When Documents Deceive: Trust and Provenance as New Factors for Information Retrieval in a Tangled Web

Clifford A. Lynch
Coalition for Networked Information, 21 Dupont Circle, Washington, DC 20036. E-mail: cliff@cni.org

Introduction

Historically information retrieval has focused on the indexing and retrieval of documents or surrogates from databases with little regard to how the indexing has been obtained or whether the surrogates are accurate. Information retrieval systems have dealt with databases that are assumed to be well behaved, consistent, and often admission controlled, and questions of trust and data accuracy have been completely implicit, to the extent that they have been considered at all.

Highly distributed information dissemination systems like the World Wide Web herald a fundamental change to these assumptions which will, in my view, have broad-reaching implications for the design and use of the next generations of information retrieval systems. These developments also motivate an entirely new research agenda for both the theory and engineering practice of information retrieval systems in the networked information environment. Among the consequences of this shift will be a new emphasis on the provenance of data and metadata, and the need for information retrieval systems to permit users to factor in trust preferences about this information.

This brief and somewhat informal article outlines a personal view of the changing framework for information retrieval suggested by the Web environment, and then goes on to speculate about how some of these changes may manifest in upcoming generations of information retrieval systems. It also sketches some ideas about the broader context of trust management infrastructure that will be needed to support these developments, and it points towards a number of new research agendas that will be critical during this decade. The pursuit of these agendas is going to call for new collaborations between information scientists and a wide range of other disciplines.

Much of what is being described here is emerging from the folklore and practical engineering knowledge of the Web and of constructing search engines for it, and is not well documented or formalized in the research literature. I am indebted to Jack Xu of Excite@Home for some very helpful presentations related to these topics (notably to the Buckland/Lynch seminar at the University of California, Berkeley School of Information Management and Systems), to Avi Rappoport, and to the very useful Searchenginewatch.com site for insights and data that support some of the information presented here. Also relevant is the NSF/ERCIM Digital Libraries working group on metadata report (http://www.iei.pi.cnr.it/DELOS//NSF/metadata.html).

Fundamental But Little Noted Assumptions in Information Retrieval System Design

Traditional information retrieval systems make several fundamental environmental assumptions that are so basic it sounds strange and a little crazy to question them. In particular:

(1) The documents that an IR system "sees" (e.g., in the indexing, retrieval, or ranking process) are the same ones that a user would retrieve if he or she chose to select those documents. How could it be otherwise? These documents are part of a database that is an integral component of the information retrieval system, and the system is internally consistent; every read operation on a given document should produce the same result.

(2) Metadata (surrogate records) for documents can be taken at face value as honest attempts to accurately describe documents, and should be treated this way in retrieval systems. A retrieval system either works with documents or with surrogates; if it works with surrogates, the relationship between surrogate and document is outside the scope of the IR system proper. For all practical purposes, the surrogates *are* the documents in this scenario.

These two assumptions are just different aspects of the same general view of the world. In one case, the creation/extraction/computation of metadata is done within the IR system as part of indexing or retrieval (indexing is just precomputation for retrieval in some sense); in the other

© 2001 John Wiley & Sons, Inc. ●

case, the development of metadata (or at least the first step) takes place "outside" the IR system, and it is assumed that it is done in a disinterested and accurate fashion (bibliographic citations, abstracts, etc), whether by computer algorithms or human beings. It is considered legitimate to discuss how much access or retrieval quality is lost by replacing documents with these externally produced surrogates (e.g., debates about full text versus surrogate retrieval), but the assumption is always that the creators of surrogates do the best job they can, subject perhaps to some fundamental constraints about economics, time, protection of intellectual property, computational resources, size of surrogate, etc.

These core design assumptions are completely at odds with the realities of the distributed information environment found on the World Wide Web today.

Digital documents in a distributed environment may not behave consistently; because they are presented both to people who want to view them and software systems that want to index them by computer programs, they can be changed, perhaps radically, for each presentation. Each presentation can be tailored for a specific recipient. Further, the information that a human takes away from a presentation of a document through mediating software such as a Web browser may be very different from what an indexing program extracts even from the identical source document, unless the indexing program is designed to consider the perceptual impact of the document on human beings.

