Identifying Reusable Resources in Digital Reference Responses

Jeffrey Pomerantz

***Jeffrey Pomerantz** is Associate Professor, School of Library and Information Science, University of North Carolina at Chapel Hill.*

Submitted for review September 21, 2009; revised and accepted for publication September 22, 2010.

Are the resources provided in answers to reference questions reusable for answering future reference questions? This study seeks to answer this question as a means to address the scalability problem of human-mediated reference work. Using the Internet Public Library's archive of over eighty thousand records of answered reference questions, this study identifies (1) what resources are provided in responses to digital reference questions, (2) the extent to which these resources are reusable in future responses, and (3) the useful lifespan of a resource that has been provided. The distribution of resources provided in these answer records was found to display a classic power law distribution. The half-life of these resources was found to be approximately eleven years, far longer than the half-lives of resources in other corpora that have been studied. The relevance of these resources was found to be remarkably high, even after more than a decade.

Ever since the advent of digital reference services, there has existed a belief that, as Coffman puts it, "if we could somehow access the work another librarian had done before, there would be no need to start over answering every question from scratch."[1] This is a seductive notion, as it offers at least a partial solution to a perennial problem of digital reference work and perhaps all reference work: scalability.[2] Human labor is time consuming and expensive, and significant savings of both could be realized if even some of the products of that labor could be reused.

On the other hand, some researchers suggest that the context that gives rise to an information need is unique for every individual.[3] A consequence of this position is that an answer that is useful to an individual in a particular context and the set of information resources provided to support that answer will not be useful to others in other contexts. If this is indeed the case, then even similarly phrased questions cannot be treated as actually being similar. It may therefore be misguided for reference services to attempt to reuse answers across ostensibly similar questions.

This is a critical issue, both practically and theoretically. As a practical matter for reference services, if answers are indeed reusable across questions, then Coffman is correct, and there is no need for reference librarians to answer every question from scratch. On the other hand, if answers are not reusable across questions, then this dramatically limits the scalability of reference services. As a theoretical matter, if answers are reusable across questions, then even

Reference & User Services Quarterly, vol. 50, no. 4, pp. 366–78

though the individual contexts out of which information needs arise may be unique, commonalities exist in how those information needs may be fulfilled. Despite this being an issue central to reference work, however, there has been no research to date on the reusability of answers provided by reference services.

The resources used in answers to reference questions are a different matter. Resources are of course reused all the time. The reference section of a library is full of materials that are used again and again. It's well known that dictionaries, encyclopedias, and telephone directories are among the most commonly used reference sources in any library—and, at the risk of being tautological, a source does not become commonly used unless it is reused. Indeed, a library's entire collection is reused, though of course some materials more than others.[4] The argument could be made that reuse of materials is one of the *raisons d'etre* of libraries. Ranganathan said it best: Books are for use.

This paper reports on a study to investigate the reusability of the information resources in answers provided by digital reference services. Specifically, the information resources investigated in this study are the URLs provided in answers from a digital reference service. The research questions guiding this project were

1. To what extent are the URLs provided in responses reusable for future responses?
2. What is the useful lifespan of URLs provided in responses to digital reference questions?

In order to answer these questions, it is necessary to have a corpus of answered questions to work with. This is, of course, why no research has been conducted on the reusability of reference answers: these corpora either do not exist or are unavailable to researchers. Questions and answers are entirely transitory at reference desks; there is a long tradition of capturing data about the interaction at reference desks, but the interaction itself is rarely, if ever, captured.[5] The entire interaction is captured by digital reference services—e-mail exchanges, instant messaging transcripts, etc.—but due to significant user privacy concerns, those artifacts are rarely made available outside of the service itself.[6]

BACKGROUND ON THE INTERNET PUBLIC LIBRARY

There are, however, two large corpora of answered questions available outside of the services that answered the questions: the QuestionPoint Global Knowledge Base, and the Internet Public Library's Archive of Reference Questions (ARQ). QuestionPoint is the most widely-used digital reference management system (www.oclc.org/questionpoint), and the Global Knowledge Base is the repository of answered questions submitted by libraries around the world that use QuestionPoint. As of February 2009, the Global Knowledge Base contained 20,061 searchable records.[7] The Global Knowledge Base is searchable by library users, though it is not clear how many libraries have implemented this feature.[8] The Global Knowledge Base is not, however, readily accessible for research, while the IPL's ARQ is, at least to the current author.

As of 2006, the IPL is partly supported by the memberships of information and library science programs in the United States and around the world. For their membership fees, these programs receive access to the ARQ. The ARQ contains a record of every question answered by the IPL, going back to the IPL's inception in 1995. This study found that as of June 2010, the ARQ contains 81,385 records, thus making it several times larger than the QuestionPoint Global Knowledge Base, even assuming growth of the Global Knowledge Base since February 2009. The author's institution, the School of Information and Library Science at the University of North Carolina at Chapel Hill is also an IPL member, thus making the ARQ readily accessible to the author.

The Internet Public Library (ipl.org) describes itself as "a public service organization and a learning/teaching environment."[9] The IPL provides two primary services: collections of vetted and annotated resources, and an Ask a Librarian service. The IPL maintains resource collections on a wide range of topics, several special collections (e.g., on topics such as the U.S. Presidents and the fifty states), and collections for young children and teenagers. These are not collections of content developed by the IPL, but rather collections of links to materials developed by others, on the free web, that have been vetted by the IPL and judged to be authoritative and trustworthy. For example, the IPL's Earth Sciences category contains resources from NASA's National Space Science Data Center, the U.S. Geological Survey, and the British Geological Survey, among many others.

The Ask an IPL Librarian service maintains a question submission webform (ipl.org/div/askus) that allows users to submit a question on any topic. Submitted questions enter a queue, and answerers may claim the unanswered question of their choice. IPL policy states that once an answerer claims a question, she should submit a response to the user within twenty-four hours.[10] IPL policy also states that answers should contain two to four sources.[11] Once an answerer provides a response to a question, the response is sent as an e-mail to the user, and the question-and-answer records are stored in the IPL's content management system, QRC.[12] Every six months or so, new question-answer records are deidentified and made available in the ARQ. The IPL has a small full-time staff, and much of the work of vetting resources and question answering is conducted by volunteers and students in information and library science programs, in particular in courses on collection development and reference.

In January 2010, the IPL merged with the Librarians' Internet Index (LII), and was renamed ipl2. The LII maintained extensive collections of vetted resources, similar to those maintained by the IPL. The LII did not, however, maintain a question answering service. While the data collected for this study spans the time period of the transition from the IPL to the ipl2, the Ask an IPL Librarian service was largely unaffected by this transition.

LITERATURE REVIEW

In the quote at the beginning of this paper, Coffman suggests that it would be useful to reuse the answers provided by librarians in response to reference questions.[13] Indeed, he seems to suggest that the only hurdle to this reuse is the technical capability to mine these answer corpora.

