Searching the World Wide Web in Low-Connectivity Communities

William Thies, Janelle Prevost, Tazeen Mahtab, Genevieve T. Cuevas, Saad Shakhshir, Alexandro Artola, Binh D. Vo, Yuliya Litvak, Sheldon Chan, Sid Henderson, Mark Halsey, Libby Levison*, and Saman Amarasinghe

thies@mit.edu, prevostj@alum.mit.edu, {tmahtab, gtcuevas, saads, aartola, bdv}@mit.edu, ylitvak@eecs.tufts.edu, {sheldons, sid, mhalsey, libby}@mit.edu, saman@lcs.mit.edu

Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02138 USA

Abstract

The Internet has the potential to deliver information to communities around the world that have no other information resources. High telephone and ISP fees - in combination with low-bandwidth connections - make it unaffordable for many people to browse the Web online. We are developing the TEK system to enable users to search the Web using only email. TEK stands for "Time Equals Knowledge," since the user exchanges time (waiting for email) for knowledge. The system contains three components: 1) the client, which provides a graphical interface for the end user, 2) the server, which performs the searches from MIT, and 3) a reliable email-based communication protocol between the client and the server. The TEK search engine differs from others in that it is designed to return low-bandwidth results, which are achieved by special filtering, analysis, and compression on the server side. We believe that TEK will bring Web resources to people who otherwise would not be able to afford them.

Keywords: Search engine, low connectivity, low-connectivity communities, information retrieval

1. Introduction

In many of the world's communities, there are no books, there are no libraries, and 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. Electricity is often intermittent. Bandwidth is so narrow that it can take the user a long time to find what she is looking for when browsing the Web, since she has to wait for each page to load. Moreover, time spent online translates to higher telephone and ISP charges, which quickly become prohibitive when baseline fees are 10% of a local wage. Finally, unreliable network infrastructures can sometimes prevent access to the Internet altogether.

It is easy to assume that the World Wide Web has made information available to everyone, everywhere. However, it simply is not so. Search engines designed for high-bandwidth connections, applications which are not backwards compatible, and predominately English-

* Corresponding author. Email: **libby@mit.edu**

language Web content imply that a Spanish speaker in Guatemala with a 1995 operating system stands little chance of reading a discussion on biodiversity. The worst connected region of the world, in terms of Internet bandwidth per person, is Africa, which has 13% of the world's population but only 0.15% of international Internet connections [Economist 2001]. In 1998, North America contained less than 5% of the world's population, but more than 50% of all Internet users [UNDP1999]; it is no surprise that Web systems focus on the niche market of North America and Europe.

If we are going to make information available to the global community, we need to design systems that account for the varied stages of information technology and Internet connectivity that exist. In this paper, we describe one part of the solution: an Internet search utility designed for low-connectivity, low-bandwidth communities. The TEK Search Engine - TEK stands for "Time Equals Knowledge" - is an asynchronous search engine that transfers both queries and query results by email. Queries are received in Boston, the Web searched, and a subset of the 'found' information is returned to the user.

Our research focuses on the development of novel technologies that are specifically designed to meet the economic and social constraints of the developing world. While this research incorporates familiar fields such as information retrieval, data compression, multi-lingual interfaces, and low-cost devices, it is distinct from most research conducted in the West in that it is driven to meet the costs and constraints of developing nations. TEK is not only an imminent solution to a social need; it is also the first step in a long-term effort to develop appropriate information technologies for developing countries.

2. The TEK System

Most search engines are designed for high-bandwidth, high-connectivity environments. That is, they optimize for speed, assuming that a user can quickly look through the returned links and immediately run a second, modified search if she is unhappy with the results of her first search. This tight feedback loop between the user and search engine is inappropriate for low-connectivity sites in the developing world where the bottleneck is the time required to transfer the information, rather than the server's delay in finding the information. In a low-bandwidth environment, exploring the various links on a search engine's results page can require a large amount of time.

Also, 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. In addition, standard search engines return all unique URLs that matched the user's query; for users in information-poor environments, receiving hundreds of thousands of URLs might be more overwhelming than useful. Learning how to manage information is a skill acquired over time. Finally, we do not expect these countries to permanently face limited access: technology changes quickly, and we expect that within a decade many more locations will have full Internet access. We did not want, in designing TEK, to impose a technology that would need to be completely changed in a few years.

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.

The resulting system has three main components:

1. An email-based communication protocol that manages the transfer of information over the unreliable connections between the client and the server [Prevost2001]. It is a store-and-forward system that runs asynchronously. The client thus does not need to ensure a permanent Internet connection.
2. The TEK server, which performs searches from MIT and sends the processed results back to the client. The server performs specialized ranking, filtering, and compression of the search results to make them as bandwidth-friendly as possible. It also records which URLs have already been sent to each client to avoid repetition and wasted bandwidth.
3. The TEK client, which provides a graphical user interface for constructing queries and viewing results. The client is a Java proxy server that runs inside a web browser. It provides the user with practice using a Web browser, clicking on links and searching.

The next section describes the details of these components.

3. TEK Details

We will describe the end-to-end implementation of the TEK system in the context of an example: a student who wants to search for information on "solar food dryers". The scenario might proceed as follows (see Appendix for illustrations):

3.1 Entering the Query: the TEK Client

We expect that the TEK client is installed on a machine in a school or tele-center and that it supports multiple users. When the student starts TEK, it appears as a set of local web pages that are viewed with a browser. The TEK Client is implemented as a proxy server that interacts with the user through a normal web browser. In the installation process, the user's web browser is configured to communicate with the local TEK proxy server instead of communicating directly to a server on the Internet. When a user wants to view a web page, she can enter the URL in the browser just as if she were connected. If the TEK Client has cached the page locally, then the proxy server displays the local page without connecting to the Internet.