Finally, in a distributed system of information publishing and accompanying metadata, the metadata may be carefully constructed by any number of parties to manipulate the behavior of retrieval systems that use it, rather than simply describing the documents or other digital objects it may be associated with.

Bluntly, these assumptions are no longer true.

Yet these assumptions underlying information retrieval system design are amazingly fundamental, pervasive, and deep; so much so that I do not recall ever seeing them explicitly stated in the traditional IR literature.

The Changing Framework for Information Retrieval in the Networked Information Environment

Traditional information retrieval deals with two types of databases: full documents, and surrogates (metadata) such as bibliographic citations or abstracts. Surrogates, when used, are assumed to be accurate, or at least not deliberately misleading; organizations producing catalogs or abstracting and indexing (A&I) databases are very fussy about who they let contribute records (indeed, this is a long-standing source of tension in community copy cataloging databases, and a competitive advantage for A&I vendors). Essentially, surrogates are assumed to be accurate because they are produced by trusted parties, who are the only parties allowed to contribute records to these databases. Documents (full documents or surrogate records) are viewed as passive; they do not actively deceive the IR system. Specifically, the designer's mental model is one of a file that can be read and reread, and that contains the same contents every time (unless there is a new version of the document, in which case the database is viewed as having been updated). And a user, having identified a document that he or she wants to inspect (e.g., by scanning a search result), should get the same document that the retrieval system examined.

Compare this to the realities of the Web environment. Anyone can create any metadata they want about any object on the net, with any motivation. Further, documents are not files—rather, they are the returned to human viewers or indexing programs as the result of a computation performed by some server within the distributed environment in response to a protocol request.

The way Web indexing systems operate is they run programs (called "crawlers" or "spiders") that visit Web sites, issue requests for pages, perform computations to evaluate and index these pages, and then place the results into Web index databases that support searching. Most of the details about the commercial Web indexing systems are proprietary: how they select the pages that they will index and how deeply they will explore the pages in a given Web site; how often they revisit sites; and precisely how they evaluate and index pages.

Sites interested in manipulating the results of the indexing process rapidly began to exploit the difference between the document as viewed by the user and the document as analyzed by the indexing crawler through a set of techniques broadly called "index spamming." For example, a document might be stuffed with thousands of words that the user would not see because they blended into the page background in a tiny font, but which would be found by the indexing crawler. The result has been an ongoing arms race between indexers and Web site developers, with the indexing services adding greater sophistication in word extraction, statistical analysis, natural language processing, and other technology. The indexing services also supplement direct indexing of content with contextual information, such as how many other sites link to a page, as a way of trying to identify important pages.

It is important to understand that when a crawler requests a page for indexing it is not simply reading a file in some sort of network file system; it is making a request for a page to a Web server through the http protocol. The request includes identification of the request source (at several levels—the software that is asking, and the machine that is sent the request), and the Web server can be programmed to respond differently to identical requests from different sources. The reasons for this may be fairly benign; for example, some servers provide pages that are tuned to index effectively with the indexing algorithms used by different crawlers. Other reasons for source-sensitive responses are more actively malicious, such as the practice of pagejacking. This is most easily illustrated by an example. Suppose you have a product X that competes with another product Y made by another company. When people issue queries to

Web search engines asking for Y you would like to get the search engine to return your page advertising X instead. You take a copy of the page for Y, and give this to the Web indexing service, but when a user (as opposed to the indexing service) clicks on the URL, you return the page for your product X instead of the copied page for Y. Competition is not the only motive; for example, perhaps you would like to ensure that the pages of an organization you do not like are returned in response to requests for explicit sexual material. Pagejacking might be defined generally as providing arbitrary documents with independent arbitrary index entries. Clearly, building information retrieval systems to cope with this environment is a huge problem, and Web crawlers are beginning to integrate a wide range of validity checks (such as looking at link networks between pages and sites) to attempt to identify and filter likely pagejacking attempts.

