

Managing Distributed Collections: Evaluating Web Page Changes, Movement, and Replacement

Zubin Dalal, Suvendu Dash, Pratik Dave, Luis Francisco-Revilla, Richard Furuta,
Unmil Karadkar, Frank Shipman

Department of Computer Science and Center for the Study of Digital Libraries
Texas A&M University
College Station, TX 77843-3112
1-979-862-3216

{shipman,furuta}@cs.tamu.edu

ABSTRACT

Distributed collections of Web materials are common. Bookmark lists, paths, and catalogs such as Yahoo! Directories require human maintenance to keep up to date with changes to the underlying documents. The Walden's Paths Path Manager is a tool to support the maintenance of distributed collections. Earlier efforts focused on recognizing the type and degree of change within Web pages and identifying pages no longer accessible. We now extend this work with algorithms for evaluating drastic changes to page content based on context. Additionally, we expand on previous work to locate moved pages and apply the modified approach to suggesting page replacements when the original page cannot be found. Based on these results we are redesigning the Path Manager to better support the range of assessments necessary to manage distributed collections.

Categories and Subject Descriptors

H.3.7 [Information Storage and Retrieval]: Digital Libraries – collection, system issues, user issues.

General Terms

Algorithms, Design, Experimentation, Human Factors

Keywords

Collection management, change detection, document location.

1. INTRODUCTION

The Web provides infrastructure for sharing a huge amount of useful information. Much effort is devoted to selecting from and organizing this information to create topical collections [1,5]. Most of these collections are distributed collections—in which a collection is a set of pointers or references to documents kept and controlled by others. But another characteristic of the Web is that

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

JCDL '04, June 7–11, 2004, Tucson, Arizona, USA.

Copyright 2004 ACM 1-58113-832-6/04/0006...\$5.00.

this information changes, moves, and even disappears over time [2]. When maintaining a digital library of Web-based documents, we constantly need to update the collection's contents, meta-data and structure to represent these changes.

Bookmarks, paths over Web pages, and catalogs such as Yahoo! and Google directories, are examples of page collections that can become out-of-date as changes are made to their components. The maintenance of these collections depends on evaluating the pages continuously, which currently relies on a high degree of human intervention and interpretation. Humans are responsible for identifying the degree of changes to these Web pages and deciding whether the current version fits the prior categorization or use. For example, Yahoo employs "surfers" to continually categorize and re-categorize Web sites in order to keep their directory up to date and others have built volunteer communities for this purpose. It is our goal to build tools to support this human activity, automating distributed collection maintenance when possible and focusing the limited human attention to where it is most needed. Section 2 presents a number of results characterizing change on the Web, motivating the subsequent system design.

There are three major types of changes that collection maintainers must cope with: edits within a page, replacement of pages, and unavailability of pages. Changes within a Web page can range from updating the number of visits to the page to replacements of large chunks of its content. In the extreme case, a page may be replaced with a completely new page. Our earlier work [9] presented algorithms comparing the content of versions of a Web page for determining the degree of change when pages were edited. These algorithms were successful at evaluating the degree of change within a page but could not distinguish when a page was heavily edited and redesigned from when a page was replaced with a similar page or when a page was replaced with a very different page. Section 3 describes the Walden's Paths Path Manager, our initial tool for managing distributed collections of Web pages.

After attempts based on more sophisticated forms of content analysis did not produce suitable results, we decided on a context-based method for comparing pages. As an example of why context is crucial in evaluating change to a Web page, consider two collections: one containing pages about French impressionist art and the other containing pages about art exhibits in Texas. Now consider a page about a visiting Monet exhibit at a museum

in Texas. This page may initially be in both collections. Now suppose that the museum's Web page changes to reflect that the visiting Monet exhibit has been replaced with an exhibit of art from the American West. The page still fits in the collection about area art exhibits but no longer fits in the context of the collection about French impressionist artists. This example shows why context (in this case the topic of the collection) must be taken into account when evaluating the degree of change to documents in the collection. Section 4 presents the context-based algorithms and their results.

Our earlier work on the Path Manager identified the changes to a page and the nature as well as the degree of these changes. When a page changed, the maintainer could let it continue to be a part of the collection if the changed page was acceptable or choose to exclude it from the collection. Also, the Path Manager would notify the maintainer when collections included broken links—i.e., URLs to pages were no longer valid. Pages identified as not accessible were classified based on whether or not the problem was likely transient or permanent.