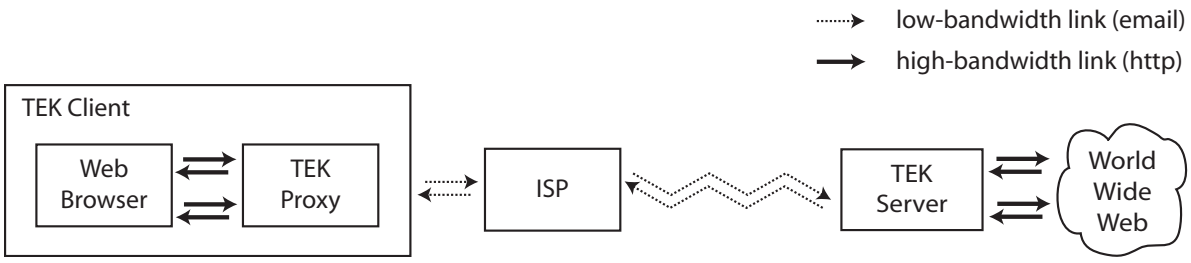

Figure 1: Components of the TEK system. Without TEK, one has to pay for a slow TCP connection as each page downloads across the low-bandwidth link. Instead, TEK uses email to bridge the low-bandwidth gap when the user is disconnected. Later, the user can browse the pages in real-time via a direct connection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

When the user is seeking information that is not already available on the client machine, she accesses the local page **http://tek/** which provides a search interface similar to standard search engines. The student must enter her username and password, after which she can also view all of her previous search results. There is a special administrator interface for managing user accounts.

The user enters the search terms, "solar food dryer", in a web-based form resembling a standard search engine. Because TEK can not assume full Internet connectivity, the TEK Client encodes each query as an email message that is stored on disk. After the student confirms her search request, the query is placed under the student's "pending query" list and scheduled for mailing. The student can now log off - everything else is done automatically by the system. At a convenient time, the site administrator connects to the Internet and emails the accumulated queries to the TEK Server (which has a high-bandwidth connection to the Internet backbone). Administrators can choose to send and receive email as frequently as the telecommunication infrastructure and their ISP's pricing scheme will allow.

In addition to full Internet searches, the TEK Client provides a utility that allows the user to request specific URLs. If, in the process of reading a Web page, the user discovers a link to a missing file, the interface prompts the user to request the page. The user can also enter the address directly. These *direct fetch* requests are also encoded as emails and queued for delivery.

Each TEK Client will begin with a set of HTML documents in its local cache. In addition, we plan to include static reference material (*e.g.* a dictionary and an atlas).

3.2 Communicating with the Server: the TEK Protocol

The next step is for the client to email the query to the server. This process can be initiated automatically by the TEK client, perhaps in the middle of the night when the telephone rates are cheapest[1] and there is less demand for phone lines.

Although email is an established means of communication, it is by no means perfect -- emails can be lost, duplicated, reordered, or corrupted, especially in low-connectivity regions that have unreliable network infrastructures. To insulate the TEK Client and TEK Server from these problems, TEK includes a special protocol that attempts to guarantee reliable communications over an unreliable medium [Prevost2001]. Generally speaking, the protocol works by keeping track of which messages have been sent, and which ones were replied to. If a reply is not

received within a given time, then the protocol resends the original message.

The TEK Protocol, loosely modeled after TCP, follows a request-reply model. That is, each message type has an expected reply message. If the client requests information from the server, and the server does not reply within a given amount of time, the protocol will automatically resend the request, assuming the original message was lost. The request-reply model allows us to eliminate acknowledgement messages which would increase both bandwidth and the number of messages required; in fact, such acknowledgements would further complicate the model since they could be lost or corrupted as well. The TEK Protocol has a fixed timeout (configurable by each TEK Client); messages are automatically resent if a reply does not arrive.

The TEK Protocol also handles additional bookkeeping: the Protocol supports special messages for client registration, tracks the **last heard from** date for each TEK Client, and can send **Ping** messages as needed. The TEK Protocol also transmits Client preferences (*e.g.* language) or information regarding changes in Client system hardware (which might enable the TEK Client to download a larger result set).

3.3 Query Processing: the TEK Server

The TEK Server searches for, finds and returns personalized, low-bandwidth results to the TEK Client. When the TEK Server receives a search query, it retrieves a set of candidate pages by invoking existing search engines such as Google and AltaVista. We send the same query to multiple search engines, as different search engines index - and therefore find - different portions of the Web [Lawrence1999].

There are two primary ways in which the TEK Server differs from other Internet search engines: 1) it keeps a full record of each client's characteristics and search history, and 2) it filters content from the resulting pages, optimizing for bandwidth rather than response time. (Because TEK is asynchronous, the server has the time to perform this post-processing.) Together, these characteristics distinguish the TEK Server as a unique information delivery service for low-connectivity communities.

3.3.1 Keeping Track of the Client

When an administrator installs TEK on a client machine, she fills out a registration form that is sent via email to the TEK Server. The registration includes such information as the geographical location of the client, the intended use of the machine, and the speed of the network connection. From this point forward, the TEK Server keeps track of all searches made by the users at that client, as well as all the URLs returned in each search result set.

The TEK Server can leverage this client database to return more relevant and lower-bandwidth results. For example, the server is careful not to waste bandwidth by sending a page that the client has already downloaded; instead, the server just sends a reference to the local page, and expands the results by sending new information that the client does not already have. In more general terms, there is potential to decrease the bandwidth requirements of pages by taking advantage of known data on the client side. Protocols that take advantage of state on the client

have been left largely unexplored in the United States, since it is simpler just to download the material again across a high-bandwidth connection. However, the savings could be significant for users in developing countries, and we hope to explore such technology as part of our future research.

3.3.2 Filtering the Result Set

As suggested by the discussion above, the TEK Server performs considerable post-processing on the set of candidate web pages in order to reduce the bandwidth requirements and to eliminate content that is redundant or irrelevant for a given client. Such processing is generally absent from mainstream search engines in the United States, since users can evaluate the results manually over high-bandwidth links and quickly submit a modified search if they are unhappy with the original. However, with users that are not expecting a synchronous reply to a query, the TEK Server has the luxury to take its time in filtering the content that is returned.