Note that selective response is used for many other reasons than dealing with indexing crawlers; for example, a Web site that offers licensed content may do access control based on request source address or host domain and simply respond “access not permitted” if the request is not from an authorized site. Some sites offering adult material may refuse requests for pages from sources that they believe belong to government or law-enforcement agencies. There are crawlers operated by services (e.g., Digimarc) that look for watermarked pictures that have been taken and reposted on other sites without the rightsholder’s permission; one could easily believe this might give rise to a selective response strategy from some sites if the crawler could be identified.

These developments suggest a research agenda that addresses indexing countermeasures and counter-countermeasures; ways of anonymously or pseudononymously spot-checking the results of Web-crawling software, and of identifying, filtering out, and punishing attempts to manipulate the indexing process such as query-source-sensitive responses or deceptively structured pages that exploit the gap between presentation and content. Down this path also lies work on competitive counterintelligence and information warfare. Fully developing these issues is beyond the scope of this article, and we will leave this line of inquiry here. But the reader should recognize that many of these things are already happening, on an ad hoc, grassroots level within today’s Web.

Of course, another alternative for indexing services is to only crawl sites that are known to behave responsibly. But this implies some system or economy of certification or rating authorities; some set of methods for these authorities to evaluate sites on a continuing basis; and the need for an indexing service to decide which of these authorities to believe, and, of course, infrastructure for obtaining ratings or certifications (such as PICS (http://www.w3c.org/PICS/))—although this is almost certainly the easiest part of the problem.

Metadata in an Environment of Systematic Deception

The absence of human-provided metadata as a base for supporting queries is painfully evident in the limitations of today’s Web search engines; with all of their power to provide access to an enormous array of information they cannot make simple distinctions (e.g., works authored by an individual as opposed to works about an individual), which are well-established expectations in traditional databases such as on-line catalogs developed by trusted sources. Markup in documents that encodes added-value semantics (another form of metadata) such as the tagging of personal, organizational, and place names similarly cannot be exploited by these search engines. The fundamental problem is that we have very little technology to allow an indexing crawler to decide whether metadata can be believed or whether it is simply attached to a page in an attempt to further manipulate the indexing process. It seems reasonable to believe that heuristics could be developed to check the consistency of some metadata against the objects it describes (e.g., subject terms could be algorithmically validated against a statistical and/or natural language analysis of the text they are assigned to, supplemented by the use of semantic networks, thesauri, and other databases), but relatively little work has been done in this area for several reasons: there is not that much metadata out on the Web to try to exploit, and there are very real limits to what we can expect from such heuristics. There is also a basic deployment problem here: if Web indexing services do not use metadata, who will go to the expense and trouble of creating and maintaining it? The only place we are seeing much use of metadata is within controlled environments—Web search engines that index sites on organizational intranets, or selected clusters of sites (such as subject gateways)—where the sites within the controlled environment can be trusted to behave responsibly.

There are other reasons why the inability to integrate metadata into Web indexing and searching is a major problem. A tremendous amount of material exists on the Web, or is accessible through the Web, which cannot be indexed simply by retrieving Web pages—this is sometimes called the “dark matter” in the Web, or “the invisible Web.” It includes both databases that manifest themselves through query forms and dynamically computed Web pages that are delivered in response to queries, and collections of proprietary material where the content owner is unwilling to permit arbitrary access by indexing services (but may still want to advertise the material by making some information available about it).

Metadata seems to be the best and most efficient way to make this dark matter visible to Web indexing and search services. There are other alternatives, but they are poorly understood research problems and also suffer from performance problems. For example, one can imagine developing protocols that permit indexing services to transverse, read out, and summarize an entire public database hiding behind a query form, but it is not clear how to perform such summarization algorithmically, and it would require huge amounts of information to be moved across the net. Further, it would likely miss many key properties of a database: scope, intent, frequency of updating and the like. Proprietary content might be indexed through some sort of trusted (controlled, quarantined) computation and transfer of the results of computation; essentially, the ability to have a crawler examine and extract indexing from the proprietary content under the computational supervision of the content provider; but the protocols and infrastructure for this do not exist today, nor do we have confidence in our ability to quarantine information in this way.

