

NSF DL Position Paper

Carl Lagoze, Cornell University

The roots of the NSF digital library initiative lie in the pre-web era. An unfortunate result of this has been an artificial separation between web research and digital library research. I believe that this separation has played out to the disadvantage of the digital library community. This is ironic on two counts. First, one of the most fundamental advances in the Web, page rank that led to the Google phenomenon, is a result of NSF DLI funding. Second, the vast majority of digital library work has been enabled advances in web technologies – Java, XML, RDF – yet almost none of these technologies are the result of DLI projects.

The cost of this separation is evident in the quality of the respective research in the fields. There is, in my opinion, a disturbing imbalance in the strength of the two areas as demonstrated by the quality of their major conferences. Over the years the quality of papers at the World Wide Web Conference have increased dramatically, with exciting results in fundamental areas such web characterization, modeling, and link analysis. In contrast, in my six years as program committee for the major digital library conferences – ECDL and JCDL – I have seen a disturbing decline in paper quality and a notable shortage of fresh new ideas and research directions. Results have been largely incremental with few signs of exciting new threads developing.

I propose in this paper that future funding for research in network-based information must address our current lack of basic understanding of the web; how it is structured, how people use it, how it affects social interaction, how it grows over time, and multiple other questions. I believe that our challenge is to learn how to exploit the ubiquity and phenomenal richness of the web environment and layer upon it the high-integrity contexts that we endow with the name ‘libraries’. Formal understanding the web space is essential for building scalable, economical, and sustainable digital libraries.

Before proceeding, I should more carefully define what I mean by ‘the web’. At one level the web is a set of protocols and languages – in particular HTTP, URLs, and HTML – that are in fact ephemeral. In this sense, the web is not something that will last forever and will surely be supplanted by future technologies. Yet at another level the web is a truly metaphysical phenomenon, an interlinked knowledge network independent of the technology. In this sense one can argue that the web has always existed because humans from pre-history have formed social networks that shared information and referred to each other and their concepts. The instantiation of this social web as a technical reality, the World Wide Web, marks an event heretofore unseen in history. That is, our inherent social and knowledge networks can now be accessed, monitored, and analyzed and, with the help of massive computation power, perhaps understood. A plausible, albeit possibly controversial hypothesis, is that by studying and understanding the web we can better understand social structures and the organization, flow, and development of knowledge. This is the web, that I propose as a research focus, rather than the purely technical web. In the remainder of this paper I enumerate a number of future research areas within this broad umbrella.

Many parts of this short paper derive from a recently submitted NSF ITR proposal submitted by Cornell and other institutions. As such, I owe credit to colleagues including Bill Arms, Jon Kleinberg, Paul Ginsparg, and Dan Huttenlocher at Cornell, Raimie Stata at UC Santa Cruz, Lee Giles at Penn State, and Brewster Kahle at the Internet Archive.

Understanding the structure of the web

Recent research has produced intriguing results related to the structure of the web including some understanding of its linkage patterns, the multiple communities represented in it, and its size and diameter. Formal understanding and modeling of these characteristics is essential to provide the baseline information for analyses described later in this paper. Without some notion of ‘normalcy’ in the web, it is impossible to recognize perturbations, spikes, and notable patterns that evidence something semantically meaningful or that can be classified as knowledge rather than as simple data.

Initial results in web characterization merely scratch the surface and the rapidly changing nature of the web calls out the need for a significant research effort in this area. Due to its size alone, the web challenges our current hardware capacities, mathematical models, and computational techniques. “Web-sized” experiments are difficult to undertake since relatively basic tools such as web crawling is still understand at only a basic level. These attempts to analyze the web demonstrate the need for research in improved graph modeling techniques and understanding of random effects in large-scale systems.

The base model for our understanding the web – as a graph where nodes are resources and edges are its hyperlinks – is proving too simplistic. For example, it is well-known that not all edges in the graph are equal. Factors such as nepotistic links, whereby sites are self-referential, are known to interfere with graph-based analysis techniques such as page rank. Furthermore, there is well-known but marginally understood polarity in links which might indicate recommendations or condemnations with varying degrees of strength – e.g., a scholarly citation to a ‘particularly good example’ in contrast to a paper that the citer strongly disagrees with. There are open research issues in natural language understanding of link anchor text and how to apply link polarity and its degree to basic web models and ranking techniques.

Another shortcoming in existing web models lies in the simplistic view of a node in the graph. Most work assumes that a node is a ‘document’ accessible via a URL. However, the modern web is far more complex than static HTML documents and in fact behind URLs lays complex and dynamic content and the large databases and repositories that make up the hidden web. The problems with incorporating these complex objects into a web model are similar to those that accompanied extending Garfield’s initial citation analysis work from journal influence to individual paper influence. If we intend to develop tools to automatically analyze the web we must develop models that disaggregate these complex nodes in the graph or aggregate them when necessary (i.e., into strongly connected components).

Understanding resource quality

Using the web as the basis for a managed information environment immediately introduces questions about resource quality. Our work in the NSDL has confronted these questions since we not only want to build a large-scale digital library but one where users can filter the library at varying quality thresholds. While quality assessment is certainly possible with human mediation, such mediation is economically infeasible at the scale of the information spaces we wish to build. Consequently, automated quality mechanisms are a fundamental area in web-based information research.