This is not, however, a widely accepted position in the reference community. Indeed, judging by the number of digital reference services that do in fact reuse answers provided by librarians, this position has been rejected wholesale by the reference community. This rejection of resource mining may be a principled stance by the reference community, a belief that information needs are subjective and unique to individuals.[14] If this is the case, then this principled stance is held despite the repeat nature of many questions submitted to reference services.[15] The rejection of resource mining may, however, simply be a technical limitation: even the most sophisticated question answering systems are not yet capable of answering many questions submitted to a reference service, which may be on any subject, ambiguous, poorly phrased, and the thousand shocks that natural language is heir to.[16]

There is, in fact, only one digital reference service of which the author is aware that reuses answers in response to new questions: the Mad Scientist Network (MADSci, www.madsci.org). Bry describes the process employed by the MADSci Ask-A-Scientist service: when the user submits a question, a CGI script searches the MADSci archive of previously answered questions.[17] Bry states that "approximately 63 percent of questions are matched with archived files"—however, "only 25 percent of users deem their questions answered by this process (15 percent of all submitted questions)."[18] While the MadSci Ask-A-Scientist service is still operational, Bry's article is a decade old now, and it is not clear how or if the algorithms used by the service have changed over that time. In the interim, other forms of question-answering systems and services have emerged.

Perhaps the most significant alternative to digital reference services to emerge in the past few years is social Q&A sites. Examples abound: Yahoo! Answers (answers.yahoo.com), Ask MetaFilter (ask.metafilter.com), WikiAnswers (wiki.answers.com), the Wikipedia Reference Desk (en.wikipedia.org/wiki/Wikipedia:Reference_desk), the now-defunct Google Answers (answers.google.com), Aardvark (vark.com, acquired by Google in February 2010), to name only a few. Due to the different mechanisms and policies according to which these sites operate, it is difficult to define precisely what social Q&A is. The most basic characteristics that all of these sites possess in common are (1) a mechanism for users to submit questions in natural language, (2) a mechanism for users to respond to submitted questions, and (3) a community built around participation in this question answering.[19] There is a great deal of variability in the types of questions asked and the quality of the answers provided on social Q&A sites; so much so in fact that some librarians maintain that social Q&A and library reference services are not even offering the same service.[20] Whether or not these two types of services are in competition will not be addressed here. The important point for present purposes is that not only the resources provided in answers, but the entire content of previous answers, are reused in answers to new questions on social Q&A sites: of Yahoo! Answers responses that refer to online sources, fully 59 percent of those online sources are previous Yahoo! Answers responses.[21]

A less recent but equally significant alternative form of question answering evolved from the information retrieval community. The Text Retrieval Conference (TREC) was instrumental in promoting the development of question-answering (QA) systems, by hosting a Question Answering Track for nine years, 1999–2007. The goal of the Question Answering Track was to develop systems "to retrieve small snippets of text that contain the actual answer to a question rather than the document lists traditionally returned by text retrieval systems."[22] These systems were also built to answer open-domain questions: in other words, questions could be on any subject, and the corpus of documents from which answers were retrieved was a newspaper collection, which could support answering on any subject. A persistent criticism of these systems, however, was that they were able to answer only "factoid" questions—such as, "Who is the conductor of the Boston Pops?"[23] In other words, these were systems designed to answer ready reference questions. One of the long-standing desiderata for question answering systems was expert-level answering of expert-level questions.[24] That was never achieved by any QA system developed for TREC, perhaps due to the difficulty of interpreting questions and formulating answers at an expert level in an open domain. At the other end of the spectrum, however, expert-level QA systems have been developed with considerable success in restricted domains, where it is feasible to develop domain-specific rules for identifying answers in texts.[25]

Out of this work on developing restricted-domain QA systems has come some research on enabling QA systems to reuse information provided in previous answers.[26] One of the categories of reuse articulated in that study is the reuse of one document to answer more than one question. This type of reuse was found to be ubiquitous in the corpus of questions used by Light et al. This finding should come as a surprise to no one; the existence and popularity of frequently-asked question (FAQ) lists attests to the usefulness of a single document in the answer to multiple questions. Indeed, a FAQ list operates as a single document that answers multiple questions in two ways: first, multiple answers are provided in a single FAQ document, and second, a single answer may be a response to multiple information needs that can be reformulated as a particular question in a FAQ list.

The popularity and usefulness of FAQs provides evidence that one person can find the answer to someone else's question useful. Almost since the advent of the web, researchers have worked on developing QA systems that make use of FAQ lists as corpora. One of the first of these was FAQ-Finder, which, even in the early days of its development, showed considerable success in retrieving QA pairs that satisfactorily answered new questions put to the system.[27] More recent work has demonstrated similar results, both in restricted and, like FAQ-Finder, in open domains.[28]

While FAQ lists are often created in response to actual frequently-asked questions, as opposed to lists of questions being created to proactively respond to anticipated questions, they are also often cleaned-up versions of those questions. Most social Q&A sites, on the other hand, make the entire content of previous questions and answers available online. Thus, unlike in a FAQ list, where multiple real questions may be reformulated into a single FAQ, in social Q&A sites, every question, no matter how similar, is retained and made available as written. Building on methods derived from FAQ-based QA systems, considerable success has been reported in answering new questions by making use of the corpus of answered questions on social Q&A sites.[29] This work provides good reason to believe that the answers provided by digital reference questions may likewise be reusable.

METHOD

This project used data from the IPL's ARQ. The ARQ contains all answered questions submitted to the Ask an IPL Librarian reference service from the inception of the IPL in September 1995. These question-answer records contain fielded data as an artifact of the IPL's Ask an IPL Librarian question submission webform. Some of these fields are closed-ended, such as the drop-down list of subjects by which the user can categorize her question. Other fields are open ended, such as the field in which the user specifies her question. The records in the ARQ contain all of the data from the webform, the response provided by the IPL answerer, and system-level administrative data supplied by the QRC (e.g., timestamps for when the question was submitted, when an answerer claimed the question, and when the answer was sent).[30] The IPL has developed an algorithm to deidentify these records to a claimed level of 90–95 percent accuracy; this algorithm is run on all records prior to their being made available in the archive.[31]

The author wrote a simple web crawler and in June 2010 downloaded every record in the ARQ. The earliest records in the archive were answered in August 1995, the latest in June 2009. The most recent data was a year old at the time of this data collection because the IPL deidentifies question-answer records every six months, and then there is a lag before they are made available in the archive; the latest batch was made available shortly before this data collection.[32]

The author also developed a parser to tokenize question-answer records. In other words, records were analyzed automatically, and certain blocks of text were identified within each record. The following blocks of text were identified in this way: the text blocks that are the question and the answer, URLs provided within answers, subject categories, and timestamps. The IPL has a policy dictating that for any URL over sixty-five characters in length, an additional shortened link be provided using TinyURL; all TinyURLs in answers were therefore eliminated as duplicates of other URLs.[33]