The following are examples of post-processing techniques that are employed by the TEK Server:

1. **Removing duplicate content.** All duplicate pages and pages that are very similar to each other (such as mirror pages from different sites) are eliminated from the candidate set.
2. **Clustering.** We use Baeza-Yates' algorithm [Frakes1992] to group the results into clusters of similar documents and choose a few representative documents from the most important clusters to send back to the user. This method eliminates redundancy and extends the breadth of the results, improving the chances that at least one page will be relevant to the user's query. For instance, in the search for "solar food dryers", we would aim to return information on: how to build a solar food dryer, how to use a solar food dryer, and what food to dry in a solar food dryer. Sending some pages from each cluster improves the likelihood that at least some of the information sent will be relevant to the aspect of solar food dryers that the student was most interested. Even in cases where all of the categories are relevant to the user's interests, clustering ensures that each page adds distinct information content to the results.
3. **Removing images.** At this point, images are removed from the resulting pages unless the client indicates that they should be preserved. In the future, we plan to support more advanced techniques for decreasing the bandwidth requirements of images.
4. **Removing background code.** The source code for HTML pages can often be distilled without affecting the appearance of the page. For instance, HTML comments and META tags can be removed. Further, there are many cases in which complex formatting code can be eliminated without affecting a page's content.
5. **Distinguishing content from links.** One goal of the TEK Server is to identify pages that are "highly informational", or that contain content, instead of references to other sources of information. In TEK's asynchronous model, authority (content) pages are preferable to hub (link) pages [Kleinberg98]. If a user is searching for an author, at least one of the results returned should contain the author's name in paragraph text instead of just in a bibliographic reference. A set of links might facilitate future browsing in the United States; TEK's goal is to deliver enough content to describe the subject of interest rather than referring the user to other sources of information.

6. **Compressing the result set.** All of the results are compressed into a zip file before sending them back to the client, thereby further reducing the bandwidth needed to download the results.

Following the server's processing, the results are emailed to the client using the communication protocol described above.

3.4 Viewing the Results

When the results arrive on the TEK Client, the student needs to login to view them. The interface for viewing results is a special front page - constructed by the server - that organizes the pages by cluster and provides a link to each. The user can then browse through the pages as if they were being retrieved from online. Because all pages are cached locally there is no delay while each page is retrieved.

An added feature of the TEK client is that it accumulates the information from each search into a local digital reference library. This library serves as a miniature, offline version of the Web, allowing users to follow links from page to page as long as the referenced pages were downloaded during a preceding search. The user interface provides a local search utility so that the user can search the collection of local pages. Only when the information is not found locally is it necessary to send a query to the TEK server. In other words, if another user had previously searched for the same information, an Internet search can be avoided.

4. Rationale

In this section we argue that the TEK system will make Internet access cheaper, more robust, and, in some respects, even more convenient for users in low-connectivity regions.

4.1 Reduced Operational Costs

There are numerous ways in which TEK will lower the cost of Internet access for the end user. In some regions, email-only accounts are much cheaper than accounts that allow full access to the World Wide Web. While the average cost in Africa in 1997 for 5 hours of dialup service was $50, the charges varied greatly -- from $10 to $100 [Akoh2001]. This compares to unlimited access in the US at $22[Inter.net2002]. Assuming 20 hours of access, the African average rises to $200 which is highly unaffordable given the differences in per capita income. As illustrated in Table 1, email-only accounts are less expensive in many places.

Thus, TEK will make Web resources available to those who could otherwise afford only email. In addition, telephone lines are often clearer, more stable, and cheaper to use during off-peak hours: Arminco Global Ltd in Armenia offers discount rates for Internet night access [Arminco]. TEK can be set up to run during those times.

Location	ISP	Unlimited Email	15 Hours of Internet Access	Extra Hours of Internet Access
Malawi [2]	Epsilon & Omega	$15 / month	$30 / month	$1.50 / hour
Sri Lanka [3]	LankaNet	$11 / month	$15 / month	$1.32 / hour (peak) $0.88 / hour (off-peak)
China [4]	n/a	n/a	$70 / month	n/a
Kenya [5]	n/a	$10 / month	$100 / month	n/a
Armenia [6]	Arminco	$8.50 / month	$42 / month (500MB quota)	$0.24 / MB traffic fee
Brazil [5]	IBM Brazil	N/a	$40 / month (20 hours)	$2.75 / hour
Nigeria [7]	Microcom Systems Ltd.	N/a	$58 / month	n/a

Table 1: Email and Internet rates as of March 2002.

The TEK system also decreases costs by shortening the duration of each phone call to the ISP. First, the connection is shortened because the client machine spends all of its time either sending a query or downloading results; unlike Web browsing, there is no idle time during which the user is reading pages or contemplating what to do next. Second, when the results are being downloaded, all of the content is available on the ISP; the user does not have to wait for the ISP to fetch information from other sources. Third, the results themselves are more compact, since they are filtered and compressed on the server side. Since calls are metered in many developing countries, a shorter connection time to the ISP will bring about a further savings [Petrazzini 1999].

Finally, there will be further savings if some TEK searches can be eliminated altogether - which will happen when the local search utility finds the sought information in the client's local digital library. To emphasize that it is a cost-effective strategy for the client to keep a persistent copy of each page that it downloads from the server, let us consider a few calculations (see Table 2). Assuming that a 75 GB hard disk costs $250 dollars, it follows that one megabyte (MB) of hard disk space costs $0.0032. On the other hand, downloading one MB of data over a 28.8 kbs modem at a rate of $1.75 per hour would cost $0.104 - more than 32 times as much as storing the data on disk! And this figure assumes a perfect utilization of the modem's bandwidth; with a more realistic utilization between 1% and 10%, retrieving pages over the phone becomes three orders of magnitude more expensive than storing them on disk. Thus, even if there is only a 1% chance that a downloaded page will be needed again in the future, it is economically advantageous to buy a hard disk on which to store downloaded pages, rather than planning to download them a second time. Note that, given the Internet prices in the United States, these results are reversed - *i.e.*, there is not an economic incentive to support an extensive client-side digital library.

Retrieval Method		Price	Price per MByte	Relative Cost per MByte
Hard disk		$250 / 75 GB	$0.00325	1.0
28.8 kbs modem (Sri Lanka)	100% utilization	$1.75 / hour	$0.104	32.0
	10% utilization		$1.04	320.0
	1% utilization		$10.4	3200.0
128 kbs Cable / DSL (USA)	100% utilization	$30 / month unlimited access	$0.00074	0.23
	10% utilization		$0.0074	2.3
	1% utilization		$0.074	23.0

Table 2: Estimated costs of local storage vs. remote fetch as of July 2001.