Indeed, there are enormous inefficiencies in the way that current Web indexing services operate; it would be much more efficient to be able to do index entry extraction at or near the content sites, in cooperation with the sites (e.g., to rely on the site to make new or changed material available to the indexing systems as changes occur, rather than reindexing the site periodically) as well as being able to incorporate metadata. The Harvest system (Bowman, Danzig, Hardy, Manber, & Schwartz, 1994a; see also Bowman, Danzig, Hardy, Manber, Schwartz, & Wessels, 1994b) provides much of the basic mechanical framework for such a restructuring of the way the Web is indexed, although it would need substantial extensions to allow different indexing services to continue to vie for competitive advantage through unique indexing algorithms; we would need to establish protocols for "landing pads" (remote execution environments) or registries of indexing algorithms either at individual sites or indexing servers for those sites, and ensure that the local execution of these indexing algorithms did not present a security risk for the sites that host such execution.

There are then a series of research questions which, if solved, might partially (but not completely) mitigate the need to integrate metadata into Web search services. But they only reduce the need, not eliminate it, and effective implementation means that hundreds of thousands of content providers need to alter their Web sites; deployment requires collaboration between the indexing services and the content providers. This is one of the problems that has proven to be a major barrier to making Web indexing more efficient. The indexing services have historically been more motivated to work with arbitrary sites than sites have been to make provision to be indexed by Web crawlers with special requirements. In a world of competitive indexing services, this balance of power is not likely to change; while numerous sites are interested in manipulating their placement in response to searches of Web indexing services, there is limited motivation for the very large-scale deployment of a complicated infrastructure that makes it more rather than less difficult for sites to manipulate their placement.

So, our best hope, at least in the near term, is probably to be able to integrate author or third-party metadata into the indices that support searching of the Web. It is in the interest of the vast majority of sites to be able to provide such metadata. And the greatest barrier is our inability to rely on the accuracy of this metadata. Independent verification is one, albeit limited (as discussed above) way to establish trust. Another alternative is to attempt to identify and validate the source of each metadata assertion, and to explicitly consider the extent to which users of a search system are willing to trust various metadata providers (including anonymous or unsourced metadata, or metadata where the alleged source cannot be validated to a requisite level of confidence).

Some of the mechanics of this are reasonably well established. We know how (at least in theory, although the specific standards for actual implementation are messy, to say the least) to use a public/private key pair to sign a metadata assertion expressed in a syntax such as RDF (the resource description framework), and to verify a signed assertion (see the work on RDF and also the joint World Wide Web Consortium—Internet Engineering Task Force working group on signed XML; information on both is at www.w3c.org). Implementing such signatures on metadata is not difficult or disruptive for content provider sites. It does not represent a significant architectural change, for example, in Web servers.

But unless you have direct knowledge of the public keys of all of the potential signatories you might be interested in, a key registry system is necessary. In the development of public key infrastructure (PKI) systems we have the basis for binding identities or identifiers ("names") assigned by, and warranted by (presumably reputable and trusted), third parties, to public/private key pairs. Companies such as Verisign operate such registries today, offering verification of identities with different levels of confidence or strength (this is based on the procedures that are used when the identity is established, such as the types of documents that need to be examined). Other companies offer software that organizations (governments, educational institutions, businesses, etc.) can use to implement their own PKI registries.

Other approaches, most notably the Pretty Good Privacy (PGP) system, treat the establishment of identity in a more distributed fashion; you begin with a series of identity/key bindings that you trust because you have established them yourself, through personal face-to-face key exchange or because you have received them directly (in a way that you feel is sufficiently secure) from a source that you trust. You then can establish trust in new, unfamiliar identity/key bindings because they are vouched for (cryptographically signed) by one or more parties that you have already established trust in, and because you trust them to evaluate other identity/key bindings. This "web of trust" can then be extended indefinitely under the user's control, where a level of trust in an identity/key binding is set by the number of chains (beginning with trusted identities) leading to the binding under evaluation, the origin parties in those chains, and the length of the chains.

There is substantial literature (much of it not formally published) on the structure of the PGP Web of trust (see, e.g., http://bcn.bolder.co.us/~neal/pgpstat). There is also significant research literature on trust management issues (see Blaze, Feigenbaum, Ioannidis, & Keromytis, 1999; Chu, Feigenbaum, LaMacchia, Resnick, & Strauss, 1997).