Broken links result from a variety of different behaviors and circumstances. Amateurism in Web site design and maintenance often leads to repeated redesigns of a Web site [8]. In other cases the Web site as a whole moves to a different location, possibly due to changes in employment or school by the site maintainers, or for financial reasons such as better hosting contracts from a competing ISP. In some cases Web sites may stop functioning altogether as site hosting contracts run out, domain names expire or the maintainers lose interest. An obvious observation is that removal, relocation, or renaming of domains, files, and folders that constitute a Web site invalidates saved links to it.

Section 5 describes an approach to help maintainers of collections in locating pages that have moved and replacements for pages that they have chosen to exclude from their collection or have vanished from the Web. Pages from Websites that have ceased to function may sometimes be mirrored at other locations. We extend Phelps and Wilensky's [13] work on algorithms to locate these page mirrors that contain identical content, yet are located at administratively unrelated locations. Additionally, if a copy of the original page cannot be found our approach returns pages that contain "similar" content. The collection maintainer retains the right to replace the links in their collections to point to the new location in order to preserve the integrity of their collections.

The next section presents data about the rate of change to Web page accessibility and content. After this we briefly describe the Walden's Paths project and the Path Manager. Following this we present context-sensitive metrics of change and methods for locating Web pages that have moved and substitutes for pages that cannot be found. We conclude with a discussion of how these approaches are influencing the redesign of the Path Manager and future work for distributed collection management.

2. CHANGE ON THE WEB

That Web pages change over time is obvious to any frequent user of the Web and has been established by a number of recent studies. Our earlier research established the need for methods to assist Web collection maintainers by establishing the degree of change within the collection's Web materials. The majority of the existing literature exploring change on the Web has focused upon determining the frequency of Web page changes [2,6,7]. When

research has focused upon the nature of the changes themselves, it has rarely provided more than some metric of the degree of change usually measured as the number or proportion of elements within a page that have changed; indeed, this characterizes the approach in our own earlier work [9]. Little work has attempted to glean the nature of Web page change with an eye toward the topicality and relevance of a changed Web page when seen in a former context.

A difference in the rate and extent of Web page change based upon their hosts' domain (.gov, or .edu, for instance) has been demonstrated previously, as discussed below. Recently, we have performed a study of page change measured across a set of pages evaluated using the Walden's Path Manager during a period spanning more than 2 years. This work confirmed several of the findings of the existing research. We present a brief overview of related studies measuring frequency of change followed by the results of our analysis in this section. The results of this study led us to the conclusion that incorporating context and topicality into an analysis of Web page change is a necessary addition to tools incorporating Web-based collections.

Brewington and Cybenko [2] downloaded over 2 million Web pages provided by 25,000 users of a clipping service. Their work is oriented toward information retrieval and cache integrity issues, that these pages were selected on the basis of their being visited by human beings for some purpose is less pertinent within their analysis. The features of the pages recorded in their model include the last-modified date, and a number of stylistic elements such as the number of tables, images, and links. Approximately 100,000 pages were downloaded every day over a period of seven months in 1999, with no query run more than once every three days. From their results they arrive at stochastic formulae for modeling how often Web pages are likely to change.

Cho and Garcia-Molina [6] collected data from 720,000 pages on 270 Web sites over a period of four months in 1999. The pages were downloaded daily and compared to a recorded checksum to determine whether a page had changed. Their primary focus in this effort was to model proposed estimators for the frequency of change against real-world data and to derive some idea of the lifespan of a Web page. Although their data does not provide much measure of the degree of change, they do find a number of interesting results related to the domain of sites and its effect upon the frequency of updates. Notably, more than 70% of the pages across all domains were unchanged for at least one month and 50% of pages in the .gov and .edu domains lasted for more than 4 months (the duration of their study). Pages in .com were the shortest-lived, with a half of all .com pages changing within 11 days, while those in .gov and .edu experienced change at a far slower rate.