URLs are relatively easy to identify using regular expressions. A regular expression is a search string that matches a specific pattern of text: as a simple example, "lib" matches the patterns "library," "librarian," "ad-lib," "calibrate," and of course many others. There are several regular expressions that one can find on the web to match URLs. This ease of identifying URLs is due to their inherent structure and syntax: all contain a top-level domain (e.g., .edu), most contain a protocol prefix (e.g., http://), many contain the character / in the middle, most end with a / or with a three- or four-letter suffix (e.g., .html, .pdf, .asp). Even many mistyped URLs may be identified using a regular expression that employs these syntactical rules. Other types of resources, such as books and journal articles, are also provided in IPL responses but far less frequently than URLs. This is the result of another IPL policy dictating that answerers should prioritize providing "freely-available sources" online to ensure "that all our patrons will have access to the information," though a related policy states that print and subscription sources may be included in answers under appropriate circumstances.[34] Future work will be required to develop regular expressions to identify types of resources other than URLs in answers.

Domains were also extracted from the full URLs. By *domain* we mean both the domain and what in other contexts would be considered a subdomain. These are defined as follows: "A domain is a subdomain of another domain if it is contained within that domain. This relationship can be tested by seeing if the subdomain's name ends with the containing domain's name. For example, A.B.C.D is a subdomain of B.C.D."[35] Here, domains were identified by simply extracting the part of the URL between the protocol prefix (e.g, http://, https://) and the next /, thus giving URLs of the form www.ipl.org, www.nces.ed.gov, etc. Common subdomains such as www were then removed from the lefthand side of the domain. We used a fairly loose definition of domain to accommodate the wide variation in the naming of domains and subdomains: for example, the URLs www.nces.ed.gov and nces.ed.gov are equivalent, but nces.ed.gov and ed.gov are not.

This study then proceeded in two phases. First, link checking was used to ensure that the URLs in question-answer records were still extant. Second, human judgement was used to evaluate the relevance of extant URLs.

Link Checking

For the first phase of this study, the author wrote a script to check every URL provided in the responses. The specifications for the Hypertext Transfer Protocol (HTTP) define a set of three-digit status codes that correspond to the possible results of an HTTP request; the best known of these is probably the code 404, which corresponds to not found.[36] There are forty-one status codes defined for the HTTP/1.1 protocol in the following five categories:

- 1xx: Informational—The request was received, continuing process.
- 2xx: Success—The action was successfully received, understood, and accepted.
- 3xx: Redirection—Further action must be taken in order to complete the request.
- 4xx: Client Error—The request contains bad syntax or cannot be fulfilled.
- 5xx: Server Error—The server failed to fulfill an apparently valid request.

URLs in responses were not corrected before they were checked; that is, if a URL was mistyped in the IPL response, it was checked in its incorrect state. For example, if the URL http://ipl.or appeared in an answer, it was left as is, even though it is perfectly clear what the error is that would cause this URL to return a 404 error. This was done because the erroneous URL was what the IPL user would have seen upon receiving the answer. Many, perhaps even most, users would know how to correct this error, but some might not. Furthermore, this study investigates the reusability of URLs provided in responses, not of those URLs as interpreted by users.

Human Relevance Judgment

For the second phase of this study, the author created a Human Intelligence Task (HIT) on Amazon Mechanical Turk (www.mturk.com). Amazon describes Mechanical Turk as "a marketplace for work that requires human intelligence."[37] This work—the HIT—may be almost anything: some HITs are as simple as tagging images, while some are as complex as transcribing recorded interviews.[38] A requester creates a HIT and sets a price that workers will be paid for completing one iteration of the HIT. Mechanical Turk is, as might be expected, increasingly being used in the social sciences, as it enables researchers to obtain certain types of data and to conduct certain types of analyses quickly and inexpensively.[39]

The HIT for this study presented Mechanical Turk workers with a deidentified question that had been submitted to the IPL, and a single webpage that had been provided in the answer to that question. These webpages were the URLs provided in answers, embedded into HITs using the iframe HTML element. Thus, each HIT included a single question and a single webpage. Multiple URLs are provided in most answers in the ARQ, so the same question may have appeared in multiple HITs. Likewise, the same webpage may have appeared in multiple HITs if the same URL was provided in multiple answers. The set of URLs from the ARQ were stratified by month, and 5 percent of URLs from each month were randomly sampled, for inclusion in the HIT.

Only URLs that returned a 2xx or 3xx status from link checking were included in HITs; in other words, only URLs that successfully resolved. A 4xx or 5xx status indicates that the webpage at a URL was not found, for whatever reason. There would therefore be no point in providing such a URL in a HIT, since, if a URL failed to resolve for our link checking script, it is unlikely that it would resolve for a Turk worker.

No URLs of search engines or pages of search engine results were provided in HITs, and likewise no URLs of IPL collections were provided. One of the research questions for this study was to determine how reusable the resources provided in digital reference answers are. While search engine results and IPL collections may be considered to be resources of a sort, they are collections of links, and therefore require the user to make decisions about which of the resources linked to are relevant. One of the selling points, so to speak, of reference services is that the librarian serves as a filter for the user, selecting relevant resources when the user may not have the knowledge or skills to do so for herself. Directing a user to a page of links—such as a search engine results list or an IPL collection—therefore runs somewhat counter to this function of a reference service.

The HIT instructed workers to evaluate the relevance of the webpage to the question. A truly enormous number of studies exist in which subjects make relevance assessments of documents, and the methodologies for making these assessments are by no means perfectly consistent across studies. Indeed, there is some disagreement as to what the phenomenon of relevance even entails.[40] Because of this variability, the definition of relevance used for the HIT was derived from the instructions provided to searchers by Saracevic et al. in their study of information seeking and retrieval behavior: a document is considered relevant if "the information it conveys is considered to be related to your question, even if the information is outdated or already familiar to you."[41] This definition of relevance was used for two reasons. First, many studies of relevance are based on this work by Saracevic et al., so using this definition aligns the current study with much prior work.[42] Second, Saracevic et al.'s instructions are simple and clear enough to convey the complex idea of relevance to Turk workers—individuals who cannot reasonably be expected to be familiar with the subtleties of the concept of relevance.

The HIT provided the following mutually exclusive options for Turk workers: relevant, partly relevant, not relevant, and broken link. This item was required; a worker could not submit the HIT until this field was filled out. As discussed above, only URLs that successfully resolved from link checking were included in HITs; despite this, the option broken link was provided, since it was possible that a URL could be offline for any number of reasons when the worker attempted to view it, even if it resolved successfully only weeks before.

