Knowledge Acquisition Processes in Internet Communities

Lee Li-Jen Chen and Brian R. Gaines
Knowledge Science Institute
University of Calgary
Alberta, Canada T2N 1N4
{lchen, gaines}@cpsc.ucalgary.ca

Abstract

With the growth of usage of List Severs and the World Wide Web the Internet has become a major resource for the acquisition of knowledge, and it has given new prominence to human discourse as a continuing source of knowledge. The society of distributed intelligent agents that is the Internet community at large provides an 'expert system' with a scope and scale well beyond that yet conceivable with computer-based systems alone. It is important to model and support the processes by which knowledge is acquired through the net. In developing new support tools is one asks "what is the starting point for the person seeking information, the existing information that is the basis for their search." A support tool is then one that takes that existing information and uses it to present further information that is likely to be relevant. Such information may include relevant concepts, text, existing documents, people, sites, list servers, news groups, and so on. The support system may provide links to further examples of all of these based on content, categorization or linguistic or logical inference. The outcome of the search may be access to a document but it may also be email to a person, a list or a news group. This articles develops a model of services and knowledge processes on the Internet, describes various forms of support tool, and categorizes them in terms of the model.

1 Introduction

The growth of the Internet has provided major new channels for the dissemination of knowledge. Increasing international connectivity has made the net accessible to special-interest communities world wide, and electronic mail and list servers now provide a major communications medium supporting discourse in these communities. Until recent years, limitations on the presentation quality of on-line file formats restricted the publication capabilities of the net to rapid dissemination of files printable in paper form. However, advances in on-line presentation capabilities now allow high-quality typographic documents with embedded figures and hyperlinks to be created, distributed and read on-line. Moreover, it has become possible to issue *active documents* containing animations, simulations, and supporting user interaction with computer services through the document interface. The major part of this functionality has become accessible through the protocols of the World Wide Web, and the web itself is seen as a precursor to an *information highway* subsuming all existing communications media.

The development of the net has been very rapid with little central planning, and, despite its widespread use, there is little information as yet on the social dynamics of net technologies. Many systems have been developed cope with the information overload generated by direct access to the net. The wide variety of indexing and search tools now available have in common that they support selective attention and awareness in the communities using the net. It would be useful to be able to analyze the design issues and principles involved in these tools in terms of the knowledge and discourse processes in the communities using these tools.

This article provides a model of the Internet in terms of discourse and awareness and uses it to classify the types of support tools existing and required.

2 Computer-Mediated Communication (CMC)

It is tempting to consider the Internet as a new publication medium in which electronic documents emulate paper ones, and where the basic human factors issues are those of indexing and information retrieval. This makes the vast existing literature on information retrieval, its techniques and human factors, relevant to the net. However, this addresses only one aspect of computer-mediated communication, neglecting its function of supporting discourse within communities. Much of the information retrieved from the net is generated as needed through discourse on list servers—the Internet is a mixed community of publications and intelligent human agents that both stores knowledge and generates it on demand. When the information needed cannot be found through retrieval then it may be requested through discourse, a phenomenon prophesied in the early days of timeshared computing:

> *"No company offering time-shared computer services has yet taken advantage of the communion possible between all users of the machine...If fifty percent of the world's population are connected through terminals, then questions from one location may be answered not by access to an internal data-base but by routing them to users elsewhere—who better to answer a question on abstruse Chinese history than an abstruse Chinese historian." (Gaines, 1971)*

The society of distributed intelligent agents that is the Internet community at large provides an 'expert system' with a scope and scale well beyond that yet conceivable with computer-based systems alone. Computer-based discovery, indexing and retrieval systems have a major role to play in that community, but are only one aspect of Internet information systems.

Krol (1993) captures the essence of these consideration in Internet RFC1462 which replies to the question "What is the Internet" with three definitions:

1 a network of networks based on the TCP/IP protocols,
2 a community of people who use and develop those networks,
3 a collection of resources that can be reached from those networks.

These are complementary perspectives on the net in terms of its technological infrastructure, its communities of users, and their access to resources, respectively. Models of computer-mediated communication must taken into account all three perspectives: how agents interface to the network; how discourse occurs within communities; and how resources are discovered and accessed.