Fetterly, et al. [7], investigated the evolution of over 150 million Web pages downloaded weekly over a span of 11 weeks in 2002. For their work they recorded variables including the length of each downloaded document, the number of non-markup words in each document, and a set of variables related to their syntactic similarity measurement technique. Their measurement of similarity relies upon fixed length feature vectors derived from the document shingling technique [3]. Each document is broken into sets of word patterns and these groups of adjacent words (called "shingles") are mathematically compared to arrive at a measure of syntactic similarity. Their paper presents a number of

findings relating Web-site domain and document size to frequency of change. Pages in .edu domains tended to be smaller than those in other domains, although they possess similar word-counts—meaning that they either tend to use shorter words or less html-markup. They found that, counterintuitively, larger pages (32KB and above) tend to change both more frequently and to a larger extent than smaller pages (4KB and below). When considering domain they found that the domain of the site and the frequency of change were strongly correlated, with pages in .com, .net, and .cn (China) changing most frequently and often, with .gov sites changing among the least frequently. They also note that most changes consist of trivial or minor modifications, often of markup tags.

We performed a longitudinal evaluation of a smaller set of Web pages using our Path Manager tool over the course of two and a half years. Using a set of 2,244 Web pages specifically provided to us for use in another project involving entomology resources online, we performed an initial crawl in March 2001. Returning to the same set of documents in October 2003 we found that 1,126 of the pages (50%) of the URLs remained valid, and of these 36% had not changed in any way. Of the invalid pages, the majority were no longer available although some server responded (76%), were down or no longer served Web pages (8%), or belonged to unknown hosts (8%). These results seem to confirm the findings that educational and government materials (both of which were heavily represented in our set) seem to change the least frequently. Nevertheless, 1,118 of the pages in our set (49.8%) were no longer available in any form, and most (64%) of the pages that remained had changed to some extent.

Using the Path Manager we were able to measure the degree of change between the most recent set of downloaded pages and their initial versions. Using the Proportional algorithm described

in [9], we associate change with four levels of granularity: checksum only; low; moderate; and high. Of the pages in our sample that had changed 10% had only changed in mark-up or tags (checksum-only), 30% changed to a small degree, 20% changed to a moderate extent, and 40% changed to a large degree (possibly entirely). These results, while confirming rough guidelines of change determined on the basis of domain, tell us little about the relevance of the changes. Specifically, because the pages in our evaluation all dealt with entomological issues, knowing merely that they had changed led us to question the utility of such knowledge to maintainers of topic-oriented digital libraries.

Our own experimentation and the studies described earlier provide clear basis for evaluating the degree and frequency of change within Web pages. These and other studies provide unambiguous insight into the relevance of a variety of factors including site domain and document length to the frequency of Web page change. Beyond establishing a rough degree of change or in some cases simply that a change has occurred at all, greater attention must be paid to whether the change is relevant to a human reader. The syntactic shingling technique used above is likely to accurately measure the similarity between two documents when seen as ordered sequences of words independent of their meaning or relevance. Web pages, however, are as a practical matter selected and later referred to for some specific use. Our work described below attempts to incorporate some greater knowledge of the user's intent when describing Web page change.

3. WALDEN'S PATHS

Walden's Paths (<http://www.csdl.tamu.edu/walden>) is an application that can be used by the K-12 educators to organize

