Persistence of Web References in Scientific Research

The lack of persistence of Web references has called into question the increasingly common practice of citing URLs in scientific papers. Although few critical resources have been lost to date, new strategies to manage Internet resources and improved citation practices are necessary to minimize the future loss of information.

Steve Lawrence
David M. Pennock
Gary William Flake
Robert Krovetz
NEC Research Institute

Frans M. Coetzee
Certus International Inc.

Eric Glover
University of Michigan

Finn Årup Nielsen
Technical University of Denmark

Andries Kruger
University of Stellenbosch

C. Lee Giles
The Pennsylvania State University

Researchers have long desired immediate access to all scientific knowledge. Although there are still major hurdles to overcome,[1] the Internet has brought this goal closer to reality. Scientists use the Internet to communicate their findings to a broader audience than ever before, and formal references to information on the Web are increasingly common. However, the lack of reliable or stable Internet publishing sources sometimes offsets the advantage of being able to easily share various materials at minimal cost. Individuals and organizations abandon Web pages, shut down servers, and rename files.

Invalid URLs do more than contribute to user annoyance and frustration. Ultimately, they can lead to the loss of important data as cited works and research findings gradually disappear from circulation. Proposed formal approaches to bypass pages with dead links during browsing[2] or reduce their ranking when presenting search engine results[3] threaten to exacerbate the problem. The lack of persistence of Web references has led many researchers to question whether articles and other published works should continue to include URL citations.

We analyzed references to Web resources in numerous computer science publications, considering the volume of citations, validity of links, and detailed nature of invalid links. We found that URL citations have increased dramatically in recent years and that many of these references are now invalid. At the same time, we determined that most missing URLs are easy to relocate. Although formal references to published articles are always preferable, we believe that Web references facilitate scientific communication and progress. However, new Internet resource management strategies and improved citation practices are necessary to minimize future information loss.

SEARCHING FOR THE MISSING LINKS

We investigated URLs cited in research papers using NEC Research Institute's scientific digital library ResearchIndex, formerly known as CiteSeer.[4,5] This database, created in 1997, indexes Postscript and PDF research articles on the Web. Aimed at improving communication and progress in science, ResearchIndex incorporates Autonomous Citation Indexing as well as the ability to quickly and easily see the context of subsequent papers in which authors refer to a given article of interest. A free service is available at http://researchindex.org/.

Search methodology

From 3 to 5 May 2000, we analyzed 270,977 computer science journal papers, conference papers, and technical reports that were available at that time on the publicly indexable Web.[1] From the 100,826 articles cited by another article in the database (thus providing us with the year of publication), we extracted 67,577 URLs. We then attempted to access each one, following redirected URLs to their new destination. We searched for strings starting with "http:", "https:", or "ftp:" and, after removal of trailing punctuation, ending with a quote or white space.

As Figure 1a shows, the number of URL citations has increased substantially since the inception of the Web. Figure 1b dramatically illustrates the lack of persistence of Internet resources. The percentage of invalid links in the articles we examined varied from

0018-9162/01/$10.00 © 2001 IEEE

23 percent in 1999 to a peak of 53 percent in 1994.

For a random sample of 300 invalid URLs, we attempted to find the new location of the page cited or, failing that, highly related information. Of these broken links, 32 percent were either extracted incorrectly from the papers, contained a syntax error such that they could never be valid, or were example links that we believe were never intended to be valid. Extraction errors were typically due to the Postscript/PDF-to-text conversion program not converting special characters correctly or inserting spaces within the URLs (our extraction routine corrects for some, but not all, easily identifiable cases). We removed these URLs from the data set; the percentages reported are for the remaining URLs.

Search results

Figure 2a shows a breakdown of the remaining invalid URLs. An initial search by five researchers found the new location of the page or highly related information 80 percent of the time. They relocated 11 percent of the broken links by guessing an alternate URL or browsing the Web, and 44 percent using a search engine. For 25 percent of the URLs, they found highly related information likely to be a good substitute for the original page (without access to the original page we cannot guarantee the success of the match). For 6 percent of the invalid URLs, they could not find the new location although a formal citation did accompany it. They were unable to find the remaining 14 percent of missing links. A single second searcher was able to locate 80 percent of the

