Providing Web Search Capability for Low-Connectivity Communities

Libby Levison, William Thies, Saman Amarasinghe
MIT Laboratory for Computer Science
libby@mit.edu, thies@mit.edu, saman@lcs.mit.edu

Abstract

There are many technical problems that exist in communities other than our own. These problems both deserve our attention and require focused research. We present one example: the TEK Search Engine. TEK is an email-based search engine designed to deliver low-bandwidth information to low-connectivity communities.

1. Introduction

In many of the world's communities, there is limited access to information. In places that have both computers and functioning phone lines, the Internet has the potential to provide access to a large amount of information electronically. However, there are obstacles. Bandwidth is so narrow that it can take the user hours to find what she is looking for when browsing the Web, since she has to wait for each page to load. Time spent online often translates to higher telephone and ISP charges. Unreliable network infrastructures can sometimes prevent access to the Internet altogether. There might be little access to technical support, software patches and anti-virus software.

At the same time, the developed world is moving to a more and more web-centric existence. Customer service is a website, not a phone call. We download school applications. We shop. Documents are exchanged. News is provided online. But the news site is image intensive, and loads very slowly over a 28K modem. For users with a low-bandwidth Internet connection, accessing these services might not be possible. High-bandwidth features that improve the appearance of a page risk to widen the digital divide even when they contain informational content.

This is an example of a set of problems that exists outside of many of our normal frames of reference. They are challenging, technical problems, and we think that problems like this deserve more of our attention. It is our responsibility to remember that not everyone is well-connected, and to work to provide access to the millions of people who have restricted or no access to the computer infrastructures – such as the Internet – that we take for granted, whether they are in the developed or the developing world.

In this paper, we will discuss the need to recognize the numerous technical challenges that exist in other settings, and how to find and identify them. Just as the field of engineering has the field of Appropriate Technology, we will discuss the need to recognize the field of Appropriate Information Technology (AIT): applications of Information Technology which are designed to suit the limits of their environment. As an illustration, we will describe the TEK Search Engine project [8,17] at MIT. TEK is an email-based search engine designed to serve low-connectivity communities. More information about TEK can be found at http://cag.lcs.mit.edu/tek/

2. Appropriate Information Technology

The Dictionary of Ecology and Environmental Science defines Appropriate Technology as "an applied science that is suitable for the level of economic development of a particular group of people" [1]. Another definition, which we like, is "Appropriate technology is technology that fits" [6]. Appropriate technology solutions arise when someone recognizes a technical problem or the lack of a resource, and designs a solution with an understanding of the environment that the solution was intended for. We list some examples below:

Problem	Solution
No electricity or batteries for radio	Wind up radios [7]
IV saline needed to treat cholera; difficult to set up and regulate an IV in an emergency setting.	Alternative IV Treatment Devices [12]
No radiologists in 14 sub-Saharan African countries (although x-ray machines and technicians available).	Store and forward Tele-radiology [4]

Within the domain of information technology, what is appropriate? What fits? Fraser's store and forward tele-radiology project provides physicians working in remote settings with an email account, a digital camera and software that allows them to send digital x-rays to radiologists in other locations. The radiologist can read the x-ray and email a diagnosis back to the physician. Satellife's HealthNet service has provided

Table 1. Regional bandwidth comparisons.

Region	2000 (Mbps)[1]	2001 (Mbps)[1]	Percent Growth[1]	Number online[2]
Africa	649.2	1230.8	89.6%	4.15 million
Asia	22,965.1	52,661.9	129.3%	143.99 million
Europe	232,316.7	675,637.3	190.8%	154.63 million
Latin America	2785.2	16,132.5	479.2%	25.33 million
U.S. & Canada	112,222.0	274,184.9	144.3%	180.68 million
Total	370,938.3	1,019,847.4		507.78 million

[1] From http://cyberatlas.internet.com. Accessed 3/14/02.
[2] From http://www.nua.ie; 8/2001. Accessed 3/14/02.

Table 2. Number of online users, and percentage of users online by region.