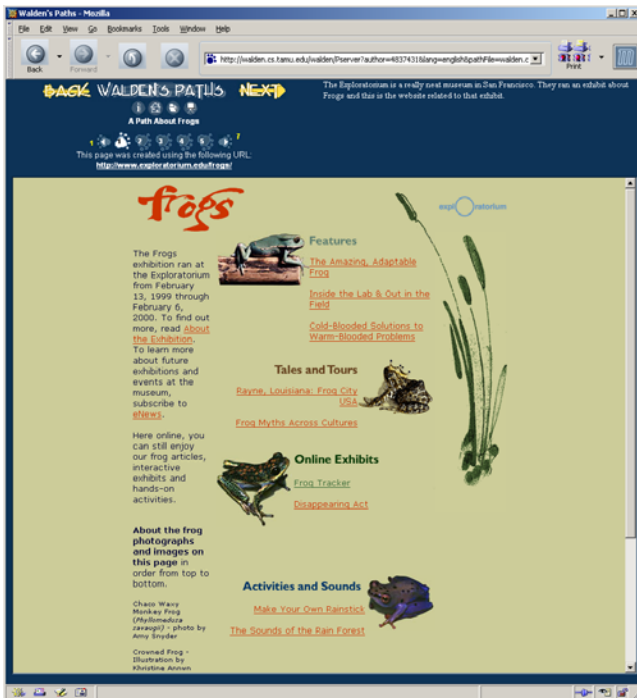


Figure 1. Walden's Paths

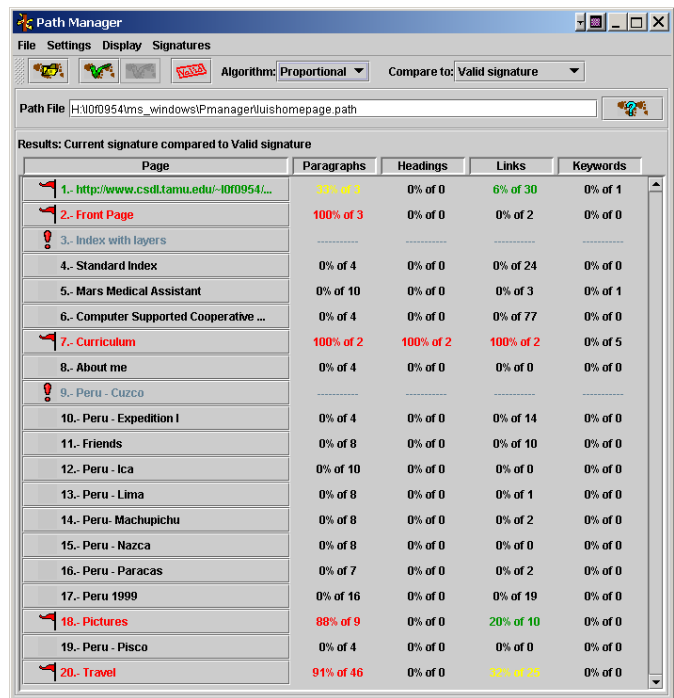


Figure 2. Path Manager

World-Wide Web material for their students' use [10]. Figure 1 shows how the user browses through the paths. It allows the creation of trails [4], paths using pages on the Web that have been created by others. The authors of the path organize the documents and add annotations or meta-data to the documents that provide contextualizing information to the original pages.

As the paths are built upon Web pages, Walden's Paths path maintainers would benefit by being able to automatically detect changes that happen to Web pages over the course of time. This is the primary motivation behind the Walden's Paths Path Manager [9], a tool that is meant to focus the collection maintainer's attention on the pages that have changed the most.

The Path Manager is a separate application that provides visual and quantitative feedback about the degree of change to a list of Web pages. Figure 2 shows how the user gets the metrics of change. Currently two algorithms evaluate the changes in the Web pages with respect to prior versions of the page. The two algorithms, Johnson's algorithm [11] and the Proportional algorithm, use signatures that abstract characteristics of a page including headers, links and text.

4. CONTEXT-BASED CHANGE METRICS

The initial change-detection algorithm included in the Path Manager uses content-based methods for evaluating similarity of Web pages. That is, it compares a computed signature of the prior page to the signature of the new page to determine the degree of change. But there was a problem when the degree of change was large—the algorithm could not distinguish between when the page had been heavily revised, and thus was likely to still be appropriate for a collection defined by topic, from when a page had been replaced by a similar page or even had been replaced by a page on a completely different topic.

This research investigates the use of context-based metrics of change to distinguish between these cases to better direct users' attention when maintaining paths. Within the Walden's Paths architecture, the context of the Web page consists of the other pages in the path and any additional annotation or meta-data provided by the path author. The current work uses only the other pages in the path as the context.

4.1 The Algorithm

The approach to calculate the context-based change metric is outlined here.

- Find the Term Vectors of the individual Web pages in the path. This is done by putting all the words present in the document (except the punctuation and stop words such as "a", "and", "the", etc.) in a vector. In addition, stemming is done on the terms present in the document.
- Find the Weight Vector for the terms present in the Term Vector for each of the pages. The term weight is calculated using many heuristics. The weight of a term is given as the logarithm of term frequency plus a constant scaling factor. The terms detected from among a list of nouns are given more weight, by multiplying a weight factor to the weight as calculated above.
- Save the Page Term Vectors and Weight Vectors for all pages into a signature file for the Path.