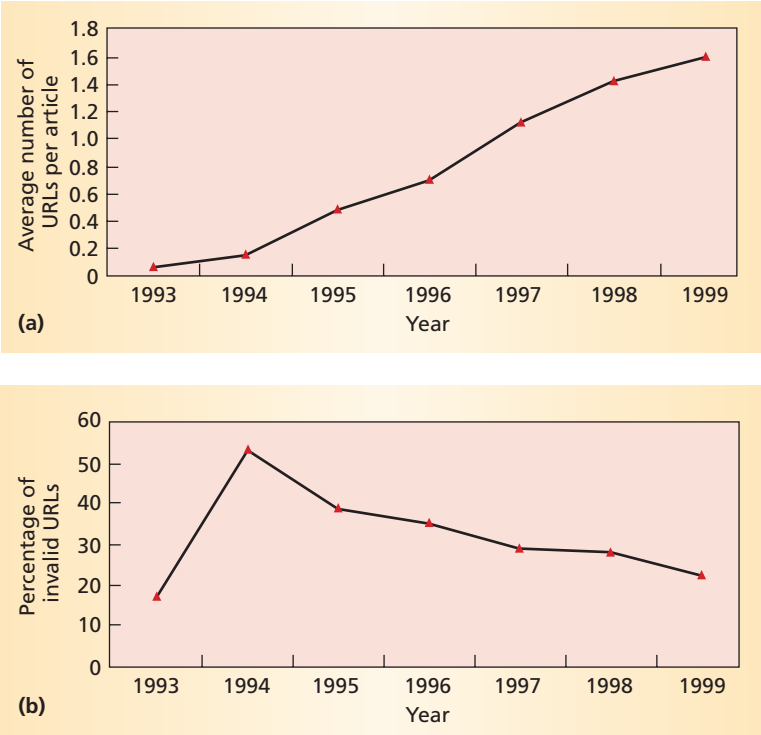

Figure 1. (a) The average number of URLs contained in the articles versus the year of publication. (b) The percentage of invalid links contained in articles versus the year of publication, corrected for incorrectly extracted URLs. The reason for the lower percentage of invalid URLs in 1993 may be that a greater percentage of references at this early stage of the Web were to relatively well-known homepages for companies or organizations—for example, http://www.intel.com/.

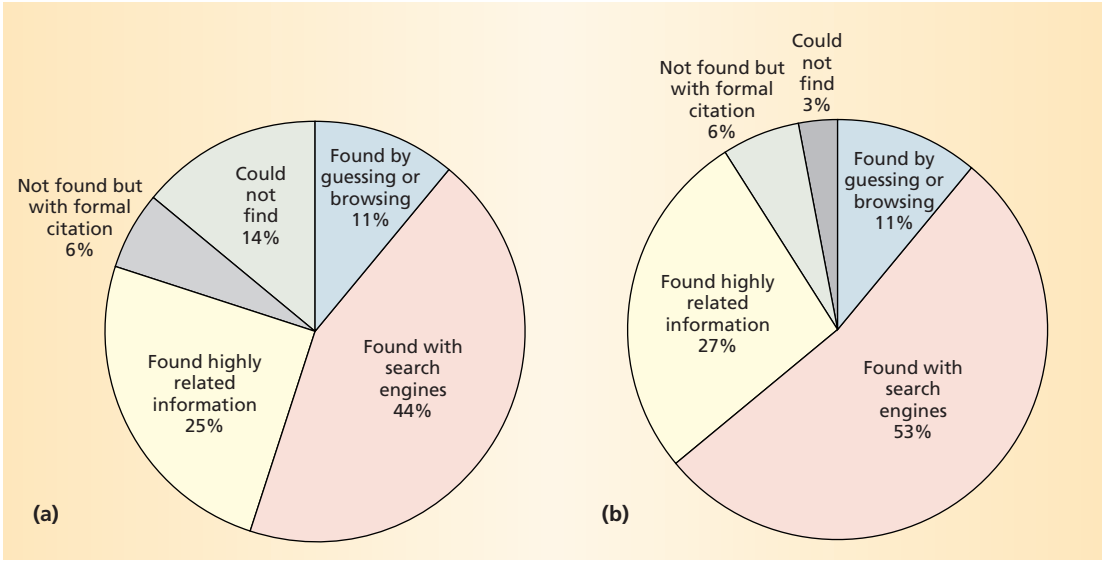

Figure 2. (a) Initial classification of a random sample of 300 invalid URLs. In many cases, it was possible to find the new location or highly related information. (b) Classification after a second search. A second searcher found most of the URLs that the first searchers could not find.