Region	Number online[2]	Percentage of online users	Regional Population[3]	Percentage of world population
Africa	4.15 million	0.82	582 million	9.7
Asia	143.99 million	28.36	3198 million	53.3
Europe	153.63 million	30.26	1194 million	19.9
Latin America	25.33 million	4.99	408 million	6.8
U.S. & Canada	180.68 million	35.58	282 million	4.7
Total	507.78 million		5934 million	

[2] From http://www.nua.ie; accessed 3/14/02.
[3] From The UNDP Development Report, 1999.

free email to health professionals in 29 countries for 12 years, providing them with access to timely and critical health information [14]. In India, a project at the M. S. Swaminathan Research Foundation uses ICT in the Information Village Research Project to dispense information from the Internet and between villages. Information such as weather conditions and market prices are changing the lives of fishermen and farmers [15].

These projects share one common theme: they are simple, innovative solutions to local technical problems.

3. Who is connected?

One domain that is full of possibilities for AIT relates to the Internet. Having lived and worked in the developing world, we know how frustrating it is not to be able to get to information. How narrow is bandwidth in various parts of the world? Looking at the bandwidth figures for various regions is illustrative.

As can be seen in Table 1, there has been a large increase in infrastructure and bandwidth supporting the Internet. However, Africa, Latin America and the Caribbean, as well as parts of Asia are still bandwidth starved. A mere 1230.8 Mbps for the African continent seems insufficient.

In Table 2 we see that in 1999, the US, with 5% of the world's population [16], had 35% of the Internet users [9]. In fact, other derivations place this number as high as 50% [16]. Again taking Africa as an example, we see that a tenth of the world's population constitutes less than 1% of the Internet users. We present and interpret these numbers with the caveat that they are constantly changing: numbers presented now will change in six month's time.

Comparing available bandwidth and the number of users in different settings can be eye-opening. Consider the figures in Table 3:

- In 1998, the country of Ethiopia had a 256Kbps connection to the rest of the Internet.
- Macalaster College, a small college in the US, has a 3 Mbps connection for 2000 students.
- In 2000, India had a total international bandwidth of 350 Mbps for 1 billion people.
- Eight Arab countries (Egypt, Saudi Arabia, Lebanon, Jordan, Morocco, Oman, Syria and the UAE) have, combined, the same bandwidth as 518 cable modem subscribers in the U.S.

Not only is there little bandwidth, but what bandwidth there is, is very expensive. One ISP in Ghana is faced with the fact that: "Internet bandwidth that would cost less than $1000 in the US must be beamed to Ghana through a

Table 3. Comparisons of available bandwidth on different continents.

Country	Provider	Date	Bandwidth	Users
Ethiopia[4]	Ethiopia Telecomm	1998	256Kbps	
	Ethiopia Telecomm	3/1999	512 Kbps	2400 users
	Ethiopia Telecomm	1/2000 (planned)	512Kbps + 1 MB	3000
USA[5]	Macalaster College	9/2001	3 Mbps	1820 students
India[6]	Various	7/2000	350 Mbps	1 billion
Eight Arab countries[7]		9/2001	777 Mbps	740,000 users

[4] http://www.bellanet.org/partners/aisi/nici/ethiopia/ethiointer.htm. Accessed 3/14/02.
[5] http://www.macalester.edu. Accessed 3/14/02.
[6] Wired. http://www.wired.com/news/business/0,1367,37232,00.html. Accessed 3/14/02.
[7] Arab Advisors: http://www.arabadvisors.com. Accessed 3/14/02.

Table 4. Email and Internet rates as of July 2001.

Location	ISP	Unlimited Email	15 Hours of Internet Access	Extra Hours of Internet Access
Malawi[8]	Epsilon & Omega	$15 / month	$30 / month	$1.50 / hour
Sri Lanka[9]	LankaNet	$11 / month	$15 / month	$1.32 / hour (peak) $0.88 / hour (off-peak)

[8] Epsilon & Omega: http://www.eomw.net
[9] LankaNet: http://www.lankanet.org

satellite link to Canada, at about $40,000 a month" [3]. At the same time, developing countries often use different pricing structures then we are accustomed to here, charging different rates for email access or full Internet access (see Table 4). Since calls are metered in many developing countries, long connection times to the ISP result in increased cost [11].

4. The Need for Different Web Applications

Given the low-bandwidth and high cost of using the Internet in low-connectivity settings, we realized that the model for Web usage that exists in this country is not compatible with the infrastructure that exists elsewhere. Downloading service packs or anti-virus upgrades is impossible at 28.8 Kbps, as it could require upwards of 5 hours. Many web pages require new plug-ins or are image-intensive. As the discussion above on the cost and shortage of bandwidth should have shown, too many 'features' on a page can make the page too costly for a low-bandwidth user to access.