- Compute a Context Term Vector and a Context Weight Vector using a composition of the Term Vectors for all the Web pages in a particular path except the page whose change is being evaluated.
- Calculate the cosine similarity angle between the Context Vector and Page Term Vector. Then compare this angle to that for the previous version of the page. The difference between these two angles is used to compute the degree of change to the Web page.

This algorithm has been tested with existing paths and Web collections to determine meaningful cutoffs for representing results to path maintainers. From these results, including those reported in the next subsection, initial values for representing changes to the user are:

- A 2 degree or greater move by the page vector towards the path vector should be indicated as a move towards the path
- A move between 2 and -4 degrees should be indicated as being similar to the prior situation, and
- A 4 degree or greater move by the page vector away from the path vector should be indicated as a move away from the context of the path.

The chosen scaling and weight factors are only initial approximations and need to be evaluated in practice. As the results in section 4.2.2 indicate, this approach at noticing changes in context produces a range of angles; no single cutoff would always generate the desired results.

4.2 Evaluation

To evaluate the described algorithm's performance in determining whether changes result in a move out of the context of a collection, we compared the algorithm's evaluations to selections made by human catalogers.

4.2.1 Method

To perform the evaluation, pages for 20 paths were selected from Yahoo! directories, which rely on human selection for category membership. Each of the paths created consisted of between 10 to 12 pages from the directory. Pages were randomly selected but were checked to ensure that they were not images, flash presentations, or otherwise lacked text for the algorithm to compare. Once the path was created and term vectors were stored for all the component pages, a page in the path was randomly selected to be replaced.

Each selected page was replaced by at least three pages. Two of the pages used to replace the page were the same for all twenty paths. The first was about elephants and the second page was from the CNN Financials Web site. These pages were chosen to be different from one another and not part of any of the collections being used to generate the paths. The page was also replaced by one page from the human-maintained directory that was not part of the original path.

Were the algorithm to match the human catalogers, it would generate small angles of change relative to the context (path) vector when the page is replaced by another from the same collection. When the page is replaced by either the page on elephants or the CNN Financials page, the angle should be greater, and directed away from the path vector.

The results, presented in the next section, indicate that this expectation is often correct but is fallible.

4.2.2 Results

We applied the prior Walden's Paths Path Manager Proportional algorithm to determine how the existing tool would rate the degree of change for the 20 paths. The Proportional Algorithm rated all replacements of pages as a "high" degree of change. This is expected as the algorithm was designed to measure the degree of change within a page and page replacements should be considered extreme changes on such a scale. This reinforced our need for further algorithms to help users distinguish between acceptable and unacceptable page replacements.

We used two algorithms to evaluate the replacement of a page in a collection. The first measure is the angle between the original page's term vector and that of its replacement. The second measure is the context-based metric described in the previous sections that determines the change in angle between the original page and context vector and the new page and the context vector. Table 1 presents a summary of these results showing the averages, ranges and standard deviations for the results from the 20 paths.

The top half of the table shows the averages for angles between original pages and their replacements. The angles are quite high (75 to 82 degrees on average) regardless of whether the replacement was with a page in the same Yahoo! Directory or the unrelated pages. These results indicate that the angle between the original and replacement page does not provide much discrimination between when the page is topically similar and when it is not.

From the values in the bottom half of the table, we can see that when the pages were replaced by the Elephants page and the CNN Financials page the result was a move away from the Path vector by approximately 8 or 9 degrees. When replaced by another page in the same Yahoo! collection the result was a move towards the path by almost 2 degrees. These results indicate that, on average, this algorithm can be used to differentiate between replacements by pages with similar content and replacements with unrelated pages.

Table 1. Content-based and context-based measures of change for replacing element in collection.

	Replaced with	Page about elephants	CNN Financial page	Page in same Yahoo! directory
Angle between original and replacing pages (in degrees)	Average	78.1	81.9	75.1
	Range	30.8 to 88.1	77.0 to 87.7	35.1 to 84.5
	Standard deviation	15.65	2.89	10.76
Difference in angle to Yahoo directory between original and replacing pages (in degrees)	Average	-7.8	-9.1	1.9
	Range	-23.2 to 1.6	-45.0 to 0.9	-15.2 to 14.3
	Standard deviation	6.95	10.57	6.80

Table 2. Number and percentage of replacements resulting in moving towards and away from context vector.

