Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities

Andrew Jackson[1], Jimmy Lin[2], Ian Milligan[2], and Nick Ruest[3]

[1] The British Library [2] University of Waterloo [3] York University

Andrew.Jackson@bl.uk, {jimmylin,i2milligan}@uwaterloo.ca, ruestn@yorku.ca

ABSTRACT

Web archiving initiatives around the world capture ephemeral web content to preserve our collective digital memory. In this paper, we describe initial experiences in providing an exploratory search interface to web archives for humanities scholars and social scientists. We describe our initial implementation and discuss our findings in terms of desiderata for such a system. It is clear that the standard organization of a search engine results page (SERP), consisting of an ordered list of hits, is inadequate to support the needs of scholars. Shneiderman's mantra for visual information seeking ("overview first, zoom and filter, then details-on-demand") provides a nice organizing principle for interface design, to which we propose an addendum: "Make everything transparent". We elaborate on this by highlighting the importance of the temporal dimension of web pages as well as issues surrounding metadata and veracity.

1. INTRODUCTION

Web archiving refers to the systematic collection and preservation of web content for future generations. Since web pages are ephemeral and disappear with great regularity [13], the only sure way of preserving web content for posterity is to proactively crawl and store portions of the web. Since 1996, the Internet Archive has captured and made publicly accessible hundreds of billions of web pages. Today, many libraries, universities, and other organizations have ongoing web archiving initiatives [7]. Although content capture is by no means a solved problem—in particular, social media and highly-interactive JavaScript-heavy pages present ongoing challenges—scholars now have at their disposal a rich treasure trove of material to study. The focus of our work is how to make these materials accessible to humanities scholars and social scientists.

This paper describes our initial experiences in providing an exploratory search interface to web archives to support scholarly activities. We describe our initial implementation and present our findings in terms of desiderata for such a system, summarized as follows: It is clear that the standard organization of a search engine results page (SERP), consisting of an ordered list of hits, is inadequate to support the needs of scholars. Shneiderman's mantra for visual information seeking ("overview first, zoom and filter, then details-on-demand" [15]) provides a nice organizing principle for the types of exploratory search interfaces that we desire. Elaborating on this, we discuss the importance of metadata and the issue of veracity—helping scholars understand the quality characteristics of the content, including biases that might be present. To Shneiderman's mantra, we propose an addendum: "Make everything transparent". We argue that a tool in support of scholarship should not have "magic". Every system decision—from the ordering of results to how aggregations are computed—should be available for inspection and manipulation by the scholar.

We view the contribution of this paper as starting a conversation with digital library and information retrieval researchers on the underexplored problem of searching web archives. While our findings are accurately characterized as preliminary, and we are by no means the first to examine the information seeking behavior of scholars (cf. [3]), to our knowledge the focus on web archiving is novel. Our discussion enriches the literature on complex information seeking and system support for such activities.

2. BACKGROUND AND RELATED WORK

Most early work on web archiving has focused on the content acquisition pipeline (collection development, crawling, storage file formats, etc.) as opposed to providing access. In fact, a question the community has perpetually struggled with is "who's using web archives and for what purposes?" The Internet Archive boasts impressive access statistics,[1] but AlNoamany et al. [1] found that most requests are actually by robots. Dougherty and Meyer [5] identified lack of shared practices, accessible tools, and clear legal and ethical guides as obstacles to advancing scholarly use of web archives. According to that article, the Internet Archive has identified three categories of current users, and, surprisingly, scholarly use *is not* one of the categories.

The most popular (and in many cases, the *only*) access method to web archives is temporal browsing, or what is commonly know as "wayback" functionality. Given the URL of a page, a user can view a particular version of a web page, move forward and backward in time to examine different captured versions, and follow links to contemporaneous

[1] twitter.com/brewster_kahle/status/364834158285041665

JCDL '16, June 19 - 23, 2016, Newark, NJ, USA

ISBN 978-1-4503-4229-2/16/06…$15.00

DOI: http://dx.doi.org/10.1145/2910896.2910912

pages. Obviously, browsing is only useful if one knows the exact URL of the desired content. Since this is often not the case, search is an obvious solution [4, 6], but unfortunately, most web archives do not support full-text search. There has been academic work on searching timestamped collections (such as web archives) [12, 10, 2, 8], but these systems have not been deployed in production to our knowledge. Regardless, most previous work has focused on technical issues such as the layout and organization of inverted index structures. In contrast, there has been relatively little work on search interfaces in direct support of users' needs.