Relocating URLs

Individual searchers in our study used different search engines to relocate missing URLs or find related information. The most commonly used were Google, ResearchIndex, and NECI's Inquirus,[1] a metasearch engine that combines the results of several search engines. Because search engines index unique sets of Web pages, combining results can significantly improve Web coverage.[2] Other search engines used included AltaVista and Northern Light.

Numerous techniques exist to relocate information on the Web. The easiest is to search for a document by title or author, if they are known. It is also often useful to search for the title as a phrase. Examining a citation's context can generate queries on project, company, or institution names. Browsing from the top-level page, researcher homepage, or another alternative starting point may help relocate pages that may have moved within a site. It is also possible to guess possible new locations—for example, if you suspect that a project moved from a specific machine to its own domain or that an individual's homepage changed to a standard notation such as http://www.x.com/~user/. Some sites have their own search engines that can be useful. If a URL has a relatively unique component, you can search for that as well.

References

1. S. Lawrence and C.L. Giles, "Context and Page Analysis for Improved Web Search," *IEEE Internet Computing,* July/Aug. 1998, pp. 38-46.
2. S. Lawrence and C.L. Giles, "Accessibility of Information on the Web," *Nature,* 8 July 1999, pp. 107-109.

URLs reported as lost. As Figure 2b shows, this lowered the overall percentage of lost URLs to only 3 percent.

The searchers spent no more than five minutes looking for each moved or related URL. They achieved mixed results, with the most successful searchers locating all invalid URLs investigated and the least successful unable to locate 16 percent of them. This variation was due to differing search experience and abilities, degrees of persistence, and opinions regarding whether or not information was highly related. More of the broken links may have been locatable if additional time was spent searching, the searchers had more search experience, or if better search tools were available (see the sidebar, "Relocating URLs").

Classifying lost URLs

The searchers estimated the difficulty of relocating or finding substitute information for the invalid URLs using the following classes: *easy, somewhat difficult,* and *very difficult*. Figure 3a shows the percentage of lost URLs in each class. Most missing links were easy to relocate.

For each invalid URL that we could not locate, we examined the citation's context in the respective paper. We estimated the URL's importance in relation to the ability of future researchers to verify or build on the paper using the following classes: *not very important, somewhat important,* and *very important*. We classified half of the URLs as not very important, 41 percent as somewhat important, and only 9 percent as very important. As Figure 3b shows, no very important lost URLs remained after the second search.

CAUSES OF INVALID URLS

Through our manual analysis of missing links, we have identified several reasons why URLs become invalid. First, personal homepages tend to disappear when researchers move. Second, many who restructure Web sites fail to maintain old links. These problems are likely to persist without improved citation practices.

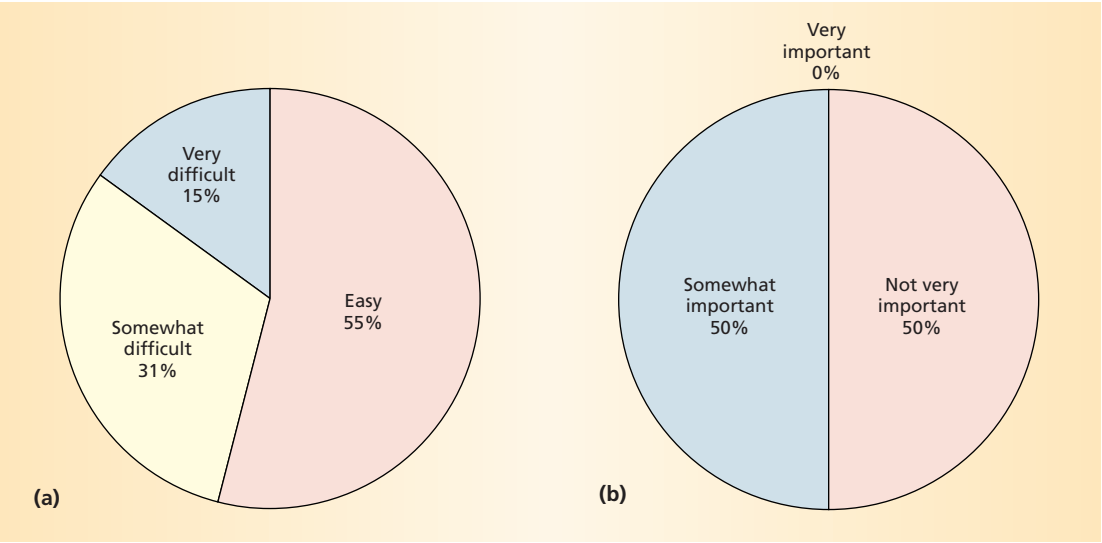

Figure 3. (a) Classification of the difficulty locating invalid URLs. It was easy to find the new location of most invalid links. (b) Importance of lost URLs. None of the lost links remaining after a second search were very important to the ability of future researchers to verify or build on the paper containing the citation.