Trust and Provenance in Retrieval

So the tools are coming into place that let one determine the source of a metadata assertion (or, more precisely and more generally) the identity of the person or organization that stands behind the assertion, and to establish a level of trust in this identity. One can have near-absolute confidence that the source possessed the requisite public/private key pair (assuming that the private key has not been compromised, and the key pair has not been revoked—and you do have to trust the party holding the key pair to guard and manage it responsibly); the level of trust is in the binding of identity to possession of the key pair.

It is essential to recognize that in the information retrieval context one is not concerned so much with *identity* as with *behavior*. Knowledge of identity creates some accountability for behavior, and observation of behavior over time allows one to form expectations about the behavior associated with an identity. This distinction is often overlooked or misunderstood in discussions about what problems PKI is likely to solve: identity alone does not necessarily solve the problem of whether to trust information provided by, or warranted by, that identity. It is only when we can use our knowledge of past behavior or our (perhaps very subjective) assessment of the character of an individual or organization to establish trust in behavior that a level of trust in identity helps us. And all of the technology for propagating trust, either in hierarchical (PKI) or web-of-trust identity management, is purely about trust in identity. PGP does make the distiction between trusting a certificate and trusting the identity established by that certificate to "introduce" or vouch for other certificates as part of establishing your trust in them within its trust model. And a similar notion is at least explicit in Certificate Authority interrelationships. But the only behavior in the vocabulary is establishing trust in someone else's identity. This is not enough to help much in the broader information retrieval context. The fact that I am willing to vouch for someone else's identity/key pair binding means that I believe the binding is true, and perhaps at most that I also believe that the person I am vouching for will manage the key pair responsibly. It certainly does not mean that I am making any general assertion about the behavior of that individual (he or she always tells the truth, is kind to animals, creates accurate and high-quality metadata, etc.).

The question of formalizing and recording expectations about behavior, or trust in behavior, are extraordinarily complex, and as far as I know, very poorly explored. There are a number of avenues: certification or rating services that might be consulted, or webs of individuals vouching for behavior of others. To make this real, meaningful taxonomies of behavior classes would have to be established (and note that in the "web of vouching" case trusting an individual to rate another party with regard to a certain class of behavior is a *different, distinct* type of behavior from the behavior that is being rated! Though many people may be willing to accept approximations: for example, someone who is known to be a good creator of metadata can also decide if someone else is a good creator of metadata); and, of course, an appeal to certification or rating services simply shifts the problem: how are these services going to track, evaluate, and rate behavior, or certify skills and behavior? To take just one case in point, I have heard suggestions that some people would be willing to use metadata "created by librarians." This would require the existence of some organization that would not only credential librarians, but also maintain a "rogue librarian" list of people who had been credentialed but subsequently were found to regularly create deceptive metadata.

The vision here is one in which personal preferences dominate; a very diverse world that empowers information seekers and rejects central control. An individual should be able to decide how he or she is willing to have identity established, and when to believe information created by or associated with such an identity. Further, each individual should be able to have this personal database evolve over time based on experience and changing beliefs. This will require powerful tools for defining and maintaining a view of the world that can be provided as input to various information retrieval services. And there are interesting and difficult architectural questions about how much of this world view actually has to be explicitly revealed as part of using information retrieval services; individuals may not wish to fully reveal or export their models of trust to external (perhaps commercial) services, but only to permit these services to consult such a model as part of query processing. While doing all this, of course, the management burden on the user needs to be kept extremely low.

The ability to scale and to respond to a dynamic environment in which new information sources are constantly emerging is also vital. For all but the most paranoid and parochial users, it should be easy to extend trust to new and unfamiliar information sources under reasonable constraints of prudence (in other words, given that a known and trusted party vouches for the new information source).

People trust different sources in different spheres of their information seeking and evaluation. Investment tips, recipies, legal advice, and health care information may come with different trust preferences. Is it realistic to contextualize searches within a topical trust framework or parameterization, and if so, how many contexts will uses need, and how should they be structured?