Mechanical Turk allows a HIT to be assigned multiple times; each HIT here was assigned three times. Thus, each question-webpage combination was seen and evaluated for relevance by three workers. In order to ensure that the worker was an actual human and not a bot, we added a final question to the HIT: "Dear human, X plus Y equals what?", where X and Y were randomly generated numbers between one and three, and provided a text field for the answer. Human workers, of course, had no trouble answering this correctly, while bots supplied nonsensical responses that were easily filtered out. Mechanical Turk provides a mechanism for rejecting responses; responses that were suspected to be from bots were rejected, and republished for other workers to complete.

RESULTS

URLs Provided in Responses

The web crawler downloaded a total of 81,385 question-answer records from the ARQ, spanning nearly fourteen years. This means that the IPL answers an average of 487 questions per month. An average value is slightly misleading, however, since the IPL, like all reference services, is subject to fluctuations in the volume of questions received over time, corresponding to the academic year. The range in the IPL's volume of questions is 13 in December 1996 to 1,182 in November 2004.

The parser identified 364,906 URLs provided in all responses in the archive. On average, each response contains 4.7 URLs. The IPL has a policy dictating that every response should contain at least three sources, so this number is not surprising, though it is gratifying that the IPL's answerers expend more than the minimum required effort in finding and providing sources, at least on average.[43] What is surprising is how wide the range is in the number of URLs provided in response to one question: a minimum of 1, and a maximum of 77, with a median of 4. To be fair, however, these 77 URLs were actually in 2 responses, the first to the user's original question and the second to a follow-up question from the same user—though this entire thread was contained in a single record in the ARQ.

The figure of 364,906 URLs is actually an overestimate. The Ask an IPL Librarian service sends the response to users as an e-mail. Most e-mail clients automatically include the text of the original e-mail in a reply. Thus, if the user replies to the IPL, the full text of the IPL's response will be included in the reply—including all URLs—unless the user deliberately deletes them. Others studying the IPL have found that between 15 and 20 percent of users reply to the IPL with unsolicited thanks.[44] This finding indicates that users do in fact reply to the IPL fairly frequently—a behavior which, while otherwise innocuous, inflates this count of URLs in the ARQ. Future work will be required to develop a more sophisticated parser that can identify duplicated text in the question-answer record, given the variability in how both e-mail clients and authors of e-mails handle replies (e.g., the reply goes above or below the original message, or is interleaved with it).

These 364,906 URLs were from a total of 74,454 domains. Not surprisingly, URLs from the domain ipl.org are provided most often, in 46 percent of all answers (n=37,488). Next is google.com (19 percent, n=15,174), followed by en.wikipedia.org (the English-language version of Wikipedia), infoplease.com, and amazon.com at 6 percent or less each (n=4,851; 3,456; and 3,306; respectively). It is interesting that Wikipedia appears in so many answers, as the IPL has a policy discouraging answerers from using Wikipedia.[45]

Fully 57 percent of domains appear in only one IPL response (n=42,242), and 95 percent of domains appear in ten or fewer responses (n=70,923). These figures mean that each domain in the ARQ is provided in a response, on average, less than once per year. Again, however, this average value is misleading, since the domain ipl.org being provided in 46 percent of all answers means that it is provided in several responses per day, while some domains were only ever provided once. The distribution of domains in the ARQ is in fact a power law distribution with an extremely long tail; so long, in fact, that it would be impossible to represent legibly in a figure here (indeed, the author and colleagues originally presented this data in a six-foot wide poster and still were able to legibly display only every five-hundreth domain).

On the other hand, some domains have been provided by answerers throughout the entire lifespan of the IPL, including:

www.ipl.org
www.findlaw.com
www.ala.org
www.nytimes.com
www.loc.gov
www.aclu.org
www.oclc.org
www.uspto.gov
www.yahoo.com
us.imdb.com

The longest lifespan for a URL (that is, the timespan between the first and most recent use of this URL in the ARQ) ending with a directory name is nine years, seven months, for The Children's Literature Web Guide (www.ucalgary.ca/~dkbrown). The longest lifespan for a URL ending with a filename is eight years, eight months, for The Complete Works of William Shakespeare (the-tech.mit.edu:80/Shakespeare/works.html).

Link Checking

As discussed above, every URL provided in responses in the ARQ was checked, and the three-digit HTTP status code returned. A total of forty-one status codes are defined, but for this analysis, returned status codes were grouped by class. The distinction between some of the codes within classes is subtle and frankly not relevant to this analysis. It does not matter, for example, if a URL requires authentication in the form of a username and password supplied by the user (401), or via a proxy (407); the consequence is that the user does not have access to the resource at the URL.

The percentage of URLs in the ARQ that returned the following status codes was as follows:

- 1xx (Informational): 0% of all URLs (n=0)
- 2xx (Success): 69.3% (n=266,636)
- 3xx (Redirection): 0.7% (n=2,781)
- 4xx (Client Error): 23.0% (n=88,551)
- 5xx (Server Error): 7.0% (n=26,939)

Link checking gives a slightly conservative estimate of the number of URLs that are valid. The web being the dynamic place that it is, there are many reasons that a URL may be inaccessible: a server may be down for maintenance, a webpage may have been moved and no redirect created, a website may have been "slashdotted."[46] Any of these and more reasons for a URL returning a 4xx or 5xx status code may have occurred at the time that this link checking was running. In order to determine how much of an overestimate the 30 percent of status codes that returned as errors was, all URLs that returned a 4xx or 5xx status code were rechecked a week later. As it turned out, it was only a very small overestimate: only 1.7 percent (n=1,913) of the URLs that returned a 4xx or 5xx status code on the first round of checking returned 2xx or 3xx on the second round. Thus, at any given time, approximately one-third of all URLs in the ARQ are "dead." Of course, the percentage of dead URLs varies with time: the farther back in time the URL was provided in an answer, the more likely it was that the URL returned an error status, as can be seen in figure 1.

It must be noted, however, that the fact that a URL is valid does not mean that the content of the webpage at that URL is the same now as it was when it was provided in a response by an IPL answerer. The content of many, perhaps most, webpages changes over time. Indeed, it is even possible that the content of a webpage may have changed so much that it is no longer relevant to the original question. This issue is what motivated the relevance assessment component of this study, which will be discussed further below.

Figure 1 shows the percentages of URLs in the ARQ that resolve to the various categories of error codes. Figure 1 is a stacked line graph, so that percentages are cumulative: the percentage of status codes that resolved as 5xx is not 100 percent, for example; rather, the percentage of status codes that resolved as 5xx is represented by the gap between the 4xx and the 5xx lines.

In figure 1, error codes are collapsed into three groups. The 2xx and 3xx statuses were combined since so few 3xx status codes returned in link checking. While, from the user's point of view, it does not matter where an error occurred—on the client or the server side—to produce a 4xx or 5xx status code, those codes are separated in figure 1 to illustrate the change over time in the difference between the relative percentages of the two status codes. This difference is especially noticeable prior to December 1999, when this difference exceeds 10 percent. In other words, the farther back in time one goes, the more server errors one gets for URLs. When put that way, this is an intuitive finding, since the farther back in time one goes, the more likely it is that a server will have gone offline.