Although all Web browsers do have an option for turning off images, many users do not know that this option exists nor how to do it. Furthermore, turning off images often results in the navigation icons within the page being obscured as well; if they are not replaced with <ALT> tags, the page can be rendered useless. Few Web sites are as well-designed for low-bandwidth users as the BBC news site [2]. The BBC offers a low-graphics version of every page; a few images remain, but alternative navigation links are provided.

4.1 The TEK Search Engine

Most search engines are designed for high-bandwidth, high-connectivity environments. The user types in a query, and the first 20 of a list of thousands of URLs are displayed. There is no actual content on this page, just a list of URLs. The user cannot judge whether a page is relevant until she has read the page, which she cannot do until the page has loaded. For users in low-connectivity settings, following each link, looking for the desired information, can require a large amount of time. In addition, for people coming from an "information-poor" setting, being faced with thousands of URLs in a search result can be overwhelming. Learning how to manage information is a skill acquired over time. Finally, mainstream search engines select pages without regard for their bandwidth requirements, a criterion that might be of primary interest to someone at the end of a slow connection.

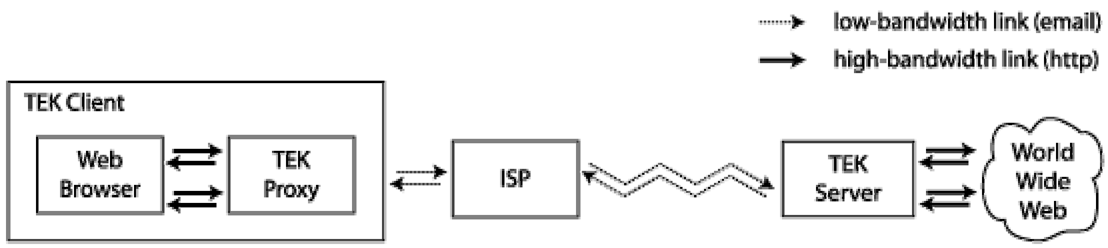

Figure 1. Components of the TEK system. Without TEK, one has to pay for a slow TCP connection as each page downloads across the low-bandwidth link. Instead, TEK uses email to bridge the low-bandwidth gap when the user is disconnected. Later, the user can browse the pages in real-time via a direct connection to the TEK Proxy.

Having recognized that searching the Web is a technical problem in low-connectivity settings, we began to look for an appropriate solution. The result is the TEK Project [8,17].

TEK – which stands for "Time Equals Knowledge" – is an email-based search engine that has three components: the *Client*, which presents a graphical interface to the user; the *Server*, which performs the searches from MIT; and a reliable email-based *Communication Protocol* between the TEK Client and the TEK Server [13]. A user enters the search terms in a web-based form resembling a search engine. The search query is automatically sent via email to the server at MIT, which performs the search using existing search engines, downloads actual pages and emails a subset of those pages back to the user. After the query results arrive on the client, the user can browse the resulting pages as if they were being retrieved from online.

The use of email as the communication medium means that the system is asynchronous and that the Client does not need to maintain (and pay for) a full Internet connection. The email protocol also introduces a time delay that the Server uses to do post-processing of search results, such as page ranking, removal of duplicate pages, page selection using various heuristics, and compression. Finally, the TEK Client is a Java proxy server that runs inside a Web browser. This makes the TEK search process as close as possible to regular Web search.

The TEK system, then, is designed to be:

1) Low-connectivity, in that it does not rely on an end-to-end connection at any point in time.

2) Low-bandwidth, in that it maximizes "information density" and only sends attachments that can be downloaded over slow connections.

3) User friendly, in that it does not overwhelm users from information-poor environments with more results than they can manage.

4) Similar to standard search engine tools, so that the skills the user acquires can be transferred to other Web tools in the future.

A discussion of the technical details and the economic justifications for the TEK architecture can be found in [17].

There are a number of email-based services that return text representations of a given web page, with some that provide an interface to search engines (e.g., GetWeb [12], www4mail [19], Web2Mail [18]). These services, however, return only the page listing the search results, instead of downloading the discovered pages and passing on the most useful ones to the client. The user must issue subsequent requests for interesting looking URLs via one of the services.