3 Dimensions of the Computer-Mediated Communications

Figure 1 is a concept map presenting the major services on the net in terms of a small set of fundamental distinctions:-

- At the top level the major net services are characterized in terms of their utility for access to resources or awareness of resources.
- Access is sub-classified as to discourse, publications or services.
- Discourse is sub-classified by whether it is:-
  - — agent-to-agent discourse or community discourse;
  - — synchronous with the agents conversing in real time or asynchronous with substantial time delays in responses.
- Asynchronous community discourse is sub-classified by whether the channel is slow or fast, and whether the community is centrally registered or not.
- Publications are sub-classified by whether they are:-
  - — just fetched or presented when fetched;
  - — text or rich media.
- Services are sub-classified by whether they are text or rich media.

- Resource awareness is sub-classified by whether it is:
  - by resource name or content;
  - by keywords or by change in contents;
  - by keywords generated manually or automatically.

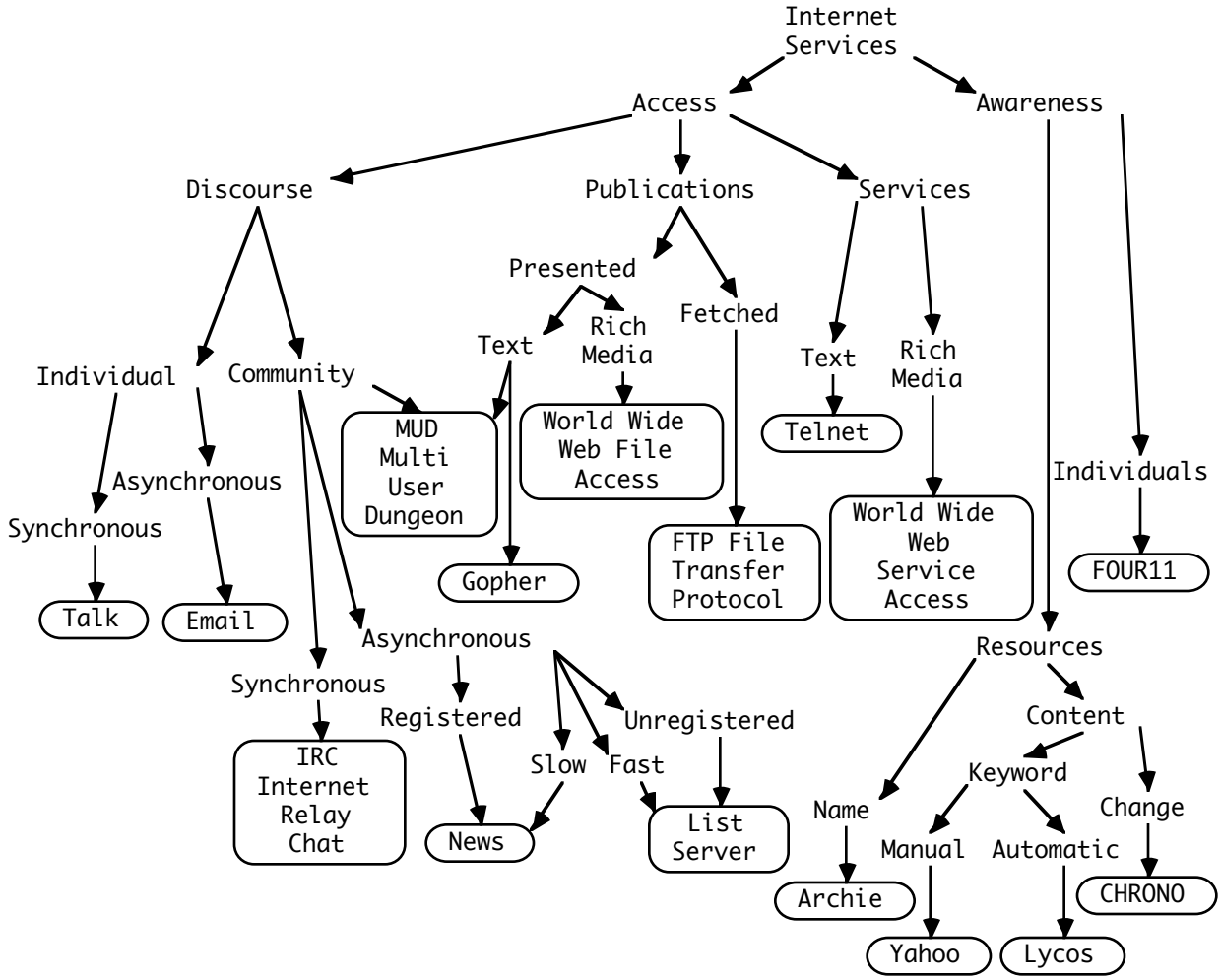

Figure 1 Internet services in terms of dimensions of computer-mediated communication