Figure 1 also shows the half-life of various types of citations. Many studies have shown that URLs tend to decay over time, a phenomenon referred to as "link rot." Koehler defines the half-life of a web document as "that period of time required for half of a defined Web literature to disappear," in other words, the rate of link rot.[47] The curve in figure 1 shows a half-life of 2 years for documents on the free web, a rate that has been shown to be consistent across multiple studies.[48] The shaded area in figure 1 shows the range of half-lives for references cited in scholarly journals, both on the web and in print, in various disciplines: a range from 1.5 to 4 years.[49]

Note that the lines representing URL status codes stop short of the far right of the graph in figure 1. This is due to the fact that the most recent questions in the ARQ were from June 2009, but the half-lives of citations were calculated from the date when this analysis was conducted, in June 2010.

The line representing 2xx and 3xx status codes is above the half-life curves for both the free web and for scholarly citations for most of its length. While there is considerable fluctuation over time in the percentage of successfully resolved or redirected URLs, the half-life of IPL answers is approximately eleven years (the point at which this line drops below 50 percent for the first time is July 1999). The resources provided by IPL answerers thus have longer half-lives than either documents on the free web or those cited in scholarly journals. Put differently, IPL answerers are skillful at selecting and providing resources that will exist on the web over the long term. Thus, even though the collection of URLs in the ARQ has experienced link rot, it has been at a slow rate. It is therefore reasonable to expect that mining this collection would yield URLs that could be reused.

Relevance Assessment

As discussed above, all URLs that returned a 4xx or 5xx status code from link checking and all URLs from the IPL and from search engines were removed from the full set of 364,906 URLs identified in the ARQ. After these were removed, a total of 198,947 URLs remained, or 54.5 percent of the original set. These were stratified by month, and 5 percent of URLs from each month were sampled for inclusion in the Mechanical Turk HIT. Thus the

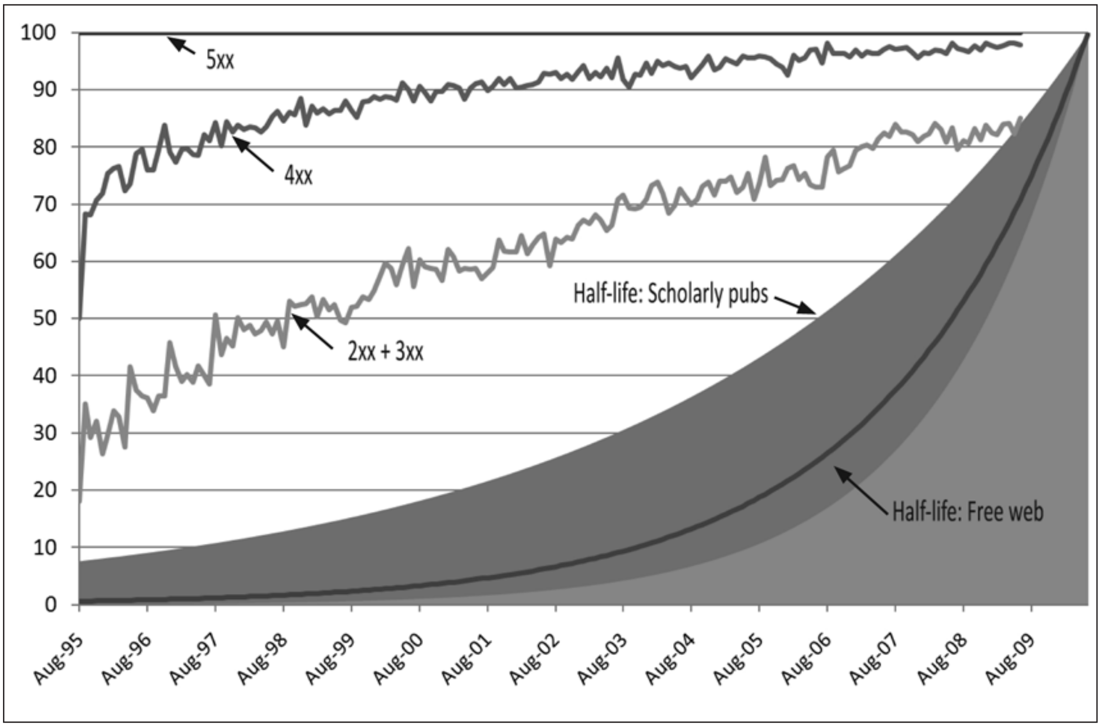

Figure 1. HTTP Status Codes for URLs in the ARQ by Month, and Decay Rate of Other Corpuses

HIT consisted of 9,947 URLs and their associated questions. Each HIT was iterated three times.

Across the entire set of 9,947 question-webpage combinations, 52 percent of those webpages (n = 5,172) were evaluated as relevant by all three workers. This average value is not terribly informative, however, since the range spans from 0 percent to 100 percent of webpages within a given month evaluated as relevant by all three workers. Agreement between all three workers may not be necessary or desirable, however. Other researchers using Mechanical Turk to assess relevance have found that the best results have been achieved when most but not all the workers agreed.[50] Therefore, figure 2 shows the distribution by month of webpages evaluated by all three workers as relevant (circles) and by two workers as relevant and one worker as partly relevant (diamonds).

There is considerable fluctuation over time in the percentage of websites evaluated as relevant or partly relevant. There is, however, fairly tight clustering around the mean values of these percentages. It was stated in the previous section that IPL answerers are skillful at selecting and providing resources that will exist on the web over the long term. This finding is evidence that IPL answerers are likewise skillful at selecting and providing resources that remain relevant to the question at hand over the long term.

DISCUSSION

IPL answerers are skillful at selecting and providing resources that remain extant on the web and that remain relevant over the long term. In fact, the author was surprised by how successful IPL answerers are at selecting such resources. The URLs provided by IPL answerers have considerably longer half-lives than documents in any other corpus that have been studied. Furthermore, of the documents that remain extant, a remarkably high percentage also remain relevant.

The second research question that this study aimed to answer was what is the useful lifespan of URLs provided in responses to digital reference questions? The answer to this question has two parts. First, the half-life of these URLs is approximately eleven years. But that number does not address the “useful” part of the question. For a URL to be useful, in the context of digital reference, it must be relevant for answering a question, and if a URL remains extant, the findings here indicate a high probability that it will also remain relevant—at least for answering the original question.

