Developmental Informatics at IIT Bombay

Anil Bahuman[1, 2], Chaitra Bahuman[1], Malati Baru[1],
Subhasri Duttagupta[1], Krithi Ramamritham[1]

[1]IIT Bombay, [2]Agrocom, India

ABSTRACT

IIT Bombay's Developmental Informatics Lab is a cross disciplinary group consisting of 6 faculty, 30 research staff and several students. The lab is working towards increasing access to information -- through the use of internet and communication technologies -- to communities in the developing world especially rural and small town India. The lab is supported by Indian Government funding sources as well as corporate and multi-lateral agencies to solve technical problems in local communities in sustainable ways. This paper focuses on two mature projects of the lab – one caters to Indian farmers while another helps with the education of tribal populations.

1 INTRODUCTION

70% of the world's poor have little access to information; over 1/3 of the world's population has never made a telephone call. The developed world has 50% telephone lines per 100 people. In developing countries the ratio is 1-4 telephone lines per 100 people. The gap is even higher when it comes to internet access. Most of the information exchanged over the Internet is in English – a language spoken by just 10% of the world's population [11].

Developmental Informatics is the study of (a) how access to internet and communication technologies (ICTs) can be increased and (b) how ICTs can help speed up the socio-economic development of these populations.

While several aspects of this large problem are being addressed by various national and international initiatives to increase access to ICTs, the research community active in the area is still very small (some top technical conferences are beginning to add "Developmental Tracks"). The Developmental Informatics Lab at IIT Bombay provides a platform for technical aspects of this broad research area that cuts across several disciplines. Areas under investigation include a) the design and evaluation of web and mobile applications for resource constrained environments, (b) visual and product design, (c) cross-lingual information retrieval and translation, (d) improving information dissemination protocols over the internet catering for low bandwidth and small devices, (e) ethnographic studies emphasizing the study of social & cultural factors influencing interaction design of applications for e-learning and use of computers in education and (f) involving rural communities and planning strategies for scaling up those innovations that have demonstrated rural user demand through government or private sector participation. The laboratory is formed around several core projects broadly in the areas of agriculture and education, each involving academic, industrial, government and village community partners.

Researchers new to the area sometimes question the application of latest technologies among such populations. A few even promote the use of older, obsolete (cheaper) technologies for such populations. Our experience especially with private sector involved in providing products and services to these populations is that if we are to design and develop technologies with the needs of these users in mind, several cutting-edge technologies, suitably designed and deployed, can provide highly affordable solutions to age-old problems. An often quoted non-ICT example is the setting up of community owned wind-driven or water driven generators with excess electricity generated being sold back to the grid. Another ICT example that the lab was involved with are community owned networks set up using wireless technologies [1].

We now discuss two projects at the lab, incorporating technical innovations in several areas.

ICT Training of Teachers of Tribal Children was undertaken after detailed ethnographic studies of the camps and schools were made [2]. A community visit oriented design phase resulted in the creation of several artifacts incorporating the use of local video footage, colorful culture-sensitive icons and teaching materials. Figure 1 shows a student attempting a sorting task. These artifacts are now being made part of the curriculum at special schools.

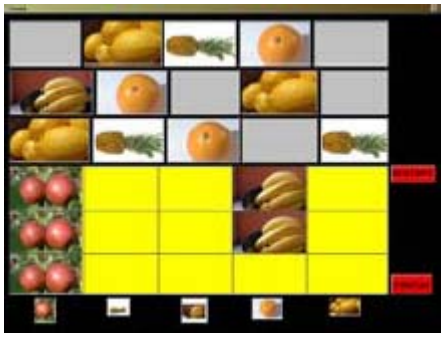

Figure 1: Colorful icons of fruits that are to be dragged using the mouse and sorted by type

The rest of the paper discusses the second project at the lab – dealing with increasing the reach of agriculture related information – the project being a good example of the confluence of several core areas of the lab's research.