Preserving Web Resources Using URNs

A major limitation of URLs as locators is that they tie Internet resources to their current network location and file path. When the resource moves, the URL breaks. To address the need for a persistent, location-independent identifier, the Internet Engineering Task Force (http://ietf.org/) developed the Uniform Resource Name (URN) specification.[1] Rather than identify where a resource resides at some instant, a URN identifies the resource itself; the name stays the same even when the location of the resource moves.

Two general-purpose systems available to Web content managers to assign, manage, and resolve URNs are Persistent Uniform Resource Locator (PURL) (http://www.purl.org/),[2] designed by the Online Computer Library Center, and Handle (http://handle.net/),[3] created by the Corporation for National Research Initiatives. Both systems use intermediate resolution servers to track the movement of digital content and other resources on the Internet. When a URN's location changes, the resolver is updated so that the URN resolves to the new location. Additional features of these systems include authentication and support for multiple instances of a resource.

Neither PURL nor Handle, however, presents an ideal solution. Handle requires installation of name resolution software, which few users currently have, to resolve persistent identifiers or "handles." PURL avoids the need for software support but is not fully location independent. An individual PURL includes the location-dependent address of a PURL resolver—for example, the PURL http://purl.org/metadata/dublin_core contains the address of the main PURL resolver, http://purl.org/. Using PURLs thus relies on the continued existence of a particular PURL resolver, as well as the continued provision of adequate response time by the resolver. Proxy servers are available that allow handles to be used in a similar way to PURLs. An additional weakness of both systems is that they require someone to maintain the validity of resources.

With more than half a million PURLs registered, PURL is more popular and currently preferable to Handle because of the latter's requirement for software support. In terms of total Web citations, however, adoption of PURLs remains poor. Of the 67,577 URLs we extracted from papers in the ResearchIndex database, we only located 11 PURLs (0.016 percent of URLs), all to the same resource—http://purl.org/metadata/dublin_core. Despite the obvious motivation of early users, some PURLs are already invalid. A search for url:purl.oclc.org using AltaVista, for example, turned up many PURLs that returned a page stating, "The requested PURL has been deactivated and cannot be resolved."

Still, we recommend using PURLs if long-term persistence is the goal, users are willing to maintain the validity of the redirection, and response time is not critical.

References

1. K. Sollins and L. Masinter, "Functional Requirements for Uniform Resource Names," Internet Request for Comments, RFC1737, Dec. 1994, http://ietf.org/rfc/rfc1737.txt.
2. K. Shafer et al., "Introduction to Persistent Uniform Resource Locators," *Proc INET 96 Conf.*, Internet Society, Reston, Va., 1996.
3. W. Arms, C. Blanchi, and E. Overly, "An Architecture for Information in Digital Libraries," *D-Lib Magazine,* Feb. 1997, http://www.dlib.org/dlib/february97/cnri/02arms1.html.

Other problems, attributable to the initial rapid growth and evolution of the Web, will likely disappear over time. Web pioneers ran their own servers on personal machines, and the links they created were lost with the disconnection of machines, relocation of servers, or change of server names. Today, however, there are more widespread conventions for setting up servers, such as using HTTP instead of FTP, using the standard communications port for HTTP servers (80), and standardizing URLs for homepages—for example, http://www.x.com/~user/. The easy availability of domain names has accelerated the movement of software and other projects from personal repositories to more stable, dedicated sites maintained by universities and corporations.

LONG-TERM SOLUTIONS

Authors who take the Web designer's perspective propose link management techniques to resolve the problem of invalid URLs.[6] Others advocate augmenting existing Internet protocols to improve link persistence. For example, David Ingham, Steven Caughey, and Mark Little[7] suggest developing an object-oriented network parallel to the Web that enforces referential integrity and performs garbage collection.

Alternatives to the Web such as the Hyper-G[8] and Xanadu[9] hypertext systems contain built-in mechanisms for enforcing link consistency. However, these are not in wide usage. In fact, the Web's success may partly lie in the relatively few requirements imposed on authors. A system that includes features such as enforced link consistency may increase overhead and complexity for the system and users, thereby limiting participation and success.