3. INTERFACE DESIGN

To begin, a word about methods: this paper represents the distillation of the authors' personal experiences working on web archives over the past several years—the authors include a historian, a librarian, a researcher in information retrieval, and a software engineer working in a major national web archiving effort. Our findings and recommendations are based on these experiences, informal interactions with colleagues at various professional events, and informal feedback from users of our working prototype.

3.1 Task Model

The development of search interfaces must begin with an understanding of who the users are and what they are trying to accomplish. In our case, we wish to support the activities of humanities scholars and social scientists. It is important to recognize that search isn't necessarily the most natural starting point for these users—search presupposes that it is possible to articulate (however poorly) an information need. In our experience, most scholars don't even know "where to start" with a web archive. To a large extent, this is because web archives are relatively novel artifacts that few scholars have had experiences with. Nevertheless, there is usually "something" that occurs before search.[2]

The chess analogy of Hearst et al. [9] for information navigation seems apt for characterizing the task model for humanities scholars and social scientists. In the "opening", they want a high-level overview of what's in a collection and how it was gathered. For the humanist, this is often called "distant reading" (aggregation and large-scale data analysis) to elicit "provocations". For social scientists, exploratory analyses are often intertwined with the process of hypothesis generation. Search is a poor tool for the "opening".

At the other end of the task model is the "end game" in our chess analogy: For a humanist, this might involve the "close reading" of several records (e.g., webpages) to construct a narrative. For a social scientist, this might involve extracting variables of interest from text or metadata and applying a regression to illustrate some hypothesized relationship. Once again, search is not particularly useful for this stage of the game.

Between the "opening" and "end game" lies the "middle game", and this is where we believe search plays a vital role. We readily concede that this three-stage model is a vast oversimplification of reality, eliding many important issues: Scholarly activities extend over many sessions, perhaps lasting months or even years. Scholarship is fundamentally iterative with false starts, backtracking, and feedback loops, e.g., consideration of pages leads to the reformulation of the hypothesis. Despite these inadequacies, we believe that our model is nevertheless helpful in situating search.

3.2 Current Implementation

We have implemented and deployed a prototype exploratory search interface for web archives, available online at webarchives.ca. The interface was originally developed by the British Library's web archiving team for the Big UK Domain Data for the Arts and Humanities (BUDDAH) project in order to facilitate access to their legal deposit crawl collection. We have adapted the tool to host the Canadian Political Parties and Political Interest Groups collection, gathered by the University of Toronto Library using the Internet Archive's Archive-It platform.[3] The collection contains 14.5 million documents from crawls performed at roughly quarterly intervals between October 2005 and March 2015. Content from around fifty organizations were collected: all of the major Canadian political parties (the Conservative Party, the Liberal Party, the New Democratic Party, the Green Party, and the Bloc Quebecois), as well as minor parties and political organizations such as the Assembly of First Nations, the Canadian Association for Free Expression, Fair Vote Canada, and beyond. The entire collection totals 380 GB in size compressed. The prototype search interface is powered by Apache Solr. The underlying Lucene indexes are generated by Warcbase [11], a platform for managing web archives built on Hadoop and HBase. All components of the system are open source.