Agriculture Extension and Farmer-Outreach Programs face three major challenges – cost-effective outreach in a large country with diverse needs, customizing information requirements of millions of individual farmers and providing excellent customer services to people who can pay very little. As has been well recognized, internet and mobile networks are now reducing the costs of creating and sharing information rendering it (i) affordable, (ii) relevant (timely and customized), (iii) searchable and (iv) up to date. The potential for such content and technologies is huge. Large sections of the farming community, particularly the rural users, are yet to gain access to the huge knowledge base acquired by agricultural universities, extension-centers and agri-businesses. While telecenters as well as several private sector initiatives are beginning to dot the Indian rural landscape, one of the major barriers is the lack of agro-content that (i) is in the language of the farmers (ii) is relevant to their needs and (iii) is delivered in a

Figure 2: Pictures are worth a thousand words for experts answering questions (text) on aAQUA

form that is of immediate use to them (iv) searchable, so as to have a multiplier effect.

With the needs of Indian farmers in mind, the Developmental Informatics Lab, at IIT Bombay has developed and deployed innovative ICT tools for dissemination of agricultural information over the internet This paper discusses three database-backended tools - aAQUA, Bhav Puchiye and a digital library of Agricultural Documents that are all accessible at the www.aaqua.org portal [3].

aAQUA - which stands for almost All QUestions Answered- is a farmer-expert Q&A database supporting Indian languages (Figure 2).

Bhav Puchiye (meaning, "ask for the price", in Hindi) is a web-based application for viewing the price and price-history information of agrarian products at the nearby wholesale markets (called mandis).

The Crops Library project was initiated to build a repository of agricultural documents. The content has been provided by Krishi Vigyan Kendra, Baramati, who have collected recommendations from agricultural universities, residential experts, journal articles, disease forecast reports etc.

These research artifacts incorporate innovations in database query optimization and caching, cross-lingual multimedia information storage and retrieval, improved usability and interaction and new information dissemination algorithms over unreliable networks and in resource constrained environments.

2 USAGE PERSPECTIVE

2.1 aAQUA

aAqua [3] provides a communication framework for rural users to get their problems solved. Solutions are shared by a larger community and are provided by either person(s) having similar experiences or by agricultural expert(s). aAQUA started as an online discussion forum and was modified over 3 years through interactions of a number of computer operator mediated internet kiosks that are being deployed in India by both commercial and government funded ventures. The computer operators earn a livelihood for themselves and also help make computers usable by a large number of users who are new to computers and provide IT enabled services even to the illiterate.

In this section, we present various stages a typical aAqua question passes through in its life cycle. An aAqua question is posted either by a registered user directly or through a telecenter/kiosk operator who has an account in aAqua. Usually the question is from a farmer whose profile information provides details such as crop, farm size, pesticides and fertilizers used, dosage etc. This provides an appropriate context for the question. If the question is clear and complete, the Agri-experts provide an answer to the question. Otherwise they ask the original user for further details about the problem. A local help desk operator follows up with the farmer over a phone, a week
after the answer has been given, to check if the answer solved the farmer's problem. In case the farmer found the solution useful he documents the impact. In case the solution was found to be ineffective the expert follows up on the phone and captures the limitations of the solution.

From a user perspective, the success of aAqua, the Crops Library and related tools lie in its incorporation of effective solutions that:

1. Provide an interface suitable for novice internet users.
2. Allow users to browse content by forum name as well as agricultural keywords and key phrases presented automatically to the user
3. Allow users access to content even over intermittent or low bandwidth connectivity.
4. Provide solutions to a user in his/her own language using translation technologies to assist manual translation.
5. Enable experts to locate answers in the continuously growing knowledge base of aAqua for recurring questions.

2.2 Bhav Puchiye

Bhav Puchiye incorporates innovations from the perspective of interface design and data provisioning. An iconic interface is presented to the user (Figure 3) with a choice of commodities, the available varieties (from which the user chooses the commodity and variety of interest) and a calendar (from which the user selects the date of interest). The prices are displayed spatially over a map. The user can decide where to sell his produce to get the maximum profit, depending on the prices and the distance of the markets. The user can create a login, store profiles of commodities and locations and create and receive e-mail alerts when prices of certain commodities of interest change in the markets of interest.