In addition to operational costs, there are setup costs that will be incurred for a system like TEK. The cost of an average personal computer is far higher in the developing world than it is in the United States: in Ethiopia, a computer costs 15 times the per capita GDP [Mannisto1998]. In India, tariffs on imported computer hardware approach 120 per cent [Panos1998]. In addition, there has been limited development of power supply networks in many countries [Akoh2001]. Teledensity - the ratio of telephones per 100 people - is .48 in sub-Saharan Africa (*i.e.*, there are 4 telephones for every 1000 people)[Mannisto1998]. The service that exists is often unevenly distributed with the majority of service available only in the cities; consequently, ISPs primarily service the urban population. Any plans for deployment will need to take this into account [Petrazzini1999].

4.2. Improved Reliability

TEK improves the robustness of Web access by reducing the user's dependence on the ISP's external network. That is, when the user wants to browse the Web in real-time, two connections need to be working: from the user to the ISP, and from the ISP to the rest of the world. However, with an email-based protocol, these connections are decoupled. First, over some period of time, there needs to be a working path from the MIT server to the user's ISP. Then, at some other time, the user needs to connect to the ISP and download the results. In other words, it is possible to obtain a page using TEK even if the page is constantly unavailable to a Web browser using the same ISP.

Assuming that the client sends and receives TEK emails once per day, the user can expect to find the results of a query within 48 hours (since the query will be sent within 24 hours, and the results received within the next 24 hours). In cases where the email is delayed or lost en route, the communication protocol automatically manages the retransmission procedures.

4.3. Improved Convenience

At first glance, it might appear that TEK is inconvenient because of the delay it imposes between searching and receiving the results. However, there are many ways in which using TEK is more convenient than using an online Web browser in a low-connectivity area. Primarily, once the results have arrived via email, one can browse through them all in real-time, instead of enduring the slow, unreliable, and frustrating process of loading each page when one is connected. Further, one can look at the results at any time that is convenient, and the results will remain available to all users of the machine as long as there is space on the hard drive. The results themselves might be more relevant to the user's query, since the TEK server spent more time analyzing and processing the results than conventional, speed-optimized search engines. Finally, TEK's nighttime download feature could free up one's phone line for other uses during daylight hours, as well as avoiding phone line congestion in trying to connect to the ISP during peak hours.

5. Related Work

There are a number of search engines that have something in common with TEK. Google [Brin1998] eliminates pages that are very similar; Northern Light and Vivisimo perform clustering of pages, and MetaCrawler invokes multiple search engines to perform the search. However, all of these search engines are optimized for speed. TEK is fundamentally different in that it is optimized for low-bandwidth and low-connectivity.

Orthogonally, 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, www4mail, Web2Mail). 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. Moreover, they lack two of TEK's key features: 1) a server that records which pages are already on the client, thereby eliminating redundant client/server communication, and 2) a series of specialized information retrieval techniques that filter, analyze, and compress the results on the server before sending them to the client.

In the context of filtering and compression, there are a number of related works that TEK can leverage. For example, [Fox96] considers reducing the resolution and color depth of images to suit low-bandwidth users, and [Douglis98] detects differences in web pages to reduce the bandwidth needed to update cached copies. In the context of bandwidth utilization, [Fan98] speculatively prefetches pages while a client is idle but connected. While many browsers maintain temporary web caches, they are hidden from the users, they are not readily searchable, and they require a full Internet connection to populate them.

Interestingly, some of the work that most resembles TEK in purpose is that of the Interplanetary Internetworking Group at JPL [Cerf2001], which is designing a network layer for communication between Internets on different planets. Like TEK, IPN faces: intermittent connectivity; narrow bandwidth; privacy concerns; a wide variety in user platforms (requiring backwards compatibility); and long communication delays that require "non-chatty" communication protocols. However, instead of building an email-based protocol on top of the

existing network infrastructure, they are designing a separate "bundle layer" in the network itself that deals with low-connectivity communications.

Finally, there have been numerous attempts to use a store and forward system to exchange information, such as Globelsud in Haiti [Peha 1999]. While this works adequately for email (an ISP might connect to an Internet hub once an hour), it does not work for Web browsing.

6. Discussion and Future Work

We have implemented a fully functional prototype of the TEK system. There are many questions that we will not be able to further research until the system is deployed and we can gather usage statistics. How broad is each location's knowledge needs? How much repetition and overlap is there among queries? What information should initially be included in the local library on the client machine? How do information needs differ in different cultures? While fascinating, these questions must all wait. We have designed the TEK system to be flexible, such that the specific information retrieval techniques it employs can be adjusted depending on observed usage patterns. There are also a number of specific items that we have identified for future work.

One obstacle in effective search is that users rarely construct adequate search descriptions. We suspect that information seekers in low-connectivity areas might be more willing to invest effort in constructing a query, since they must wait longer to receive the results. Thus, we plan to include a more sophisticated query builder to help ensure that a query is appropriate - for instance, by detecting spelling errors. A word-frequency table would let the TEK Client evaluate the semantic content of a query term or the number of likely 'hits'; the system could prompt the user for a different, or an additional, term. In addition, we will investigate what kind of information is most useful to include in a search query, and how this information can be effectively utilized from within the server to direct the Web search. For instance, if the user provides a description of the type of information desired when sending the query, we could dispatch the search to a specialized search engine -- *e.g.* to MapQuest for maps and to the BBC or CNN for news. In addition, the user might include a reference to an "example result" with the search description -- she could send a document on growing corn, when she needs one on growing rice. The example document can be used to identify a new result set, either by returning to the Web site where the original document came from, or by doing similarity matching between the example and found documents.

On the server side, we would like to issue our queries to more search engines. Different search engines index different portions of the Web; [Lawrence 2000] and [Notess 2000] have found that results of a query vary among search engines and are, in fact, often non-overlapping. By querying different search engines we can achieve better coverage of the result space.

In order to lower bandwidth, we are investigating extending search functionality with document summarization. The user will specify whether she would like a small number of full documents or a larger number of summarized documents. In the latter case, a request form based on the returned results will allow the user to identify which documents she wants full versions of. In this approach the user is again trading time for knowledge: it also provides an iterative step in which the user further specifies her search request.