In developing this prototype, we faced a classic chicken-and-egg problem: scholars have a difficult time articulating what capabilities they desire in a search interface for web archives, and without some notion of requirements, it is difficult to build a prototype. We attempted to break out of this cycle by implementing, at least in the beginning, an interface similar to what most users today have come to expect from a web search engine: a simple search box and results organized as an ordered list, just like a standard search engine results page (SERP). Note, however, that the results are *not* algorithmically ranked, but simply presented in archival (i.e., temporal) order; we discuss this decision later. A screenshot of the interface is shown in Figure 1 (left) for the query "recession". Running down the left edge of the interface are controls for faceted navigation, which lets users filter content type (HTML, PDF, etc.), year of crawl, site, and a few other facets. Next to each facet we show the number of documents that match the filter criterion. We believe that faceted navigation is sufficiently commonplace today (in sites like Amazon.com) that users will be able to manipulate the controls without requiring instructions or training. In the current results display, different versions of the same document are treated as if they were different documents; we have an experimental feature deployed elsewhere that groups different versions of the same page together. However, as we discuss later, there are issues with both approaches.

As an alternative interface, we developed a "trends visualization" inspired by Google's Ngram Viewer,[4] shown in Figure 1 (right). In this view, we are able to concurrently visualize the prevalence (i.e., frequency) of term matches over time for multiple queries. Here, we show trends for the

[2]This is a deliberately vague statement because, as Dougherty and Meyer [5] point out: "researchers have trouble deciding what they want methodologically before they begin".

[3]`archive-it.org/collections/227`

[4]`books.google.com/ngrams`

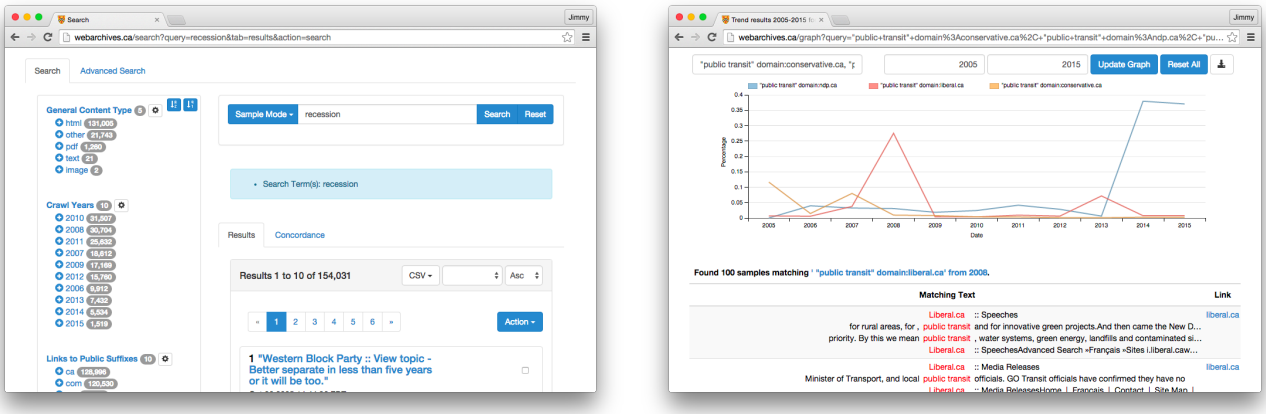

Figure 1: Screenshots of our exploratory search interface: on the left, a standard SERP layout and controls for faceted navigation; on the right, the trends visualization showing the prevalence of the phrase "public transit" on the websites of the Conservative Party, the Liberal Party, and the New Democratic Party.

phrase "public transit" on the websites of the Conservative Party, the Liberal Party, and the New Democratic Party. The user can click on the line graph to obtain sample results, shown as a keyword-in-context concordance. In this example, a scholar might use this interface to explore how each party's views on public transit evolved over time.

4. DISCUSSION

A Standard SERP Isn't Enough. One immediately obvious finding about our existing prototype is the inadequacy of the standard SERP organization. There are two important interacting issues:

First, our choice *not* to rank the results bears some discussion. Although it would be simple enough to provide a ranking model (e.g., BM25), some scholars we interacted with questioned the very idea of relevance ranking. To them, *all* of the matching results are relevant from a scholarly perspective. Often, they are looking for information beyond the documents themselves, i.e., for traces of evidence that allow them to infer information about the authors of those documents or the societal context in which they were created. Because these scholars are treating the tool primarily as a "lens" through which to examine society, all matching documents should be considered. However, without relevance ranking, scholars are frequently presented with overwhelming numbers of results and no way to prioritize their attention. This observation about the needs of scholars deserves further study because it challenges the probability ranking principle [14], one of the central tenants of modern information retrieval research.