Of course, when investigating the reusability of URLs, a URL’s relevance to the original question is not truly the issue: it is far more important for a previously provided URL to be relevant to a future question. The first research question that this

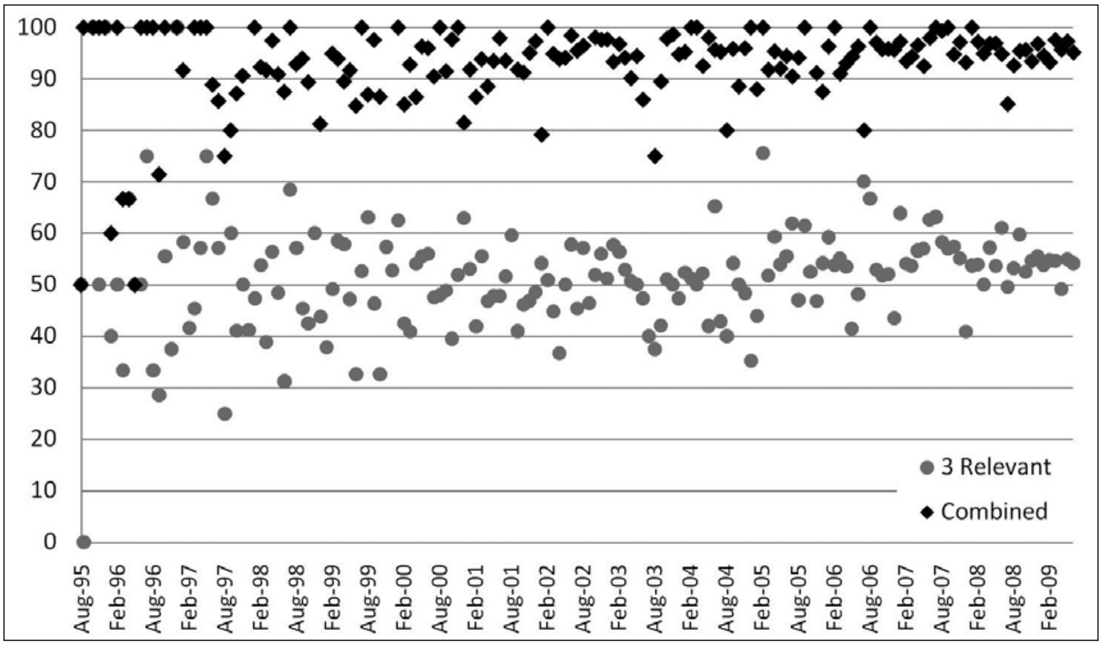

Figure 2. Percentage of URLs Judged Relevant by Month

study aimed to answer was to what extent are the URLs provided in responses reusable for future responses? This study was, unfortunately, only able to answer half of that question. As discussed at the start of this paper, it is an open question in reference work, whether entire answers are reusable. Are information needs truly unique? Is an answer, and the resources that support that answer, provided to a particular individual in a particular context useful to others in other contexts? Answering that question has as much to do with the individuals and the contexts as with the answer and the resources. The findings of this study indicate that the URLs provided in responses are likely to be reusable in response to future questions. There are, however, two areas in which research is needed to fill out this picture: determining the degree of similarity between ostensibly similar questions, and determining the relevance of a previously provided answer to a new question.

Measures of query similarity already exist. Query similarity is closely related to, and may be considered a subset of, document similarity, a concept which has existed in information retrieval for decades. A full treatment of document similarity is considerably beyond the scope of this paper: for a thorough treatment of measures of association, see van Rijsbergen.[51] In brief and vastly oversimplified, however, document similarity is often computed pairwise: the frequency of all terms (words and phrases) is identified in the documents, and an algorithm is computed over these frequencies. This algorithm provides a measure of semantic distance between the two documents: the closer two documents are, the more terms they have in common, and therefore the more similar. Query similarity can be computed using some of the same techniques, though not all; some document similarity measures do not work well for questions because questions are generally so much shorter than documents.[52]

To date, no work has been done to compute the similarity of questions from reference services. As discussed above, it has been difficult for researchers to gain access to corpora of reference questions, either because such corpora did not exist (as for desk reference services) or due to user privacy concerns (as for digital reference services). With the availability of the IPL's ARQ and the QuestionPoint Global Knowledge Base, however, this becomes possible. Future research on the similarity of questions asked of digital reference services over time is needed, in order to fill out the picture of whether previously provided answers are reusable for new questions. The hypothesis for such work would be that the more similar two questions are, the more relevant the answer to one would be to the other.

As discussed above, however, some answers on social Q&A sites are reused whole cloth.[53] Although we do not know what percentage of reused answers are voted by the asker or by the social Q&A community as "best answers," it is reasonable to assume that some are. An evaluation of best answer status is an explicit indication, especially if the evaluation is made by the asker, that a previously provided answer is relevant to answer a new question. Research on social Q&A answering and evaluation behavior thus has the potential to inform the behavior of digital reference answerers. Still, digital reference and social Q&A services are quite different; enough so that it is possible that users of the two types of services have different standards for evaluating the answers provided. In order to truly determine the relevance of a previously provided answer for a new question in the context of digital reference, it must be tested in the context of a digital reference service.

The author therefore proposes the following experiment: that one or more digital reference services implement an algorithm to automatically suggest resources. All digital reference services must have a question submission webform of some kind. Upon a user submitting a question, that question can be used as the query to search for relevant answers and URLs from the corpus of previously provided answers. Answers and URLs would then be returned to the user, similar to a page of search engine results. The user could then be presented with the option to submit their question to be answered by a human answerer if none of the retrieved resources were satisfactorily relevant. It would then be an empirical question: what percentage of submitted questions were satisfactorily answered by reused resources, and what percentage were submitted to a human? The interface and presentation of the results and the submit option might of course also affect this submission rate. But it seems clear that it would be preferable to collect relevance assessments from the questioners themselves, rather than from proxies, as the present study did with Mechanical Turk workers.

The author hypothesizes that the larger the corpus of answered questions, the greater the probability that a previously provided answer will exist that is relevant to any new question. Another important future direction for this type of development is to make use of other corpora of answered reference questions. While the IPL's ARQ is extensive and possibly the largest archive of answered digital reference questions in existence, it is hardly the only one: in addition to the QuestionPoint Global Knowledge Base, nearly every online reference service stores its answered questions, even if only in its e-mail outbox. The IPL is also only one service; it would be useful to mine resources across multiple services or from a consortial service. Further, the Ask an IPL Librarian service is asynchronous: the user submits a question via a webform and receives a reply by e-mail. It would be useful to mine resources from services using synchronous media, such as chat, instant messaging, and SMS. Finally, and perhaps more controversially, it might be useful to mine resources from social Q&A sites such as Yahoo! Answers, Ask MetaFilter, and the Wikipedia Reference Desk.

The idea of creating a large corpus of answered questions is hardly new: the Digital Reference Electronic Warehouse (DREW) was proposed in this journal as a project to collect reference transactions from multiple services into a single large database.[54] The authors who proposed the DREW articulated three uses of such a repository: support of teaching and research; informing management of and decision making for reference services; and modeling the flow of traffic, topics, and other factors in the world of digital reference. The current author proposes a fourth use: support for automated question answering, or more accurately automated answer suggestion. Indeed, the author would go so far as to suggest that automated answer suggestion is the key to addressing the problem of scalability of digital reference services.