There is considerable initial work in this area. Research and development of annotation systems provide the basis for tools where people and agents may impose quality assessments on web resources and portions thereof. The utility of these annotations beyond simple display for human perusal remains to be studied and developed. Link-based search algorithms such as page rank and HITS attempt to infer quality by recursively quantifying links. Research leading to better understanding of the structure underlying the links, both in terms of node granularity and edge polarity, may improve these automatic quality assessments. Finally, important open questions remain about reputation systems that facilitate ‘reviewing the reviewers’. While a number of algorithms exist to assess reputation, they have proven fragile to intentional attacks and need to be improved to be a reliable basis for automated quality assessment.

The Dimension of Time

Certainly one of the interesting distinctions between bricks and mortar libraries and networked information systems is the dynamic nature of the objects and collections. Digital objects and collections are inherently mutable and there is a need for substantial work on the time dimension in digital information. This work lies in two main areas.

There is the need for work on the time dimension at the individual object level. Besides quality, discussed in the previous section, the two most critical integrity issues with web-based information are provenance – where does an object come from and how has it changed (been tampered with) over time – and persistence – how long will an object last and how to make it last longer. Recent DLI2 work provides some interesting foundation work in this area. This includes data provenance work under Peter Buneman at University of Pennsylvania and our own work on event-based knowledge models (the ABC model) and digital object risk management. The first two results focus on models for recording changes in digital objects and databases over time, while the latter aims to develop techniques for automatically monitoring web resources (sites and pages) and enforce formally expressed preservation policies. At present results from these projects are still preliminary and focused research initiatives are necessary to move them to the state where practical and reliable digital libraries such as the NSDL can be built on top of the web layer.

Perhaps more intriguing is research that addresses the flow and growth of information on the web. If, as I proposed earlier, the web indeed increasingly provides a mirror of the broad scope of human activity (cultural, social, economic, legal, intellectual), then research results in the ‘the time axis of the web’ may give us important tools for understanding past events (e.g., the development of an important intellectual idea) or even predicting future events (e.g., a potential terrorist attack). Projects like the Internet Archive and Wayback Machine lay some basis for this work by taking periodic snapshots of the web as a whole and especially high-frequency snapshots of focused sites during special events – presidential elections and September 11. In addition, Jon Kleinberg’s work on text-based burst analysis has proven effective in a number of contexts including news burst recognition (www.daypop.com) and research trend recognition (in the context of arXiv). But this work needs to be moved up to web scale and into real time (in order to monitor events as they are happening), and herein there is opportunity for considerable research. One area of particular interest is synchronization of data from multiple resources. As pointed out earlier, underlying the web as a simple document structure are databases, news feeds, weblogs, etc. Thus, burst analysis is more than simple text basis but the recognition and relation of heterogeneous events in distinct data types. Doing this at a very large scale and rapidly will require powerful new formalisms and algorithms.

Understanding and adopting to the user

An increasing number of commercial web sites increase their level of service by profiling users, manually and automatically, and adapting to their behaviors. Any user of Amazon will notice the sensitivity of search results to both global user behavior (what is popular at the global level) and individual user behavior (what have I ordered in the past). Techniques for understanding how to effectively exploit user behavior are still at their initial stages and need to be generalized at the level of academic research rather than simply as a commercial activity. Of particular interest are the means of doing this without compromising user privacy.

Automatic Classification and Organization of Web Information

At the heart of any library problem is the need to organize information for diverse purposes such as management, discovery, and browsing. Perhaps due to inertia from the library cataloging tradition, there has been far too much attention to metadata (in its human-generated form) in digital library activities. A good bit of this work is motivated by dubious or outdated assumptions (metadata is critical to resource discovery) and the results have disappointing (the quality of non-professional produced metadata is degraded to the point of non-utility) and non-scalable (human cataloging even in its most simple form simply doesn’t scale up to the web).

In response to this reality we have seen increasing interest in both automatic metadata generation and collection aggregation. Liz Liddy’s work on ‘breaking the metadata bottleneck’ is an excellent example of the former and my own group has produced notable results in automatic collection aggregation and organization. Both areas deserve substantial greater research

attention, and can especially benefit from increased understanding of basic web structure as described at the beginning of this paper.

Many of the techniques for automatic metadata generation have heretofore been restricted to analysis of a textual object and inferences from its contents. The success of Google demonstrates the importance of contextual information (i.e. link structure) for ranking search results and projects such as Google image search and the CLiMB project at Columbia show how context can provide ‘aboutness’ information relative to an object. It appears that there is substantial opportunity to improve that state-of-the-art of automated metadata generation (and thereby break through a major scalability issue in information management) by understanding how to fully exploit information context for generation of descriptive information. A special benefit of this work would be descriptive metadata generation for non-textual objects. This work is related to the efforts to understand ‘communities’ in web space. These communities have multiple motivations, but the community I am proposing here is effectively an ‘aboutness community’ – the set of web sites graphically neighboring a selected web site that offer descriptive information about the respective site.

Moving beyond the object level, libraries consist of collections and users of information (both managers and consumers) are often more comfortable in dealing with information in the aggregate rather than at the finest granularity. Research work on automatic inference of these aggregations on the web can make an important contribution to our ability to make the web a more effective knowledge environment. Even the most commonly used information aggregation on the web – the notion of a ‘site’ – is poorly understood. Work to automatically discover more abstract aggregations, such as groupings into semantically-based collections or equivalence classes (where equivalence is more complex than simple bit equality) has shown some progress but needs considerable more focus before it can be effectively deployed in real information systems.

Conclusions

As stated at the beginning of this paper, digital library research has an unfortunate distinction from web research. In fact, there has been tremendous and exciting progress in the creation of a ubiquitous knowledge environment in web space that has progressed independent of DLI-funded research. The opportunity now exists for a refocus of NSF funding in this area and make substantial progress in our understanding of the web phenomenon. I look forward for discussions at the workshop on this issue.