Figure 3. Bhav Puchiye: Market prices are shown on a map of Maharashtra state, commodities, varieties and dates can be chosen on the panel on the left.

2.3 Crops Library

Crops Library consists of collections of crop diseases (Criop Doctor), crop recommendations and translated aAQUA threads. The collections are built centrally using open source software called Greenstone chosen after comparing 3 open source digital library software [6] – DSpace, EPrints, and Greenstone.

A user of Crop Doctor browses through a gallery of images classified by Crop and Disease names. He/she can choose a matching image and read the symptoms, causes, prevention and control of the disease.

A user of Crop recommendations chooses a crop of interest classified as vegetables, fruits, cereals and pulses, flowers and "Others" and reads through tested recommendations collected from agricultural universities.

A user of aAQUA Translations collection can choose to view the questions and answers in English, Hindi or Marathi. The question-answers are also classified by crop names.

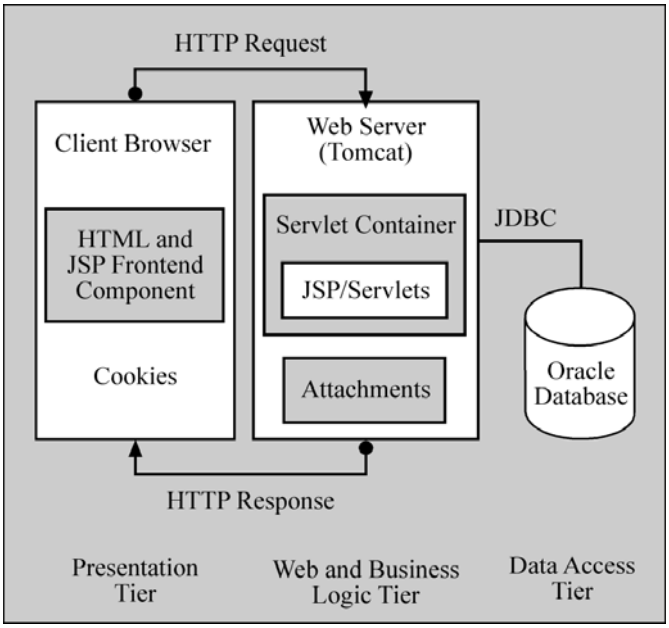

Figure 4. Three Tier Web Architecture of the aAQUA system

3 TECHNICAL PERSPECTIVE

Figure 4 explains pictorially how aAQUA and Bhav Puchiye [5] employ the three tier web architecture using Java technology (Java Server Pages/Servlets) and Oracle (aAQUA) or MySQL (Bhav Puchiye) databases.

Based on the standard MVC (Model View Controller) architecture, they are compatible with any Servlet container which supports JSP 1.2 and Servlet 2.3. They are currently being deployed within the Servlet container (Catalina) of the Tomcat Web Server. The systems use Unicode UTF-8 compliant databases. The aAQUA database comprises mainly of tables that include attributes for Member, Farmer, Expert, Moderator, Category, Forum, Posts, Thread, Permissions and Attachments. The Bhav Puchiye database comprises of tables that include attributes for Member, APMC (Agricultural Produce Marketing Committee) user, Commodity names, Categories, Variety names, Market names, Arrivals and Prices.

The Crop library uses Dublin Core standard of metadata storage [6]. Documents are uploaded to the server and indexed by meta-data provided by a librarian. Alternatively, meta-data may also be extracted from the documents.

3.1 Enhancing Usability

The target users are predominantly non-English speakers, semi-literate and new to the internet. Our tools provide a simple, yet rich interface suitable for new internet users. A web-based soft keyboard is also available to assist users. The questions on aAQUA are answered by agricultural experts who provide answers in the local language (or a combination of languages), paying special attention to the language used. Use of technical jargon is generally avoided, for e.g., instead of prescribing quantities measured in *parts per million (ppm)* or *grams*, common measures such as the *teaspoon* are made.