The less well-know systems classified are FOUR11 which provides an index of email addresses, and CHRONO (Chen, 1995) which indexes a web site in reverse chronological order to provide an automatic "what's new" page. MUDs, multi-user dungeons/dimensions, are interesting in providing a mix of services supporting both discourse and resource access. Web browsers such as Netscape are interesting in providing a single tool accessing nearly all the services shown except talk and chat.

4 A Punctuated Discourse Model of Computer-Mediated Communications

Figure 1 presents a conventional model of Internet services in terms of their utility, but it does not provide an integrative model of the way in which they support communities. Such a model can be developed by noting that what distinguishes discourse from publication is that in discourse it is expected that the recipient responds to the originator, whereas publication is generally a one-way communication. However, on list servers some material is published in that the originator expects no specific response, and material published in electronic journals or archives often evokes a response. Computer-mediated communication offers a very flexible

medium that breaks down the conventions of other media. The following diagrams show the different characteristics of the main Internet services in terms of these issues.

Figure 2 shows email discourse as a cycle of origination and response between a pair of agents communicating through a computer-mediated channel.

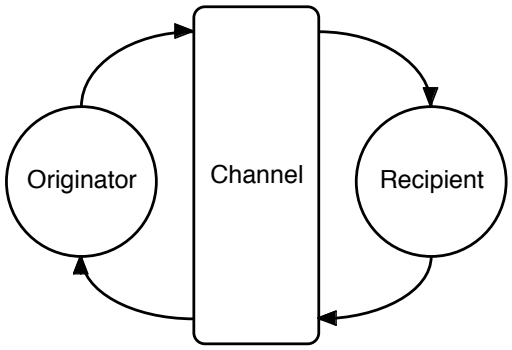

Figure 2 Email discourse

Figure 3 extends Figure 2 to show list server discourse as a cycle of origination and response between agents that is shared with a community through a computer-mediated channel. The community involvement leads to more complex discourse patterns in that: the originator may not direct the message to a particular recipient; there may be multiple responses to a message; and the response from the recipient may itself trigger responses from others who did not originate the discourse. For a particular discourse sequence this leads to a natural division of the community into active participants who respond and passive participants who do not.

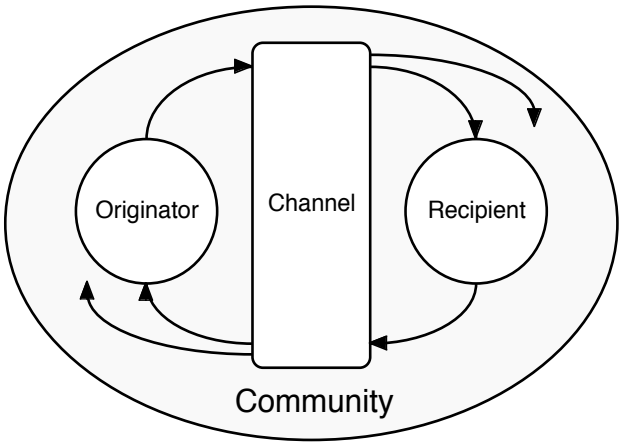

Figure 3 List server discourse

Figure 4 modifies Figure 3 to show web publication as an activity in which the channel is buffered to act as a store also. The material published is available to a community and the originator is unlikely to target it on a particular recipient. Recipients are not expected to respond direct to the originator, but responses may occur through email, list servers or through the publication of material linked to the original. Because the published material is not automatically distributed to a list, recipients have to actively search for and discover the material.