In addition, a mechanism to gather feedback from the users of TEK on the usefulness of each returned page will be critical for evaluating the effectiveness of the heuristics employed by the server. The server could even use this information on a client-by-client basis to choose the search methodologies that are best suited to a given user. Also, we will have to expand TEK to support other languages.

Finally, TEK as it stands now is an information retrieval system. We hope to investigate ways to extend TEK to make it an information delivery system as well. We do not believe that connectivity-poor populations should be restricted to read-only access to the Internet. Isn't the user in Malawi more likely to be an authority on Malawi than someone in Boston? Users need to be able to contribute information into the global community. To meet this need, we envision an extension of TEK that allows users to email web pages to be published on our server.

7. Conclusion

If we are going to make information available to the global community, we need to design systems that account for the varied stages of information technology and Internet connectivity that exist. In this paper, we have described the TEK Search Engine: an Internet search utility designed for low-connectivity, low-bandwidth communities. TEK is part of a broader initiative within the MIT Laboratory for Computer Science to find engineering "shortcuts" to help narrow the technology gap faced by the developing world [Dertouzos2000, Dertouzos2001].

But is access to the Internet the best thing to provide the developing world, where food, safe drinking water, essential drugs and housing are also desperately needed? We would argue that TEK is providing access to information, information that is selected by the user. As the final report of the DOT Force stated: *"ICT cannot of course act as a panacea for all development problems, but by dramatically improving communication and exchange of information, they can create powerful social and economic networks, which in turn provide the basis for major advances in development."* [DOT Force2001; p.3]

TEK is a technical solution to a social need. From its conception, TEK was based on an understanding of the cultural and global context it needs to serve. While cutting-edge information technology tends towards "more information, faster," TEK is designed to work in a low-connectivity, low-bandwidth setting, where the aim is to guarantee the delivery of "better information, slower."

We do not consider TEK to be a permanent solution to the problem of providing Internet access in developing countries. Instead, we believe that there is a need for an interim solution - a more reasonable way for people to access the Internet - while more ambitious and long-term telecommunication initiatives are implemented. By its gains in affordability, reliability, and convenience, we believe that TEK will meet exactly that need: it will bring Web access to some people who would otherwise be without it. As such, TEK is also our first step in a long-term effort to develop appropriate information technologies for developing countries.

8. Acknowledgements

We have had advice and discussions with: Damon Berry, David Clark, Michael Dertouzos, David Karger, Jaime Teevan, Lynn Andrea Stein, and Peter Szolovits. Thank you.

This work was partly funded by a Faculty Fellowship from Singapore University, a Graduate Fellowship from Siebel Systems, and the Summer Undergraduate Research Opportunity Program at the MIT Laboratory for Computer Science.

9. Appendix

9.1 TEK Screenshots: Sending a Query

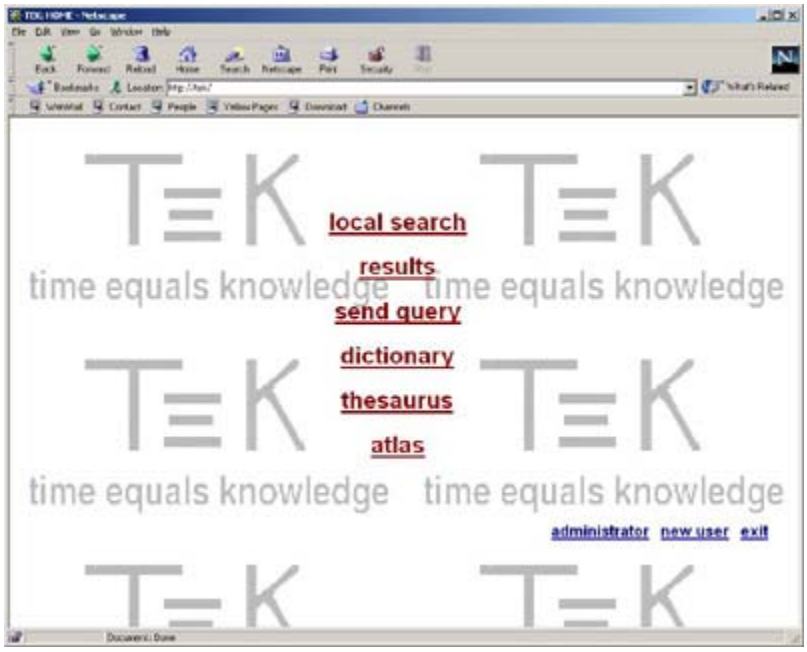

Figure 1: The TEK front page allows users to conduct different types of local searches and remote queries, as well as to view results.

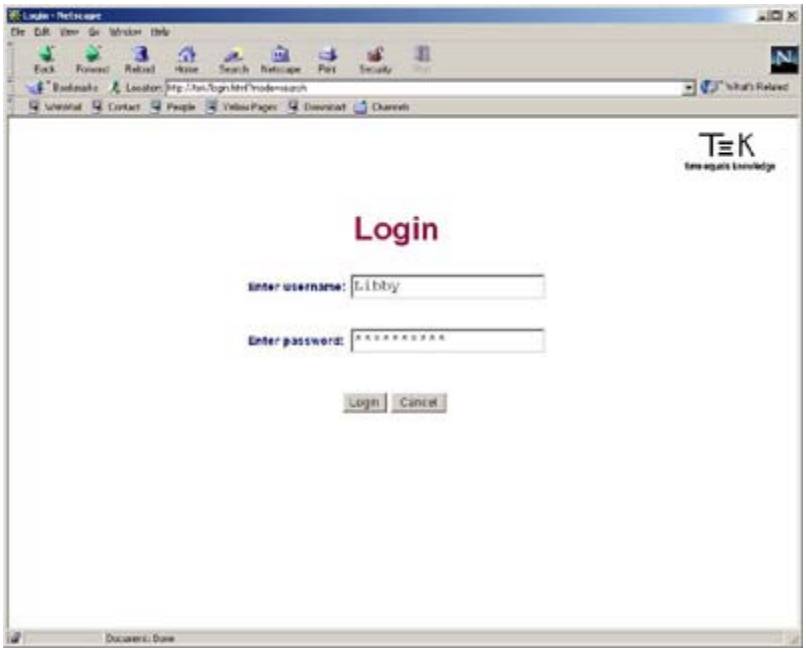

Figure 2: To place a remote query or view results, the user must first login.