Second, regarding the choice between grouping different versions of the same page in the results view or treating them as if they were different documents: both options are problematic. Without grouping, results are frequently populated by duplicates or near duplicates. Grouping versions creates a different issue: pages change over time, and it is often these changes that are of interest (for example, when a particular phrase is removed from a page). An interface that performs grouping can inadvertently hide insights.

The fundamental deficiency with the standard SERP organization is that it offers only a single linear dimension with which to organize search results, whereas scholars desire a view into a multi-dimensional document space. Faceted browsing helps, but does not solve the problem.

Shneiderman's Mantra. We believe that Shneiderman's mantra for visual information seeking ("overview first, zoom and filter, then details-on-demand" [15]) provides a nice organizing principle for exploratory search interfaces to web archives. The trends visualization in Figure 1 (left) appears to be a useful starting point based on our own experiences and feedback from colleagues—scholars find the trends view preferable to the standard SERP organization. In Shneiderman's mantra, trends provide the "overview first". Interactions with the trends provide "zoom and filter" capabilities. The user is able to click on a particular year and query (e.g., the Liberal party in 2008) and bring up sample results that match the criteria. Finally, "details-on-demand" is supported by the user clicking on a particular sample in the concordance view to bring up the archived page.

Shneiderman's mantra brings to focus what humanities scholars refer to as "distant reading", or understanding text through massive data analysis, and "close reading", which is the careful interpretation of select passages. Exploratory interfaces, especially for humanities scholars, need to support seamless movement between the two modes and to adjust the "distance" of reading: although many trends only become apparent during distant reading, criticism and interpretation requires close reading.

Metadata and Veracity. Although the trends visualization is a good starting point, we have already identified several areas for improvement, the most salient of which is better support for faceted navigation of metadata. Even in the traditional SERP organization, scholars commented positively about the support for faceted browsing. Beyond obvious facets such as content type and source, there are a number of facets that can be straightforwardly derived from page content. For example, we can group sites into categories such as "news" and "social media". This requires only a modest amount of effort, and often these metadata are already available during collection development. We can also break down pages in terms of the number of incoming or outgoing links (appropriately bucketed). Another facet we have been experimenting with is based on named entities. For example, a scholar might be interested in mentions of op-

ponents in a particular party's website. The current trends visualization lacks support for faceted browsing, which represents one future direction for development.

Related to metadata is the issue of veracity. Scholars consistently express concerns about the veracity of whatever insights our interface purports to show—for example, is a trend reflective of underlying shifts in content, or merely an artifact of the crawl or decisions made by the system? Many of these concerns stem from unfamiliarity with web archives as objects of study: when a historian approaches a traditional archive to examine the personal papers of an important figure, for example, he or she has a fairly good idea of "what to expect", including potential biases that might be inherent in the collection. Not so with web archives. Part of the solution is to better educate scholars on the technical nuances of web crawling and related technical issues.

We provide a concrete example: mention aggregation, one of the most common and useful tools for distant reading, can be insightful but is fraught with peril. We can learn from our collection, for example, that Liberal Party of Canada leader Michael Ignatieff appeared 160k times on the liberal.ca website in 2008 at the peak of his leadership (approximately 5% of the pages), versus only 22k times or 1.4% of documents in 2012, the year after he resigned. Broadly, we can see the fall from favor of a politician, but to what extent is this finding an artifact of biases in content selection? Crawl volume (in absolute number of documents) can widely vary (for example, due to minor changes in scoping rules), thus distorting frequencies. Even if we take these counts at face value, the aggregates don't tell us the context of the mentions: What fraction of the mentions were from boilerplate (e.g., page footers)? What fraction were mentioned in a policy context? Surely these distinctions would be relevant for a political scientist or a historian. At present, our interface does not provide any mechanism to explore these questions. However, it may be possible to answer many of the types of questions posed above via faceted browsing, provided that we are able to accurately extract the relevant facets.