Some agricultural and veterinary problems are better described by photographs or audio and video files which provide details to the expert. aAQUA allows attaching images taken by a digital camera or scanner and experts can zoom to specific portions of the images. As shown in Figure 2, users can add images and also attach audio and video files which are automatically played back on viewing.

There are many ways of organizing aAQUA content and it remains an open research issue especially when catering to different sets of users- farmers, experts, agri-businesses - both from literacy as well as expertise point of view. The two approaches in use so far are organizing (a) by question category – crops, animals, market price and others and (b) by agricultural keywords (in categories) chosen by experts.

The interface design of Bhav Puchiye employs the "Inverted Pyramid Approach"[5], which aims at maximizing the relevant results with minimal inputs from the users. The principle involves providing some results, by assuming default options, as soon as the web page loads. The user can change the options (usually provided on the same page) to view other results. Every mouse click in Bhav Puchiye refines the results.

The Crop Library interface is also designed using similar principles. In addition, the retrieval of documents is near instantaneous since the documents are converted into XML after uploading and displayed as HTML within the browser at runtime. This also eliminates the need for installing proprietary editor and viewer applications.

3.2 Offline Access, Database Synchronization and Caching

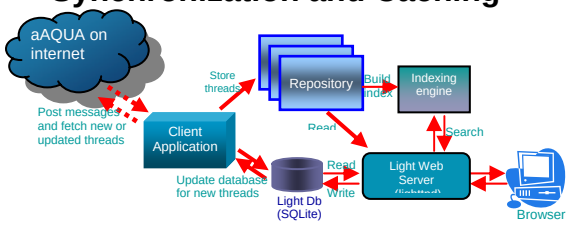

Figure 5. Architecture of the offline-aAQUA system

The underlying assumption in almost all web based applications is the availability of a continuous connection to the server over the Internet. However, aAQUA has to cater to users who connect to the internet on unreliable dial-up connections where the bandwidth is low or intermittent and the user is usually charged by the number of hours online. aAQUA pages are thus designed to have a lower payload and can also be installed as a standalone application which connects to the internet whenever available. The offline version of aAQUA is created for such users and can be personalized based on individual or group profiles and access patterns. It incorporates a store-and-forward mode, delaying authentication and allowing users to login and ask questions. The application periodically updates the main server and also receives the latest updates. The offline version can also be searched using the keyword search interface described later without connecting to the network. The Crop Library can also be configured for offline browsing and offline search and can be distributed in Compact Discs.

In order to improve the response time and robustness in delivery of content, we also have deployed aAQUA mirrors closer to the users. These mirror sites periodically synchronize with the main aAQUA server with the help of our synchronization tools for databases of different vendors (Oracle and MySQL). Query logs are stored at either end and conflicts are resolved at the time of synchronization. The synchronization tool is designed to be tolerant to network and server failures during synchronization so that over time the synchronization process will complete.

Web caching, database query-optimization strategies and Web 2.0 techniques are also being exploited to make our application more responsive. Frequently accessed fragments of pages as well as query responses are cached in the (i) browser (ii) one or more application server tiers. Our work in [10] studies the performance of our offline user activities comparing some caching approaches and popular approaches – online use, RSS feeds and Newsgroups.

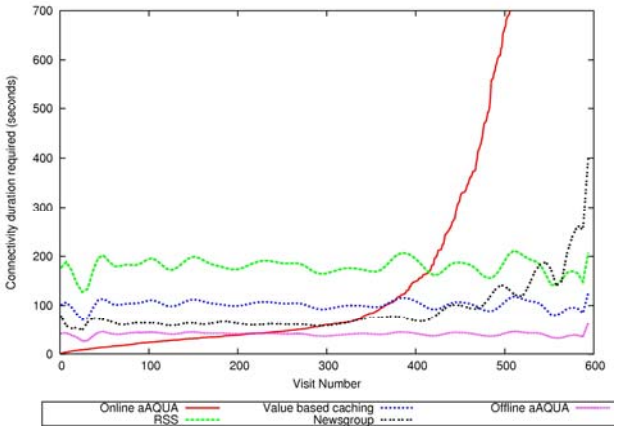

Figure 6. Comparison of some approaches that enable offline access

3.3 Language Independent Semantics-driven Retrieval

We are converting the knowledge contained in the Questions and Answers in the forum to a language independent form – Universal Networking Language(UNL) - so that Semantic Based Search techniques can be used for multi-lingual retrieval. This performs better than keyword-based approaches [4]. When a user types a query, the system converts it to a UNL query graph and looks up the UNL document base to find matches. It responds with the answers (also in UNL) and translates them into the language chosen by the user in the AgroExplorer [4] tool.