CONCLUSION

This paper began with a quote in which the author takes it for granted that reusing the resources provided in answers to reference questions would be useful. Despite that quote being nearly a decade old, and the sentiment expressed in it being even older than that, no research had been undertaken to determine if this sentiment is well founded or misplaced. This study sought to test the hypothesis that the resources provided in answers to reference questions may in fact be useful to answer future questions. The findings of this study support this hypothesis and support the sentiment expressed by Coffman.

The resources provided in responses to digital reference questions follow a power law distribution, like most information-related phenomena. In the case of the Ask an IPL Librarian reference service, this means that a few URLs and domains were provided in answers frequently—at least once and in some cases several times per day—and most URLs and domains were provided few times, and many only once ever in the service's history. The half-life of this corpus of resources is approximately eleven years, which is considerably longer than the half-life of any body of literature previously studied. The resources in this corpus also remain relevant over the long term.

This study was exploratory, to identify whether it would be useful to reuse the information resources provided by reference librarians in answers. While there are some significant challenges to collecting and making use of these resources, this study has determined that it would be useful to reuse these resources. The author suggests an experiment by which digital reference services can test the degree of usefulness of these resources to new questions submitted to services.

ACKNOWLEDGMENTS

The author wishes to thank the staff of the Internet Public Library for providing access to the Archive of Reference Questions. Thanks also to Stephanie Haas and Eliah Hecht for collaborating on research that led to this paper; Eliah Hecht and Chirag Shah for significant assistance with building the parser, and Terrell Russell and Jacob Kramer-Duffield for originally bringing Mechanical Turk to my attention. The author received assistance from the staff of the Odum Institute for Research in Social Science at UNC-CH (www.odum.unc.edu) in using SAS. This material is based upon work supported by the University Research Council of the University of North Carolina at Chapel Hill.

References and Notes

1. Steve Coffman, "We'll Take It from Here: Developments We'd Like to See in Virtual Reference Software," *Information Technology and Libraries* 20, no. 3 (2001): 152.
2. R. David Lankes, "The Foundations of Digital Reference," in *Digital Reference Service in the New Millennium: Planning, Management, and Evaluation*. The New Library Series vol. 6, R. David Lankes, John W. Collins, and Abby S. Kasowitz, eds. (New York: Neal-Schuman., 2000), 246.
3. Brenda Dervin, "Useful Theory of Librarianship: Communication, Not Information," *Drexel Library Quarterly* 13, no. 3 (1977): 16–32.
4. Richard L. Trueswell, "Some Behavioral Patterns of Library Users: The 80/20 Rule," *Wilson Library Bulletin* 43, no. 5 (1969): 458–61.
5. Kenneth D. Crews, "The Accuracy of Reference Service: Variables for Research and Implementation," *LISR* 10, no. 3 (1988): 331–55; Matthew L. Saxton, "Reference Service Evaluation and Meta-analysis: Findings and Methodological Issues," *Library Quarterly* 67, no. 3 (1997): 267–88.
6. Paul Neuhaus, Connie Van Fleet, and Danny P. Wallace, "Privacy and Confidentiality in Digital Reference," *Reference & User Services Quarterly* 43, no. 1 (2003): 26–36.
7. QuestionPoint: 24/7 Reference Services, "Did you know..." Feb. 19, 2009, http://questionpoint.blogs.com/questionpoint_247_referen/2009/02/did-you-know-that-the-global-knowledge-base-has-over-20000-searchable-records-in-it-in-fact-as-of-this-writing-1.html (accessed Oct. 11, 2010).
8. OCLC, "Let Patrons Search the Global Knowledge Base," www.questionpoint.org/support/documentation/templates/search_globalkb.html (accessed Oct. 11, 2010).
9. Internet Public Library, "About ipl2," http://ipl.org/div/about/index.html (accessed Oct. 11, 2010).
10. *Ask an ipl2 Librarian Digital Reference Service Student and Volunteer Training Manual*, Section 1, http://training.ipl.org/div/backroom/refvols/students (accessed Oct. 11, 2010).
11. Ibid, Section 3.3.
12. Nettie Lagace and Michael McClennen, "QRC: We Call It Quirk," *Computers in Libraries* 18, no. 2 (1998): 26–27.
13. Coffman, "We'll Take It from Here," 152.
14. Birger Hjørland, "Information: Objective or Subjective/situational?" *Journal of the American Society for Information Science and Technology* 58, no. 10 (2007): 1448–56.
15. Georgina F. Payne and David Bradbury, "An Automated Approach to Online Digital Reference: The Open University Library OPAL Project," *Program: Electronic Library and Information Systems* 36, no. 1 (2002): 5–12.
16. Hoa Trang Dang, Diane Kelly, and Jimmy Lin, "Overview of the Trec 2007 Question Answering Track" (paper presented to the Thirteenth Text REtrieval Conference (TREC 16), Gaithersburg, Md., Nov. 5–9, 2007), http://trec.nist.gov/pubs/trec16/papers/QA.OVERVIEW16.pdf (accessed Oct. 11, 2010).
17. Lynn Bry, "Simple and Sophisticated Methods for Processing Large Volumes of Question and Answer Information through the World Wide Web," in *Digital Reference Service in the New Millennium: Planning, Management, and Evaluation*. The New Library Series vol. 6, ed. R. David Lankes, John W. Collins, and Abby S. Kasowitz (New York: Neal-Schuman, 2000), 111–23.
18. Ibid., 118.
19. Chirag Shah, Sanghee Oh, and Jung Sun Oh, "Research Agenda for Social Q&A" (under review).
20. Soojung Kim and Sanghee Oh, "Users' Relevance Criteria for Evaluating Answers in a Social Q&A Site," *Journal of the American Society for Information Science and Technology* 60, no. 4 (2009): 716–27.
21. Sanghee Oh, Jung Sun Oh, and Chirag Shah, "The Use of Information Sources by the Internet Users in Answering Question," in *Proceeding of the 71st Annual Meeting of the American Society for Information Science and Technology* (2008).
22. Ellen M. Voorhees, "The Trec-8 Question Answering Track Report" (paper presented to the Eighth Text REtrieval Conference (TREC 8), Gaithersburg, Md.:, Nov. 16–19, 1999), 77, http://trec.nist.gov/pubs/trec8/papers/qa_report.pdf (accessed Oct. 11, 2010).
23. This is one of the test questions from the 2007 TREC QA Track. Test questions and data sets for the TREC QA Tracks are available at http://trec.nist.gov/data/qamain.html.
24. John Burger et al., *Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A)*, Gaithersburg, Md.: National Institute of Standards and Technology, 2001, www-nlpir.nist.gov/projects/duc/papers/qa.Roadmap-paper_v2.doc (accessed Oct. 11, 2010).
25. Anne R. Diekema, Ozgur Yilmazel, and Elizabeth D. Liddy, "Evaluation of Restricted Domain Question-Answering Systems" (paper presented to the ACL 2004 Workshop on Question Answering, Barcelona, Spain, 2004), www.clt.mq.edu.au/Events/Conferences/

acl04qa/papers/diekema.pdf (accessed Oct. 11, 2010); Elizabeth D. Liddy, Anne Diekema, and Ozgur Yilmazel, "Context-Based Question-Answering Evaluation" (paper presented to the SIGIR 2004, Sheffield, UK, 2004), www.cnlp.org/publications/04Liddy.SIGIR.Poster.2004.v2.pdf (accessed Oct. 11, 2010).