There have been numerous attempts to use a store and forward system to exchange information, such as Globelsud in Haiti [10]. While this works adequately for email (an ISP might connect to an Internet hub once an hour), it does not work for browsing the Web without a system like TEK.

5. Conclusion

In this paper, we have discussed the need for Appropriate Information Technologies. There are many interesting, challenging, technical IT problems waiting to be solved – problems that we do not see on a daily basis because we have adequate infrastructure and resources – that deserve more attention. One set of these problems relate to low Internet connectivity. We believe that once one takes into consideration the paucity and cost of Internet infrastructures in developing countries, a new class of technical research problems presents itself.

As an example of an AIT project, we presented the TEK Search Engine. TEK is designed to deliver information using the most cost-effective and reliable Internet communication structures that we could find: asynchronous email. Because we do not consider TEK to be a permanent solution, we have designed it to be as much of a “stepping stone” to true Web browsing as possible, building the user interface as a proxy server inside a Web browser. We have a working implementation of the TEK system and we are expanding our user base.

We believe that TEK is a good example of AIT research because the TEK design was based on an understanding of the context it needs to serve. If we are to convince more people to become involved in AIT projects, it will be necessary to teach people to identify technical problems in environments in which they have never lived. This remains an important topic for discussion. We cited research projects from the fields of Appropriate Technology and Appropriate Information Technology that might entice people to think more about some of these problems.

As computers become more and more central in our lives, new applications will threaten to exclude users who lack access, funds, or training. It is our responsibility to remember that not everyone has the same resources or is as well-connected. But this responsibility is not necessarily a burden: we should use the <ALT> tags in our HTML pages, transmit files in manageable sizes, and we should continue to look for the exciting, technical challenges that – while not right under our noses – are out there.

6. Acknowledgements

The TEK project is being coded with the help of Damon Berry, Tazeen Mahtab, Saad Shakhshir and Binh D. Vo. David Clark, Hamish Fraser, and Peter Szolovits are always available to offer discussion and support.

7. References

[1] H. Art, ed. The Dictionary of Ecology and Environmental Science. Henry Holt, New York, 1993.

[2] BBC. http://news.bbc.co.uk/

[3] H. Bray. Critics don’t deter Internet visionary. The Boston Globe, 7/24/2001.

[4] H.S. Fraser, D. Jazayeri, L. Bannach, P. Szolovits, S.J. McGrath. Telemedmail: free software to facilitate telemedicine in developing countries. Medinfo. 2001:10(Pt 1):815-9.

[5] http://digitalmass.boston.com/news/packages/africa/part3_sidebar.html. Accessed 3/2002.

[6] Journey to Forever. http://journeytoforever.org/at.html

[7] InnovativeTechnologies http://windupradio.com/

[8] L. Levison, W. Thies, S. Amarasinghe. The TEK Search Engine. Position paper at the Development by design workshop, MIT, Boston, MA. July 2001.

[9] NUA Internet Serveys. http://www.nua.ie/

[10] J. Peha. Lessons from Haiti's Internet Development. Communications of the ACM. vol 42, no. 6; 1999.

[11] B. Petrazzini, M. Kibati. The Internet in Developing Countries. Communications of the ACM. vol 42, no. 6; 1999.

[12] T. Prestero, M. Height, R. Hwang. Rapid Cholera Treatment: Exploring Alternative IV Treatment Devices. Presented at the Development by Design conference, MIT, July 2001.
http://www.mit.edu/~tprester/DtM/dyd01_cholera.pdf

[13] J. Prevost. A Reliable Low-Bandwidth Email-Based Communication Protocol, Master's Thesis, Massachusetts Institute of Technology, 2001.

[14] SatelLife. http://www.satellife.org

[15] M. S. Swaminathan Research Foundation, Information Village Research Project. http://www.mssrf.org/

[16] UN Development Programme (UNDP). Human Development Report. New York: Oxford University Press, 1999.

[17] W. Thies, J. Prevost, T. Mahtab, G. Cuevas, S. Shakhshir, A. Artola, B. Vo, Y. Litvak, S. Chan, S. Henderson, M. Halsey, L. Levison, and S. Amarasinghe. Searching the World Wide Web in Low-Connectivity Communities. To appear: Proceedings of The 11th International World Wide Web Conference, Global Community Track, May, 2002.

[18] Web2Mail. http://www.web2mail.com/

[19] WWW4Mail. http://www4mail.org/