Make Everything Transparent. Shneiderman's mantra provides a nice organizing principle, to which we propose an addendum: "Make everything transparent". The "magic" of modern search engines in placing relevant results at the top of a ranked list is *not* what scholars want—they are instinctively distrustful of any mechanism they don't understand. If documents are presented in a particular order, scholars will want to know exactly how the ranking was generated. Something simple such as date ordering may be suboptimal, but at least it is understandable. If the system presents an aggregation, it should explain what information is potentially lost. Quite simply, transparency increases veracity.

Note that this recommendation does not preclude the use of machine learning or other automated techniques for ranking, classifying, clustering, etc., simply that the output must be transparent to the scholar. They are not the only group of users with this requirement: lawyers share similar requirements in the context of electronic discovery and health professional in searching the medical literature.

5. ONGOING WORK

Recognizing that search support is directly applicable to only a relatively small portion of scholarly activities (Section 3.1), we are currently exploring tighter integration of our prototype with Warcbase [11], a related project that provides a general platform for analyzing web archives. Using Spark, a framework for large-scale data processing, users can perform arbitrarily complex filtering and aggregations, ranging from gathering page statistics to extracting the hyperlink structure and named entities mentions from page contents. These capabilities complement what scholars can accomplish with search alone.

We fully recognize that what we currently have is a case study focused on a particular collection. While illuminating, there may be idiosyncratic characteristics that prevent us from making generalizations across different web archives. Thus, we hope to engage more scholars to explore different types of collections. We have only begun to scratch the surface in developing tools to support scholarly access to web archives. We hope that this paper begins a conversation with computer scientists and generates community interest in tackling the many challenges in this space.

Acknowledgments. This research was supported by the AHRC under grant number AH/L009854/1, the U.S. National Science Foundation under awards IIS-1218043 and CNS-1405688. Additional support was forthcoming from the Social Sciences and Humanities Research Council of Canada under Insight Grant 435-2015-0011, and the Ontario Ministry of Research and Innovation's Early Researcher Award program. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the sponsors.

6. REFERENCES

[1] Y. AlNoamany, M. Weigle, and M. Nelson. Access patterns for robots and humans in web archives. *JCDL*, 2013.

[2] K. Berberich, S. Bedathur, T. Neumann, and G. Weikum. A time machine for text search. *SIGIR*, 2007.

[3] G. Buchanan, S. J. Cunningham, A. Blandford, J. Rimmer, and C. Warwick. Information seeking by humanities scholars. *ECDL*, 2005.

[4] M. Costa, D. Gomes, F. Couto, and M. Silva. A survey of web archive search architectures. *WWW Companion*, 2013.

[5] M. Dougherty and E. Meyer. Community, tools, and practices in web archiving: The state of the art in relation to social science and humanities research needs. *JASIST*, 65(11):2195–2209, 2014.

[6] D. Gomes, D. Cruz, J. Miranda, M. Costa, and S. Fontes. Search the past with the Portuguese web archive. *WWW Companion*, 2013.

[7] D. Gomes, J. Miranda, and M. Costa. A survey on web archiving initiatives. *TPDL*, 2011.

[8] J. He, J. Zeng, and T. Suel. Improved index compression techniques for versioned document collections. *CIKM*, 2010.

[9] M. Hearst, P. Smalley, and C. Chandler. Faceted metadata for information architecture and search. *CHI*, 2006.

[10] M. Herscovici, R. Lempel, and S. Yogev. Efficient indexing of versioned document sequences. *ECIR*, 2007.

[11] J. Lin, M. Gholami, and J. Rao. Infrastructure for supporting exploration and discovery in web archives. *WWW Companion*, 2014.

[12] K. Nørvåg. Space-efficient support for temporal text indexing in a document archive context. *ECDL*, 2003.

[13] A. Ntoulas, J. Cho, and C. Olston. What's new on the web? The evolution of the web from a search engine perspective. *WWW*, 2004.

[14] S. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

[15] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. *IEEE Symposium on Visual Languages*, 1996.