26. Marc Light et al., "Reuse in Question Answering: A Preliminary Study" (paper presented to the Proceedings in AAAI Symposium, New Directions in Question Answering, Palo Alto, CA, 2003), 6, www.cnlp.org/publications/reuseFinal.pdf (accessed Oct. 11, 2010).
27. Robin D. Burke et al., "Question Answering from Frequently Asked Question Files: Experiences with the FAQ FINDER System," *AI Magazine* 18, no. 2 (1997): 57–66.
28. Chung-Hsien Wu, Jui-Feng Yeh, and Ming-Jun Chen, "Domain-specific FAQ Retrieval Using Independent Aspects," *ACM Transactions on Asian Language Information Processing (TALIP)* 4, no. 1 (2005): 1–17; Valentin Jijkoun and Maarten de Rijke, "Retrieving Answers from Frequently Asked Questions Pages on the Web," in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management* (Bremen, Germany: ACM, 2005), 76–83.
29. Xiaobing Xue, Jiwoon Jeon, and W. Bruce Croft, "Retrieval Models for Question and Answer Archives," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Singapore, Singapore: ACM, 2008), 475–82.
30. Lagace and McClennen, "QRC: We Call It Quirk," 26–27.
31. Internet Public Library, "Membership Information," http://ipl.org/div/about/IPLconsortium/IPLMembershipInfo.pdf (accessed Oct. 11, 2010).
32. William Doran, personal communication, April 22, 2009.
33. *Ask an ipl2 Librarian Digital Reference Service Student and Volunteer Training Manual*, Section 6g.
34. Ibid, Section 4d; Section 6i.
35. P. Mockapetris, "Network Working Group Request for Comments 1034: Domain Names—Concepts and Facilities," 1987, http://tools.ietf.org/html/rfc1034 (accessed Oct. 11, 2010).
36. "Hypertext Transfer Protocol -- HTTP/1.1," www.w3.org/Protocols/rfc2616/rfc2616.html (accessed Oct. 11, 2010).
37. Amazon.com, "FAQ > Overview," 2009, www.mturk.com/mturk/help?helpPage=overview (accessed Oct. 11, 2010).
38. The company CastingWords (http://castingwords.com) has built an extremely successful transcription business on top of Mechanical Turk.
39. A body of evidence is emerging that results from Mechanical Turk are comparable to many established findings in the social sciences. The Experimental Turk blog has been aggregating these findings in fields such as experimental and social psychology and economic behavior: http://experimentalturk.wordpress.com.
40. Tefko Saracevic, "Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II," *Advances in Librarianship* 30 (2006): 3–71.
41. Tefko Saracevic et al., "A Study of Information Seeking and Retrieving. I. Background and Methodology," *Journal of the American Society for Information Science* 39, no. 3 (1988): 161–76.
42. For example: Louise T. Su, "A Comprehensive and Systematic Model of User Evaluation of Web Search Engines: I. Theory and Background," *Journal of the American Society for Information Science and Technology* 54, no. 13 (2003): 1175–92; Youngok Choi and Edie M. Rasmussen, "Users' Relevance Criteria in Image Retrieval in American history," *Information Processing & Management* 38, no. 5 (2002): 695–726.
43. *Ask an ipl2 Librarian Digital Reference Service Student and Volunteer Training Manual*, Section 4j.
44. Lori Mon and Joseph W. Janes, "The Thank You Study: User Satisfaction with Digital Reference Service" (Dublin, OH: OCLC Online Computer Library Center, Inc., 2003); David S. Carter and Joseph Janes, "Unobtrusive Data Analysis of Digital Reference Questions and Service at the Internet Public Library: An Exploratory Study," *Library Trends* 49, no. 2 (2000): 251–65.
45. *Ask an ipl2 Librarian Digital Reference Service Student and Volunteer Training Manual*, Section 6b.
46. http://en.wikipedia.org/wiki/Slashdot_effect.
47. Wallace Koehler, "A Longitudinal Study of Web Pages Continued: A Report after Six Years," *Information Research* 9 (2004), Introduction section, ¶ 3, http://InformationR.net/ir/9-2/paper174.html (accessed Oct. 11, 2010).
48. Ibid; Wallace Koehler, "Web Page Change and Persistence—a Four-year Longitudinal Study," *Journal of the American Society for Information Science and Technology* 53, no. 2 (2002): 162–71.; Wallace Koehler, "An Analysis of Web Page and Web Site Constancy and Permanence," *Journal of the American Society for Information Science* 50, no. 2 (1999): 162–80.
49. Daniela V. Dimitrova and Michael Bugeja, "The Half-life of Internet References Cited in Communication Journals," *New Media Society* 9, no. 5 (Oct. 2007), 811–26.; Dion Hoe-Lian Goh and Peng Kin Ng, "Link Decay in Leading Information Science Journals," *Journal of the American Society for Information Science and Technology* 58, no. 1 (2007): 15–24.; Diomidis Spinellis, "The Decay and Failures of Web References," *Communications of the ACM* 46 (2003): 71–77.; Mary Rumsey, "Runaway Train: Problems of Permanence, Accessibility, and Stability in the Use of Web Sources in Law Review Citations," *Law Library Journal* 94 (2002): 27–39.
50. Eugene Agichtein, Yandong Liu, and Jiang Bian, "Modeling Information-seeker Satisfaction in Community Question Answering," *ACM Transactions on Knowledge Discovery from Data* 3, no. 2 (2009): 1–27.
51. C. J. Van Rijsbergen, *Information Retrieval* (Boston: Butterworths, 1979).
52. Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee, "Finding Similar Questions in Large Question and Answer Archives," in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management* (Bremen, Germany: ACM, 2005), 84–90.
53. Oh, Oh, and Shah, "Use of Information Sources by the Internet Users."
54. Scott Nicholson and R. David Lankes, "The Digital Reference Electronic Warehouse Project: Creating the Infrastructure for Digital Reference Research through a Multidisciplinary Knowledge Base," *Reference & User Services Quarterly* 46, no. 3 (2007): 45–59.