In addition to enhance multi-lingual retrieval, we are currently integrating functionality to (a) automatically pick agricultural keywords and (b) perform a multilingual keyword based search on the database. This is especially useful for users

fluent in two or more languages (which is quite common in India). The original search query is expanded with their counterparts in each language. It allows users to search in their own language and retrieve content in other languages.

Since aAQUA serves many novice users, a popular search feature is a categorized list of hyperlinked keywords (also called Virtual fora) that appears on the home page. On clicking these hyperlinks, aAQUA performs a search on the corresponding keyword (e.g., "Onion" invokes all onion related threads – independent of language).

Ongoing efforts address challenges in organizing and harvesting meta-data from aAQUA's heterogeneous and multimedia content. Separate resources like agri-ontologies and corpora for cleaning, organizing and identifying good metadata are being developed.

3.4 Reuse of knowledge in previous answers

Unlike most online discussion fora and web portals, a review of our search query logs has indicated that our users infrequently use the Search feature. The "Browse by Agricultural keywords" was a feature motivated by the same observation. When a question is asked, the expert uses this tool to find previous answers that can be reused. e.g., If the expert is searching for *Powdery mildew* disease on *tomato* he could look under *Crops* and refine his search to *Powdery mildew* (Figure 7).

Figure 7: Refining search by crop and disease name

The agricultural keywords are in the noun form and are tagged to enable classification. Retrieval of documents by these keywords is being improved by incorporating stemmers and spell-checkers (measured by "precision" and "recall"). The limitation is that the precision may be compromised. E.g., If "powdery" is stemmed to "powder" the number of hits increase several times.

A feature has also been provided for experts to help them give quick answers while they formulate more detailed answers. The quick answer contains a hyperlink of keywords chosen by the expert allowing the user to click and browse through the results while he is waiting for the answer. Although technically possible, we have avoided giving automated answers. There are a number of reasons for this. A review of the question-answer database has shown that while farmers' questions are imprecise (short, sometimes erroneous, wrong grammar, spelling, local language typed in English alphabet, question captured as scanned text, as pictures, audio or video), they expect precise, descriptive answers. Another issue is the temporal validity of historical answers in the database. Yet another challenge is the number of variables that makes each farmer's question unique even though they may be expressed in similar ways on aAQUA. Thus, techniques those assist experts in reviewing past Q&A while answering new questions are more useful and are being incorporated.

3.5 Integrating with email and mobile phones

aAQUA has also been tailored to cater to users with limited access to the internet in several ways. Users can post their questions over email as well as mobile phone text (SMS) and receive their answers back. We are also looking at the possibilities of providing rich multimedia applications over the mobile phone are currently being investigated. One example is using a combination of photos taken over a mobile phone and GPS information to spatially record images of crop infestation or reported issues on a GIS. Another example is sending alerts such as weather forecasts in real-time to a large number of farmers.

3.6 Integration with Related Content Repositories

aAQUA has been integrated with other sources of information including government schemes for farmers, monthly forecasts and cropping suggestions, market price information (Bhav Puchiye), crop diagnostics (crops library), crop recommendations (crops library) and a glossary of agricultural terms and keywords.

We track usage history to decide which

repositories need frequent updates. Bhav Puchiye's market price information is updated daily. aAQUA content is indexed immediately after posts are made.

The quality of aAQUA is measured in the aAQUA-QoS module. The time taken to answer the question is tracked for every question and the minimum, average and maximum time for every month is computed. The reports allow us to track usefulness and response time of answers by farmer, kiosk and expert. Questions were received from 290 of India's 604 districts.