Figure 3: After logging in, the user can perform remote searches, including advanced search, specific URL request, and image search.

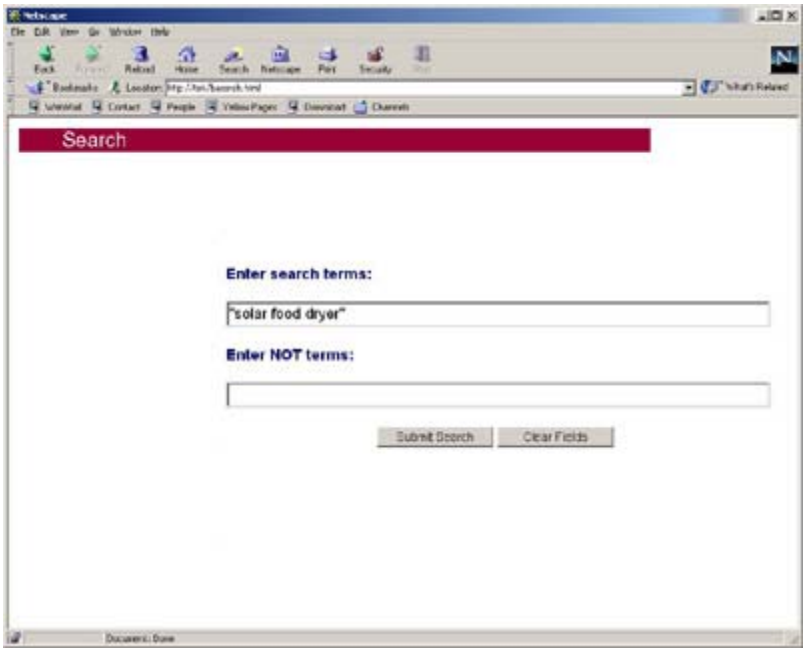

Figure 4: The basic TEK search interface has two fields: one for terms that must appear, and one for terms that must NOT appear.

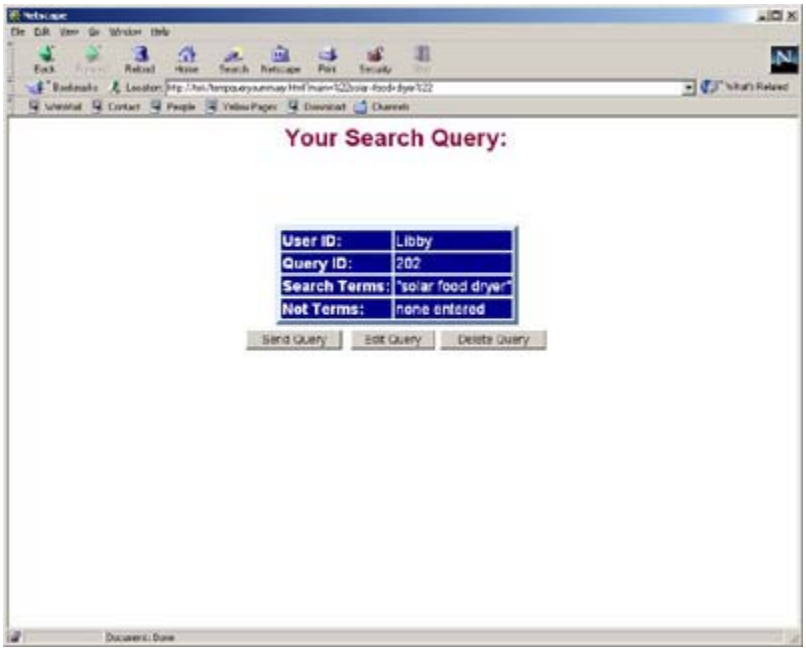

Figure 5: The user is asked to confirm the query.

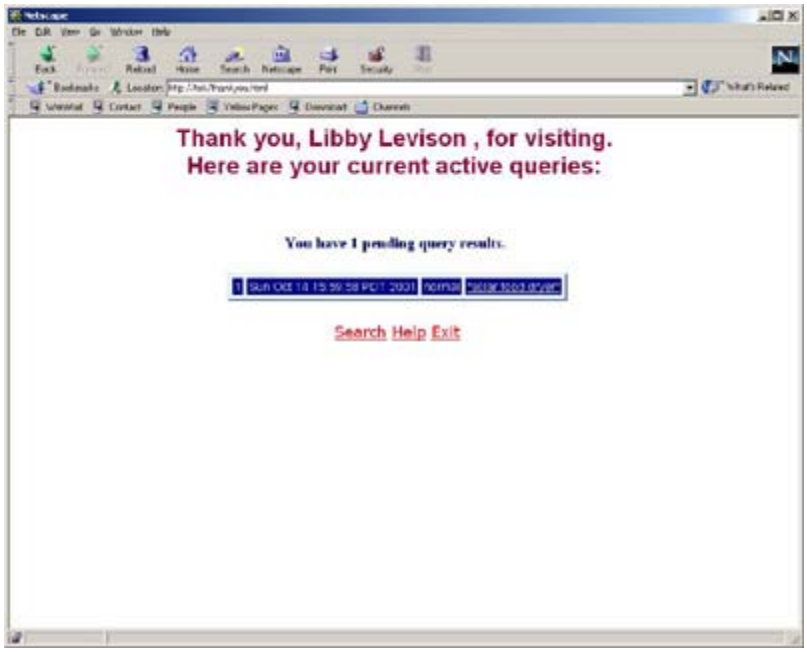

Figure 6: Confirmation that the query is complete.