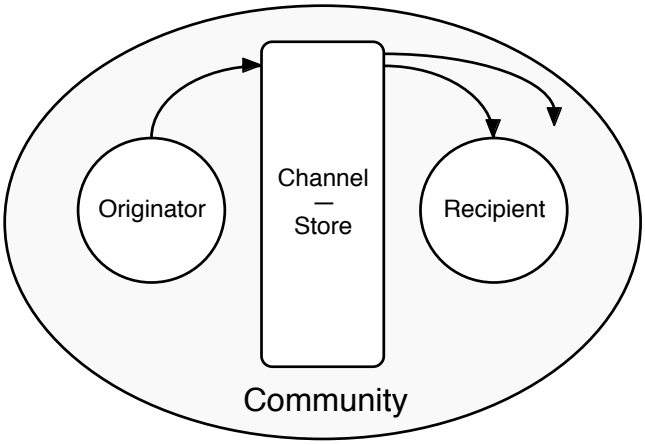

Figure 4 World Wide Web publication

The common structure adopted for the diagrams is intended to draw attention to the commonalities between the services. List server discourse is usually archived and often converted to hypermail on the web. Web publications do trigger responses through other services or through links on the web. A search on the web may not discover a specific item but rather a related item on a newsgroup, list or by an author, and result in an request for information to the newsgroup, list or author. Individuals and communities use many of the available Internet services in an integrated way to support their knowledge processes.

Figure 5 subsumes Figures 2 through 4 to provide an integrated model of Internet knowledge processes that captures all the issues discussed. It models the processes as discourse punctuated by the intervention of a store allowing an indefinite time delay between the emission of a message and its receipt. It introduces two major dimensions of analysis: the *times* for each step in a discourse cycle; and the *awareness* by originators of recipients and vice versa.

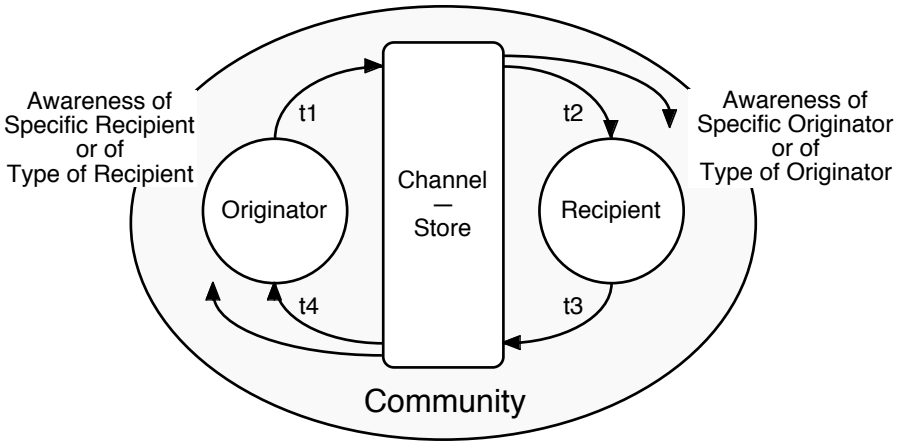

Figure 5 Punctuated discourse

4.1 Time Structure of Punctuated Discourse

The four times shown in Figure 5 are:

t1: the origination time—the time from a concept to its expression and availability
t2: the discovery time—the time from availability to receipt
t3: the response time—the time from receipt to expression and availability of a response
t4: the response discovery time—the time from response availability to receipt

Note that agent processing times and channel delays have been lumped. A study focusing on the impact of communication delays would want to consider them separately, otherwise there is no significant distinction—a general principle might be that communication delays should not be greater than agent processing times. Note also that the diagram is to a large extent symmetrical—the recipient becomes an originator when responding.

An important overall parameter is the round-trip discourse time, t1+t2+t3+t4. If this is small, a few seconds or less, we talk in terms of synchronous communication. If its is large, a few hours or more, we talk in terms of asynchronous communication. If it is infinite, so that there is no response, we talk in terms of publication. However, this analysis shows that there is a continuous spectrum from synchronous through asynchronous to publication.

The discovery times, t2 and t4, are very significant to publication-mode discourse, and attempts to reduce them have lead to a wide range of awareness-support tools that aid potential recipients to discover relevant material and originators to make material easier to discover.

4.2 Awareness Structure of Punctuated Discourse

One can regard a community as a set of agents that provide resources to one other with the most significant dimension relating to the coordination of the community being that of the *awareness* of who is providing a particular resource and who is using it. In the tightly-coupled team, each person is usually aware of who will provide a particular resource and often of when they will provide it. In logical terms, this can be termed *extensional awareness* because the specific resource and provider are known, as contrasted to *intensional awareness* in which only the characteristics of suitable resources or providers are known.