These observations suggest that information retrieval in the distributed environment is going to become a very complex process. In determining what data a user (or an indexing system, which may make global policy decisions) is going to consider in matching a set of search criteria, a way of defining the acceptable level of trust in the identity of the source of the data will be needed. Having gained sufficient confidence in the identity of the source, methods of deciding whether the expected behavior of that source are acceptable will need to be employed. Only if the data is supported by *both* sufficient trust in the identity of the source and the behavior of that identity will it be considered eligible for comparison to the search criteria. Alternatively, just as ranking of result sets provided a more flexible model of retrieval than just deciding whether documents or surrogates did or did not match a group of search criteria, one can imagine developing systems that integrate confidence in the data source (both identity and behavior, or perhaps only behavior, with trust in identity having some absolute minimum value) into ranking algorithms. Obviously, there are numerous open research problems in designing such systems: how can the user express these confidence or trust constraints; how should the system integrate them into ranking techniques; how can efficient index structures and query evaluation algorithms be designed that integrate these factors.

Conclusions: Societal Implications of the Integration of Trust

The very idea of formalizing and systematizing trust is complex and alien to most people. Information retrieval systems such as Web search engines are themselves complex and hard to understand; in general, they are just treated as “black boxes” and more or less trusted. Almost nobody understands *why* they get the results that they do from a search engine; they just deal with the results that they do get. The networked information environment is already so complicated that most users simply accept defaults established by software and service providers, often without knowing that they are doing so. As a case in point, Web browsers today contain tables of certificate authorities that are trusted to establish identity; very few users even know that these tables exist, much less explore and examine them critically and customize them. The power to assign default values is tremendously powerful.

As we integrate trust and provenance into the next generations of information retrieval systems we must recognize that system designers face a heavy burden of responsibility. Defaults about who should be trusted amount to a de facto censorship mechanism and a very strong editorial voice. New design goals will need to include making users aware of defaults; encouraging personalization; and helping users to understand the behavior of retrieval systems—why a given result was retrieved and ranked the way it was, and the interaction between this outcome and the trust-related parameters supplied to the as input. Powerful paternalistic systems that simply set up trust-related parameters as part of the indexing process and thus *automatically* apply a fixed set of such parameters to each search submitted to the retrieval system will be a real danger; such systems will be appealing to designers because they can be simpler and more efficient and equally seductive to users because they conceal, and thus apparently minimize complexity—after all, the user just wants an answer.

The integration of trust and provenance into information retrieval systems is clearly going to be necessary and, I believe, inevitable. If done properly, this will inform and empower users; if done incorrectly, it threatens to be a tremendously powerful engine of censorship and control over information access. Undoubtedly, some commercial and even political interests will choose to try to deploy systems that censor and restrict rather than empower; but we may hope that system developers will not make such choices on grounds of avoiding technical difficulties, and that research will permit us to gain sufficient understanding to be able to develop and deploy systems that will truly empower users to deal with an environment that is characterized not only by information overload but active deception by information providers.

References

Blaze, M., Feigenbaum, J., Ioannidis, J., & Keromytis, A. (1999). The role of trust management in distributed system security. In J. Vitek & C Jensen (Eds.), Secure internet programming: Security issues for distributed and mobile objects (Lecture Notes in Computer Science, vol. 1603) (pp. 185–210). Berlin: Springer, Berlin. http://www.research.att.com/~jf/pubs/sip99.ps.

Bowman, C.M., Danzig, P.B., Hardy, D.R., Manber, U., & Schwartz, M. F. (1994a). The Harvest information discovery and access system. Proceedings of the second international World Wide Web conference (pp. 763–771), Chicago, IL.

Bowman, C.M.; Danzig, P.B., Hardy, D.R., Manber, U., Schwartz, M. F., & Wessels, D.P. (1994b). Harvest: A scalable, customizable discovery and access system. Technical report CU-CS-732-94. Boulder, CO: Department of Computer Science, University of Colorado.

Chu, Y.-h., Feigenbaum, J., LaMacchia, B., Resnick, P., & Strauss, M. (1997). REFEREE: Trust management for Web applications. World Wide Web Journal, 2, 127–139. Reprinted in 1997 from Proceedings of the 6th international World Wide Web conference (pp. 227–238), Cambridge, MA; World Wide Web Consortium. http://www.si.umich/edu/~presnick/papers/Referee/www6-referee.html.