4 FUTURE

India in 2006 emerged as one of the world's fastest-growing wireless telecom markets, with the number of mobile-phone service subscribers in the nation growing to 149.5 million, up from 85 million in 2005. Indian farmers are beginning to go online for the first time, not on a PC but on a mobile phone [7].

The Developmental Informatics lab continues to find socially-relevant applications of technical research undertaken by faculty at IIT Bombay on ICT. Examples are *Siksha* – exploring ICTs for Education and *Galla* – exploring ICTs for Small & Medium sized Retail Enterprises (SME clusters). Another example is *Akashdoot* – Deployment of climate sensors for recording weather data and developing crop disease prediction algorithms supported by Agrocom Private Ltd, a spin-off company of our lab based in IIT Bombay's SINE [7] incubation facility.

Our software engineering team is evolving a design & implementation methodology to develop small footprint applications that are deployed in a device-independent manner. An example is developing applications for localized weather or market prices that can be integrated with larger PC-based web applications and usable over handhelds as well. Extending the same application to handhelds where memory is constrained involves a trade-off between user functionality and delays perceived by the user.

Our community oriented projects which use some Web 2.0 principles [8] are inspiring the software teams both at Agrocom and the Developmental Informatics Lab to further explore Web 2.0 principles such as (a) lightweight programming languages (b) incremental daily releases (c) content syndication and loosely coupled, cooperating data services, (d) data and code located closer to user using AJAX-like approaches and (e) feeds, user data and community blogs to sustain applications that get richer with use and age.

Community-oriented projects provide a wealth of opportunities to do cutting-edge research. For example, porting aAQUA on handheld devices led to DeLite, a small footprint database [9] that employs a novel storage scheme. The scheme is selected based on data characteristics, nature of queries, and updates. Also, DeLite's query execution plan is chosen depending on the amount of available memory and the underlying storage scheme.

Another example is contextualizing content to the user's location, profile, usage history and current date and time. Other areas of interest are effective cache management of mobile devices and prediction of contextual information to be presented to the user just before the occurrence of an event. Increasingly, such projects are finding corporate sector funding.

5 REFERENCES

[1] A Bahuman, S Inamdar, R Swami and K Ramamritham, "Robust Network for Rural Areas: study of two of Nlogue's ICT projects (in Maharashtra) and a compilation of the weakest links in their services" http://www.dil.iitb.ac.in/docs/Interim%20Report-Feb%202005-IIT-Bombay.pdf

[2] Dr. M V Ananthakrishnan, "Educating Nomadic Children: An experiment with the Convergence of Technologies" ,IEEE TENCON 2005, Melbourne, November 2005

[3] www.aaqua.org

[4] S Kagathara, M Deolalkar, P Bhattacharyya, "A Multi Stage Fall-back Search Strategy for Cross-Lingual Information Retrieval", Symposium on Indian Morphology, Phonology and Language Engineering, Kharagpur, Februrary 2005, http://agro.mlasia.iitb.ac.in

[5] K Ramamritham, S Duttagupta, A Joshi, G Mathur and T Vilankar, "An Interface-Driven Approach to Information Provision for Wired and Wireless Customers", Mobile Commerce Proc Int'l Workshop, pp. 102-109, 2003, www.dil.iitb.ac.in

[6] A.Bahuman, C. Rao, S. Nair, "Building Open Source Digital Libraries", National Seminar on empowering the masses through technology application and skill training, http://www.dil.iitb.ac.in/dil.htm

[7] www.sineiitb.org, Society for Innovation & Entrepreneurship. Also www.agrocom.co.in

[8] http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[9] Rajkumar Sen and Krithi Ramamritham, Efficient Data Management on Lightweight Computing Devices, Proc. of 21st IEEE International Conference on Data Engineering, Tokyo, Japan, April, 2005.

[10] Saurabh Sahni and Krithi Ramamritham, Delay Tolerant Applications for Low Bandwidth and Intermittently Connected Users: the aAQUA Experience (poster paper), WWW2007, Banff, Canada, May 2007.

[11] http://www.learningpartnership.org/resources/facts/technology