In a special interest community resource providers usually do not have such extensional awareness of the resource users, and, if they do, can be regarded as forming teams operating within the community. Instead, resource providers usually have an *intensional awareness* of the resource users in terms of their characteristics as *types* of user within the community. The classification of users into types usually corresponds to social norms within the community, such as the ethical responsibilities in a professional community to communicate certain forms of information to appropriate members of the community. Resource users in a special interest community may have an extensional awareness of particular resources or resource providers, or an intensional awareness of the types of resource provider likely to provide the resources they require. This asymmetry between providers and users characterizes a special interest community and also leads to differentiation of the community in terms of core members of whom many users are extensionally aware, and sub-communities specializing in particular forms of resource.

In the community of Internet users at large, there is little awareness of particular resources or providers and only a general awareness of the rich set of resources is available. Awareness of the characteristics of resources and providers is vague, corresponding to *weak intensional awareness*.

These distinctions are summarized in Figure 6 and it is clear that the classification of awareness can lead to a richer taxonomy of communities than the 3-way division defined. Analysis of awareness in these terms allows the structure of a community to be specified in operational terms, and in complex communities there will be complex structures of awareness. The coarse divisions into sub-teams and sub-special interest communities provides a way of reducing this complexity in modeling the community.

Locus of responsibility	Team	Special-Interest Community	Community at Large
Originator	*Extensional awareness of actual recipients.* Use email to notify. Use CHRONO to index.	*Intensional awareness of types of recipient.* Broadcast to list server. Establish HTML links. Use CHRONO to index.	*No awareness of recipients, or only weak intensional awareness of types of recipients.* Broadcast to news groups. Register in Yahoo. Initialize Lycos.
Recipient	*Extensional awareness of actual resources and originators.* Use email to inquire. Check CHRONO index.	*Extensional awareness of actual resources and originators, or intensional awareness of types of resources and originators.* Subscribe to list server. Follow HTML links. Check CHRONO index. Use WebWatch, Katipo or URL-Minder	*No awareness of resources or originators, or only weak intensional awareness of types of resources and originators.* Read news groups. Browse Yahoo. Search with Lycos. Search with MetaCrawler.

Figure 6 Communities and tools distinguished in terms of awareness

The differentiation of communities in terms of awareness draws attention to the significance of supporting various aspects of awareness in a CMC system. *Resource awareness*, the awareness that specific resources or resources with specified characteristics exists, may be supported by various indexing and search procedures. However, there is also a need to support *chronological awareness*, the awareness that a resource has changed or come into existence. Figure 6 also shows the way in which current tools for awareness support are classified within this framework.

5 CHRONO: Chronological Awareness Support Tools

CHRONO is an HTTPD server-side systsem which generates chronological listings of Web pages that have been changed recently at specific sites. It provides a basic awareness-support that let of a Web site visitors (e.g., members of a group, an organization, or other Netsurfers) see which Web pages have been modified since their last visit. Currently, the CHRONO system is implemented on a UNIX platform and has been made widely available for use at other sites (see http://www.cpsc.ucalgary.ca/~lchen/cpsc.html#chrono). As shown in Figure 7, CHRONO presents to the visitors an HTML document that lists the titles of Web pages at the site in reverse chronological order. This chronological listing of Web pages also functions as a collection of hyperlinks to the listed pages.

This time-line dimension allows frequent visitors of a Web site an immediate awareness on what have been changed since their latest visit. The changes they see may be some Web pages in which they have particular prior-interests of or may be some pages that they have never seen before but now appeal to them. Hence this *chronological browsing* characteristic is analogue to *spatial (subject-category) browsing* characteristic that library patrons have often experienced when looking for books on open book-shelves (i.e., accidentally finding (more) relevant books near by the books that they are looking for originally).

Figure 7 CHRONO in use at a PC user group site

What is different here is that instead of finding relevant information via browsing the near by subject-categories, now the users may find relevant information via browsing the concurrently modified/created web pages. Sometimes, conceptually related documents are created (or

modified) around the same time, however their author(s) may not remember to update the HTML links to them. Unlike a manually updated *what's new* page in which the users have to rely on the timely updates made by a Webmaster (or by the document authors), CHRONO provides the time-line dimension to the users automatically, in a reliable, periodic fashion.