Another interesting idea is to replace HTML with protocols that provide support for improved content-based indexing and retrieval. In principle, search engines can use content summarization and indexing to recognize materials that move on the Web. Thomas

Even when a Web citation's location is stable, its contents can change so that subsequent readers may not view the exact same cited material.

Phelps and Robert Wilensky[10] have shown that a small set of words can serve as unique identifiers for most Internet documents. These words can be used to augment URLs and thereby help locate mobile documents.

Substituting persistent, location-independent Uniform Resource Names (URNs) for URLs is a promising solution (as shown in the sidebar, "Preserving Web Resources Using URNs"). However, to date relatively few researchers have adopted this syntax in their Web citations. Various public systems exist to name and manage URNs such as the Persistent Uniform Resource Locator (PURL) system and the Handle system, but they require human maintenance and, in some cases, retrofitting Internet software.

Even when a Web citation's location is stable, its contents can change so that subsequent readers may not view the exact same cited material. One way to address this issue is to periodically archive the entire Web, as Brewster Kahle does in his Internet Archive (http://www.archive.org/). However, the efficacy of this approach is unclear. Alexa Internet (http://www.alexa.com/), which creates Internet navigation software and studies trends in Web content and behavior, estimates that Web pages disappear after an average time of only 75 days. Further, many pages can change during the large amount of time required to take a snapshot of the Web.

Another related issue is that the software or hardware required to read specific data formats may become unavailable or difficult to access.

Other efforts to improve Web permanence include NECI's Intermemory project,[11] which aims to create highly survivable and available storage systems using widely distributed processors that, individually, may be unreliable and untrustworthy (http://www.intermemory.org/). Stanford University's LOCKSS (Lots of Copies Keeps Stuff Safe)[12] system is an effort by multiple libraries to work together to redundantly cache copies of specific documents (http://lockss.stanford.edu/).

IMPROVING CITATION PRACTICES

All of these approaches can at best redirect attention to material that has moved but not disappeared. Although few critical resources cited in computer science articles appear to have been lost to date, we believe that improving citation practices is necessary to minimize future information loss. Based on our experiences labeling missing URLs, we recommend several citation practices to make it easier for future readers to relocate information that may move:

- Provide formal citations along with URL citations whenever possible, but include valuable URL citations even when formal citations are unavailable. The loss of *some* links over time is preferable to depriving readers of *all* links. Even when formal citations are available, providing an accompanying URL can significantly improve the information's accessibility.
- Provide enough context information so that readers can pose adequate search engine queries to track down invalid links. For example, when giving the URL for a preprint, provide the document's full title along with full details of the authors—instead of, say, using "et al." We found many examples of URLs for which the contents could not be inferred from the context.

In the case of URLs that cite repositories controlled by the author:

- Place materials in a reliable central repository such as a preprint or software archive. This is particularly important for links to complete versions of papers, omitted proofs, and supporting data or results.
- Name the repository and include the name in citations. This name is then available for later searches. For software distributions, include a file with the name of the software package; some search engines may index the filename. Provide a documented homepage for software and establish a domain name.
- When referencing software or software manuals, reference a URL for the entire project rather than URLs for specific software or manual versions. Version files frequently become unavailable after updating of the software or manual.
- Avoid URLs that depend on a personal directory, specific machine, or subnet name.

We believe that it is impractical in the long term to expect individuals or small organizations to provide persistent access to online resources. Such material will probably ultimately move or disappear. The general problem of persistence and disappearance requires a combination of technical solutions and peer policies. Professional societies such as the IEEE and the ACM, and funding agencies such as the NSF, could help by proposing and enforcing acceptable standards for citations. Ideally, all cited materials—but especially those important to building on or verifying research—would be available from a stable site mirrored worldwide such as the Netlib Repository (http://www.netlib.org/). These societies and agencies could sponsor or host preprint and software repositories, or at least request that authors use appropriate repositories when possible. ✷

