian language content is very less compared to the official languages of United Nations. The following chart gives an idea about the number of web servers available in various Indian languages as of 2006.



Creating content in local languages is a challenge particularly because of the sophistication of Indian language

scripts. There was a huge delay coming out with a standard for Indian languages that was accepted by everyone. Unicode standard for Indian languages was relatively delayed and even when it arrived the publishers of Indian language content were slow to adapt it because the support from operating systems and browsers in rendering Indic scripts was also delayed. For example, Hindi is rendered properly only on Windows XP and beyond. Though there are some efforts to create Indian versions of the Linux, largely there is very little support for Indian languages.

Lack of tools and other support for Indian language scripts creates a major bottleneck for web publishers in terms of creating content. It resulted in most of the content to be in proprietary encodings. Every publishing house had its own set of encodings. For rendering view point, using multiple encodings while authoring such websites served the purpose but this move resulted in having about 95% of the Indian language content to be in non-standard, proprietary fonts. This affects the viewership because most of the available content is not searchable and hence not reachable.

The technology challenges for better Internet penetration of Indian languages can be clas-

Keyboards should be designed in such a way that they stick to the standard layout of a Roman keyboard

sified into three major areas: better input methods for Indian language content creation, better rendering and display of the content and better search technologies for reaching the right content which require good amount of language processing. In this article I will mostly focus on the first set of challenges - namely the Indian language text input methods.

Text input is a major part of user interaction between the man and the machine. Text entry in Indian languages is an interesting technical problem owing to the nature of the Indian language scripts compared to English. Apart from having a larger number of alphabets compared to English, the consonant-vowel combinations generate new symbols in Indian languages, which make it impossible to use the English keyboard as it is, for data input in Indian languages.

One way to enable input using the same QWERTY keyboard is to create a mapping between English and the target Indian language and use this mapping while typing is done by the user. These applications are usually

referred to as Input method editors. Another possible method is to design specialized keyboards for Indian languages, which will have a number of keys and facilitate typing all possible combinations easily. However, keyboards should be designed in such a way that they stick to the standard layout of a Roman keyboard. Other possible options are to use a soft keyboard or an auto-complete like prediction system.

The problem of text input for Indian languages in to five categories.

1. Keyboard layouts: These refer to all QWERTY like keyboards designed both in English as well as Non-English languages. Using specially designed keyboards, we should be able to enter the Indian language text and switch back and forth to English when necessary.

2. Input method editors: These refer to all the systems which use a mapping between two scripts (typically Roman and a target language) to facilitate typing in one language, using the script of another.

3. Prediction systems: These refer to the input systems which attempt to predict the user needs by means of various statistical and non-statistical techniques and generate suggestions for the user as he/she types in using a stan-

Text input is a major part of user interaction between the man and the machine. Text entry in Indian languages is an interesting technical problem owing to the nature of the Indian language scripts compared to English

dard keyboard. The options like T9, Google Suggest are some of the examples of text prediction systems. The aim of a text prediction is to reduce the typing load on the users' part. Prediction becomes an important data input method especially in case of Indian languages owing to the possibility of using more than one key-stroke per one character of Indian script.

4. Soft keyboards: Soft keyboards refer to the input sys-

Our goal should be to break the language access barrier of the users in the digital world who can speak only an Indian language

tems in which data is entered using an alternative input device like mouse or stylus. Virtual keyboards are very useful for security related applications such as bank account access or for mobile device or touch screen based applications where the full keyboard facility is not available.

5. Input through Speech

Recognition: This is the ultimate goal of speech research. If we can interact with computers and provide our input in the form of spoken language, without using keyboard, mouse or any other artificial devices, that would make life so much easier. However, it is a huge challenge to come up with a speaker independent, domain independent speech recognition system. We can be hope-

ful to see such systems in next three to five years.

Besides text input methods, the challenges of processing Indian language content and ability to render them well are also important. Rendering problem can be expected to be solved automatically with each new release of operating systems and word processing tools. We need to pay attention to develop tools and technologies to address Indian language processing. This includes development of speech recognition and synthesis systems, spell checkers, search engines for Indian languages, cross language information access systems, machine translation systems among Indian languages and English.

There are several research

and development efforts across India (and some outside India) looking into these challenges. Some of them are now being available for commercial use as well as in the form of free software in the open source. We are hoping that very soon we will be able to see the development of technology where Indian language content creation and access becomes easy and natural. Our goal should be to break the language access barrier of the users in the digital world who can speak only an Indian language.

Prof. Vasudeva Varma is a faculty member at IIIT Hyderabad. He is heading search and Information Extraction Lab and Software Engineering Research Lab at IIIT Hyderabad. A part of the work reported in this column is based on research by him and his students VB Sowmya and Prasad Pingali.



Yes! Please send me one year at the special new subscription rate of Rs 500

NAME: _____

INSTITUTION: _____

ADDRESS: _____

STATE: _____

COUNTRY: _____

I WISH TO PAY BY: CASH CHEQUE/DEMAND DRAFT CHEQUE/DD NO DATED: _____

Send this form to Ravi Kanta, Digital Empowerment Foundation, 3rd floor, 44 Kaalu Sarai, New Delhi - 110 016. +91-11-26532786