WebWatch (Specter, 1995), *Katipo* (Newberry, 1995), and *URL-Minder* (NetMind, 1995) are other chronological awareness tools that track changes in specified documents. WebWatch is a client-side chronological awareness system for keeping track of changes in selected Web documents. Given an HTML document referencing URLs on the Web, it produces a filtered list, containing only those URLs that have been modified since a given time. Katipo is another client-side chronological awareness system built for Macintosh that shares many similar concepts as WebWatch. It reads through the Global History file maintained by some Web browsers checking for documents that have changed since the last time a user viewed them. It writes a report file (in HTML format) listing all such documents in a format that allows you to easily visit the updated documents. URL-Minder is a centralized system that keeps track of resources on the Web, and sends registered users e-mail whenever their personally registered resources change. Web users can have the URL-minder keep track of any Web resource accessible via HTTP. It can be anything, not just Web pages that users personally maintain.

6 Further Developments

The CHRONO research program is now at a stage of measuring the structures and time constants of discourse on the Internet from empirical data. Studies are being carried out of the rates of diffusion of information and the various knowledge acquisition paths and processes wherby individuals become aware of information on the net. List server archives are being analyzed to determine the fine structure of discourse and to track the trajectories of ideas.

CHRONO was issued in May 1996 and is now being used at a number of sites. A new program META-CHRONO is under development which will collect and collate information from multiple sites running CHRONO and provide awareness of activities being carried out on a distributed basis.

It is expected that more details of these further developments will be available for reporting in the final version of this article.

7 Conclusions

The purpose of the research reported in this article has been to develop a finer-grained model of the knowledge acquisition processes that occur in Internet communities in order to support and improve those processes through new and better services. The model developed suggests three levels of analysis of services:

- *Message quality*—the improvement of the multimedia capabilities of the basic message channel—there are been continuous improvement from simple text to typography, images, movies, sounds, animations, simulations, and so on.
- *Relationship modeling*—the incorporation of linkage information preserving discourse relationships—the hypertext links of the original web technology introduced this capability and clickable maps extended it—there is scope for further extension based on greater understanding of the roles the links play in enable people to grasp the argument forms of information on the web.
- *Awareness support*—the systematic reduction of the time (t2 and t4 in Figure 5) for a potential recipient to become aware of relevant information—manual and automatic indexing and various forms of search engine have made massive advances in coping with the information overload resulting from the growth of the web—however, there is scope for many different tools supporting the various ways in which people manage their awareness.

The key question to ask in developing new awareness support tools is “what is the starting point for the person seeking information, the existing information that is the basis for their search.” A support tool is then one that takes that existing information and uses it to present further information that is likely to be relevant. Such information may include relevant concepts, text, existing documents, people, sites, list servers, news groups, and so on. The support system may provide links to further examples of all of these based on content, categorization or linguistic or logical inference. The outcome of the search may be access to a document but it may also be email to a person, a list or a news group.

The net is a vehicle for discourse in which the goals of individual agents are supported through social knowledge processes, and support tool design needs to be based on increasingly refined models of those processes. Much of our current research is concerned with the empirical studies of discourse processes on the net through analysis of information diffusion, list server archives, and so on. We conjecture that tools that develop models of such processes and make them available to the participants may themselves result in improved usage of net resources.

Acknowledgments

Financial assistance for this work has been made available by the Natural Sciences and Engineering Research Council of Canada.

References

Chen, L.L.-J. (1995). CHRONO: A Chronological Awareness Tool. Knowledge Science Institute, University of Calgary. http://www.cpsc.ucalgary.ca:80/~lchen/cpsc.html#chrono.

Gaines, B.R. (1971). Through a teleprinter darkly. **Behavioural Technology 1**(2) 15-16.

Krol, E. (1993). FYI on "What is the Internet?". Internet. RFC 1462.

NetMind (1995). The URL-Minder: Your Own Personal Web Robot. NetMind. http://www.netmind.com/URL-minder/URL-minder.html.

Newberry, M. (1995). Katipo—a Web Lurker. Victoria University of Wellington, New Zealand. http://www.vuw.ac.nz/~newbery/Katipo.html.

Specter (1995). WebWatch. Specter Communications. http://www.specter.com/.