References

1. S. Lawrence and C.L. Giles, "Accessibility of Information on the Web," *Nature,* 8 July 1999, pp. 107-109.
2. P. De Bra, G-J. Houben, and Y. Kornatzky, "A Formal Approach to Analyzing the Browsing Semantics of Hypertext," *Proc. Computing Science in the Netherlands (CSN94),* Stichting Mathematisch Centrum, Amsterdam, Nov. 1994, pp. 78-89.
3. J. Shavlik and T. Eliassi-Rad, "Intelligent Agents for Web-Based Tasks: An Advice-Taking Approach," *AAAI/ICML Workshop on Learning for Text Categorization,* AAAI Press, Menlo Park, Calif., 1998, pp. 63-70.
4. S. Lawrence, K. Bollacker, and C.L. Giles, "Indexing and Retrieval of Scientific Literature," *8th Int'l Conf. Information and Knowledge Management (CIKM 99),* ACM Press, New York, Nov. 1999, pp. 139-146.
5. S. Lawrence, C.L. Giles, and K. Bollacker, "Digital Libraries and Autonomous Citation Indexing," *Computer,* June 1999, pp. 67-71.
6. M.L. Creech, "Author-Oriented Link Management," *Computer Networks and ISDN Systems*, May 1996, pp. 1015-1025.
7. D. Ingham, S. Caughey, and M. Little, "Fixing the 'Broken-Link' Problem: The W3Objects Approach," *Computer Networks and ISDN Systems*, May 1996, pp. 1255-1268.
8. F. Kappe, K. Andrews, and H. Maurer, "The Hyper-G Network Information System," *J. Universal Computer Science*, Apr. 1995, pp. 206-220.
9. T. Nelson, *Literary Machines*, Mindful Press, Sausalito, Calif., 1993.
10. T. Phelps and R. Wilensky, "Robust Hyperlinks Cost Just Five Words Each," tech. report UCB//CSD-00-1091, University of California, Berkeley, Jan. 2000.
11. A. Goldberg and P.N. Yianilos, "Towards an Archival Intermemory," *Proc. IEEE Advances Digital Libraries (ADL 98),* IEEE CS Press, Los Alamitos, Calif., Apr. 1998, pp. 147-156.
12. D.S.H. Rosenthal and V. Reich, "Permanent Web Publishing," *Freenix Track, Usenix Ann. Technical Conf.*, Usenix, Berkeley, Calif., June 2000, pp. 129-140.

Steve Lawrence *is a research scientist at NEC Research Institute. He received a PhD in computer science from the University of Queensland, Australia. His research interests include information retrieval and machine learning. He is a member of the AAAI, the AAAS, the ACM, and the IEEE. Contact him at lawrence@research.nj.nec.com.*

Frans M. Coetzee *received a PhD in electrical engineering from Carnegie Mellon University. His research interests include computer learning and signal processing. He is a member of the ACM and the IEEE. Contact him at fmcoetzee@bigfoot.com.*

Eric Glover *is a PhD student at the University of Michigan. He received an MS in electrical engineering from the University of Michigan. His research interests include Web information retrieval. He is a member of the AAAI, the ACM, and the IEEE. Contact him at compuman@eecs.umich.edu.*

David M. Pennock *is a research scientist at NEC Research Institute. He received a PhD in computer science from the University of Michigan. His research interests include electronic commerce and artificial intelligence. He is a member of the AAAI, the ACM, the IEEE, and INFORMS. Contact him at dpennock@research.nj.nec.com.*

Gary William Flake *is a research scientist at NEC Research Institute. He received a PhD in computer science from the University of Maryland. His research interests include machine learning, data mining, and complex systems. He is a member of the ACM and the IEEE. Contact him at flake@research.nj.nec.com.*

Finn Årup Nielsen *is a PhD student at the Technical University of Denmark, where he received an MS in engineering. His research interests include neuroinformatics and analysis of functional neuroimages. Contact him at fn@imm.dtu.dk.*

Robert Krovetz *is a scientist at NEC Research Institute. He received a PhD in computer science from the University of Massachusetts, Amherst. His research interests include computational linguistics, intelligent information retrieval, and artificial intelligence and law. He is a member of the ACL, the ACM, and the ASIST. Contact him at krovetz@research.nj.nec.com.*

Andries Kruger *is an MSc student at the University of Stellenbosch, South Africa. His research interests include special-purpose search engines and associated technologies, including information extraction, focused crawling, document classification, and indexing. Contact him at akruger@cs.sun.ac.za.*

C. Lee Giles *is the David Reese Professor of Information Sciences and Technology, Professor of Computer Science and Engineering, and Associate Director of Research at the eBusiness Research Center at Pennsylvania State University, and a consulting scientist at NEC Research Institute. He received a PhD in optical sciences from the University of Arizona. His research interests include e-commerce and intelligent information processing, retrieval, and management. He is a member of the ACM, the AAAS, the IEEE, and the INNS. Contact him at giles@ist.psu.edu.*