9.2 TEK Screenshots: Viewing Results

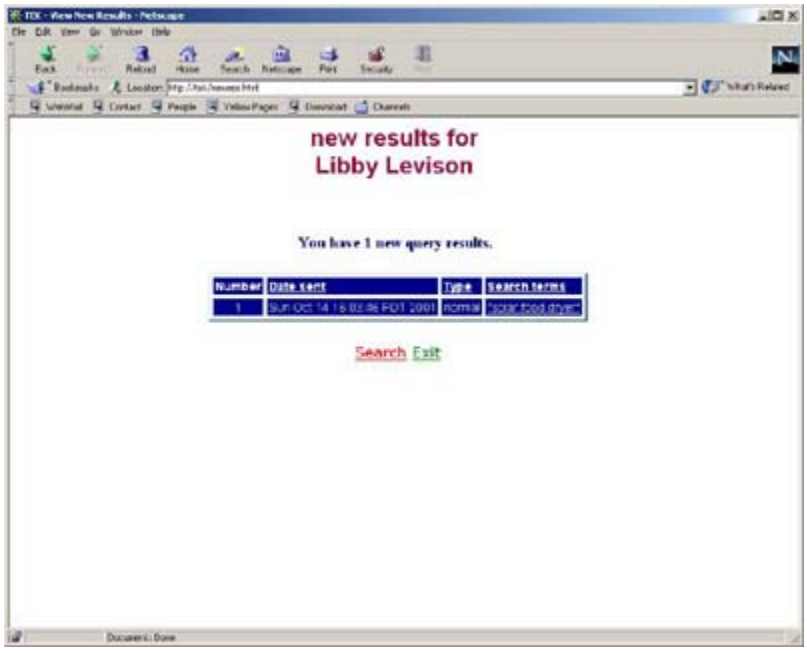

Figure 7: After logging in, the user can see a list of recently returned query results.

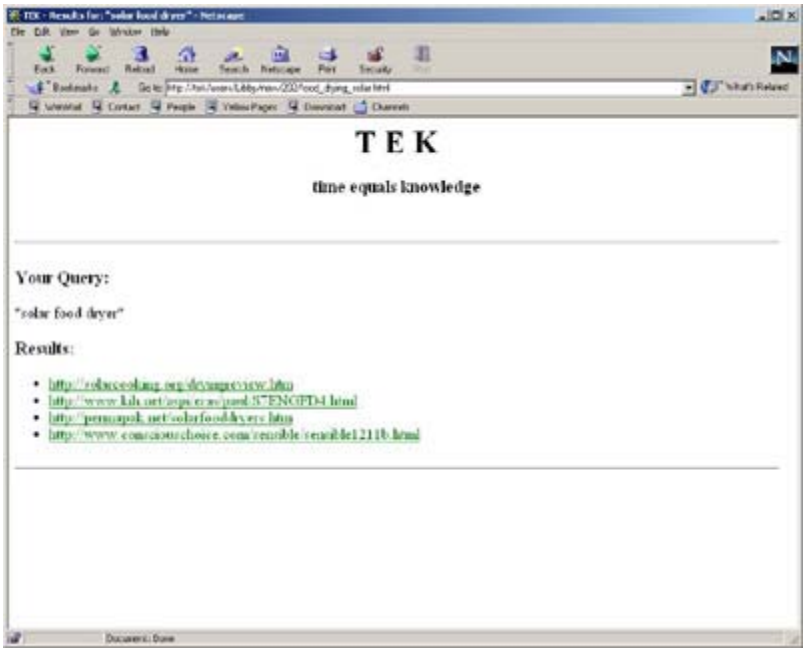

Figure 8: TEK presents the set of pages corresponding to the user's query.

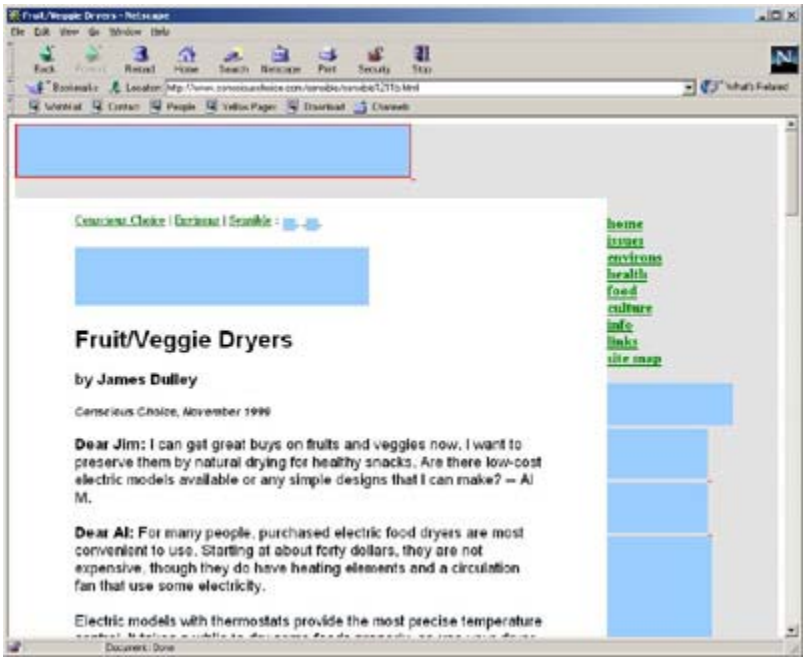

Figure 9: A resulting page, as seen by the user. TEK refines pages, removing images to save bandwidth.

Figure 10: The original, unrefined version of the page seen in Figure 9.

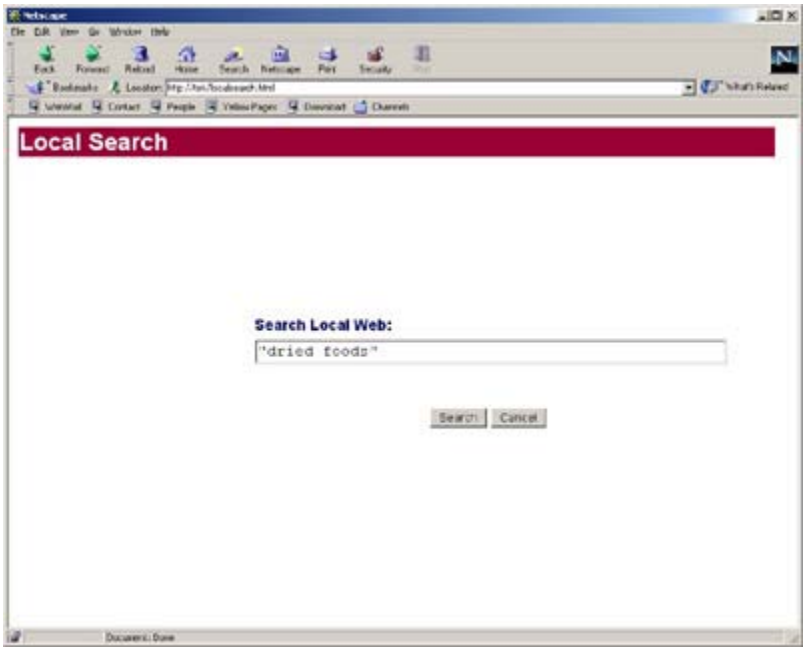

Figure 11: Returned results are stored in the local database, which can be searched with a local engine.

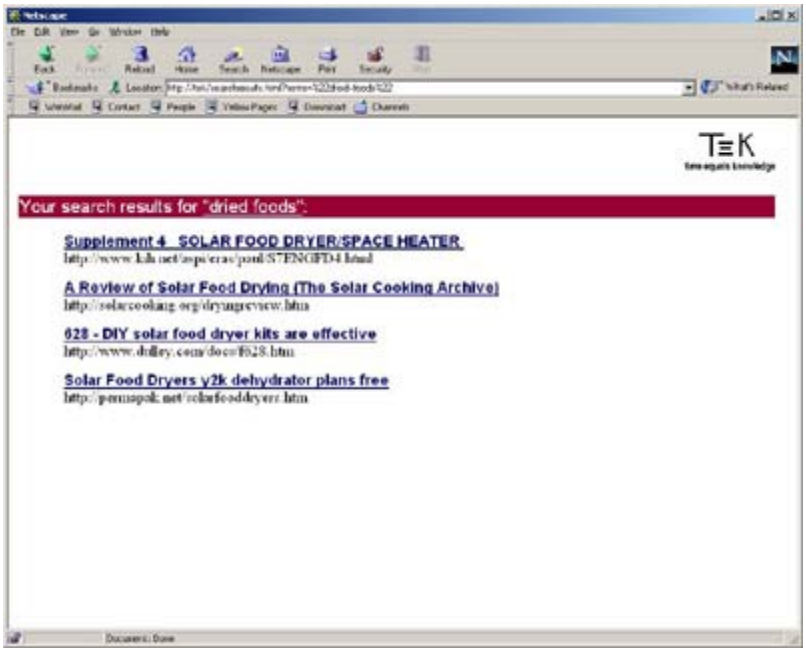

Figure 12: Results of a local search.

10. Endnotes

1 Arminco Global Telecommunications offers users a "night surfer" rate; between the hours of 8pm and 8am the fee is reduced by more than 50% [Arminco].
2 Malawi: http://www.eomw.net
3 Sri Lanka: http://www.lankanet.org
4 China: [Panos1998]
5 Kenya, Brazil: [Petrazzini1999]
6 Armenia: http://www.arminco.com/services.html.
7 Nigeria: http://www.micro.com.ng/

11. References

[1] Ben Akoh. E-Business in the Developing World, Africa and Ethiopia. Presented at the Conference on Information and Communication Technology and Development. 2001
[2] AltaVista. http://www.altavista.com.
[3] Arminco Global Telecommunications. http://www.arminco.com/services.html.
[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In Seventh International World Wide Web Conference, Brisbane, Australia, 1998.
[5] Vinton Cerf. Interplanetary Internet (IPN): Architecture Definition. Work in progress. Internet Working Group. May 2001.
[6] M.L. Dertouzos. Personal communication. 2000.
[7] M.L. Dertouzos. The Unfinished Revolution: Human-Centered Computers and What They Can Do for Us. New York, NY, HarperCollins Publishers, 2001.

[8] Digital Opportunities Task Force. Digital Opportunities for All: Meeting the Challenge. 2001.
[9] Fred Douglis, Tom Ball, Yih-Farn Chen, and Eleftherios Koutsofios. The AT&T Internet Difference Engine: Tracking and Viewing Changes on the Web. World Wide Web, January 1998, pp. 27-44.
[10] Li Fan, Pei Cao, and Quinn Jacobson. Web Prefetching Between Low-Bandwidth Clients and Proxies: Potential and Performance. In Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '99), Atlanta, GA, May 1999.
[11] Armando Fox and Eric Brewer. Reducing WWW Latency and Bandwidth Requirements by Real-Time Distillation. Computer Networks and ISDN Systems, Volume 28, Issue 711, p. 1445. May 1996.
[12] GetWeb. http://www.satellife.org/webcontent.php
[13] Google. http://www.google.com/
[14] William Frakes and Ricardo Baeza-Yates. Information Retrieval: Data Structures and Algorithms. Prentice-Hall. 1992.
[15] Inter.net http://www.us.inter.net/
[16] Jon Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms.* 1998.
[17] Steve Lawrence and C. Lee Giles. Accessibility of information on the web. Nature, 400, 107, 1999.
[18] Steve Lawrence. Context in Web Search. IEEE Date Engineering Bulletin, Volume 23, Number 3, pp 25-32, 2000.
[19] Laura Mannisto, Tim Kelly and Ben Petrazzinni. Internet and Global Information Infrastructure in Africa. ITU. 1998.
[20] MetaCrawler. http://www.metacrawler.com/
[21] Micro.com Systems Limited. http://www.micro.com.ng/
[22] Northern Light. http://www.northernlight.com/
[23] Greg Notess. Search Engines Statistics: Database Overlap. Available at http://www.searchengineshowdown.com/stats/overlap.shtml. February, 2000.
[24] Panos Briefing. The Internet and Poverty: Real Help or Real Hype? Panos Briefing No. 28. 1998.
[25] Jon Peha. Lessons from Haiti's Internet Development. Communications of the ACM. vol 42, no. 6; 1999.
[26] Ben Petrazzini and Mugo Kibati. The Internet in Developing Countries. Communications of the ACM. vol 42, no. 6; 1999.
[27] Janelle Prevost. A Reliable Low-Bandwidth Email-Based Communication Protocol, Master's Thesis, Massachusetts Institute of Technology, 2001.
[28] The Economist. International Internet Bandwidth. November 3, 2001. Reported from Packet Geography 2000.
[29] UN Development Programme (UNDP). Human Development Report. New York: Oxford University Press, 1999.
[30] Vivisimo Clustering Engine. http://www.vivisimo.com.
[31] Web for Mail. http://www4mail.org/
[32] Web2Mail. http://www.web2mail.com/