	Moved away by more than 4 degrees	Moved to between -4 and 2 degrees of prior	Moved towards by more than 2 degrees
Replacement by member of same Yahoo! directory	1 (5%)	10 (50%)	9 (45%)
Replacement by non-member	25 (62.5%)	15 (37.5%)	0 (0%)

4.3 Discussion

The results from this study indicate that the context-based algorithm for evaluating the validity of page replacements worked better than the simple content-based algorithms we have been using in the Path Manager. While the context-based algorithm worked on average, it did generate false positives and false negatives.

As summarized in Table 2, of the 20 page replacements by another page in the Yahoo! collections, only one moved away from the context vector by more than 4 degrees. The replacement in this collection, on ozone depletion, moved away 15 degrees while the pages on elephants and CNN financials moved away by 23 and 26 degrees respectively. This is likely a case where the terminology on the replacing page was highly specialized and not included on other pages, even when they dealt with the same topic. An additional 10 of these replacements resulted in an angle between -4 and 2 degrees of the prior angle.

Of the 40 page replacements by either the page on elephants or the CNN financials page, 25 caused the angle with the context vector to move away by more than 4 degrees, and 15 caused the angle to move between -4 and 2 degrees towards the context vector. None caused a move towards the context vector by more than 2 degrees.

Collections where the page on elephants was not recognized as being out of context were: Indian history, zoos, allergies, climate change, sociology, scientists, nanotechnology, and education. In some cases, e.g. Indian history and zoos, it is easy to see the overlap between terminology on elephants and the collection. Collections where the CNN financials page was not recognized as being out of context were: networking basics, email, software history, Texas history, climate change, nanotechnology, and education. In this case, the technological focus of the financials page caused most technology-related topics to be similar.

In general, these results show that our heuristics could be used to focus the collection maintainer's attention on those pages that are highly unlikely to be reasonable replacements and away from the more obvious cases of topical cohesion. Depending on the application context, the risk/reward ratio of effort saved to making the wrong decision, a system could choose to automatically reject replacements moving more than 4 degrees away and automatically accept those moving towards by more

than 2 degrees—thereby making best use of the maintainers’ limited available time to determine the validity of the middle set.

5. LOCATING PAGE REPLACEMENTS

The Path Manager supports collection maintainers in visualizing the nature and level of changes to pages contained in their collections. When pages in a collection change its maintainer has the option of accepting these changed pages or of removing them from the collection [9]. However, in collections such as paths each page contributes to both the meaning of the collection and the continuity of the narration; removal of pages may break the flow of the path or render it semantically incomplete. Thus, removal of a page often requires the path’s creator to change the path in order to restore the path’s consistency by filling these gaps.

A path’s integrity also is threatened when page(s) that it points to cannot be located. The Path Manager informs the maintainer whether the page is unavailable due to a potentially temporary cause or a permanent one. Inability to retrieve pages from a Web site within a reasonable amount of time is likely to be a temporary situation due to heavy load experienced by the site. On the other hand, inability to find a Web site or a particular page on the Web site—the infamous 404 error, or the broken link problem—is likely to be a long-term issue as Web site managers and developers frequently redesign the structure of their sites [8]. Often Web sites themselves may cease to exist as users move between jobs, change service providers or let domain names expire. Currently the Path Manager has no mechanism to deal with the situation caused by pages that have been deemed unacceptable by the path maintainer, moved or disappeared. The path maintainers are left to fend for themselves to preserve the coherence of their collections.

The Path Manager has the potential to aid maintainers in finding suitable replacements for the invalid pages in their collections. This can be done by locating copies of the original page to replace the missing page or by providing a set of similar pages that can substitute for the original page.

5.1 Approach

Location of replacement pages is achieved in two distinct phases. In the first phase, the Path Manager extracts information about pages in the collection soon after the path is created, before any of the pages included in the path change. At this point, all pages in the path are in the state that the path’s author viewed when deciding to include them in the path. The Path Manager retrieves the page text and extracts phrases that are representative of this page. These keyphrases are used to locate page substitutes in the second phase, after some pages in the path have been deemed unacceptable or have become unreachable.

Our approach is similar to the one employed by Phelps and Wilensky to create robust or location-independent hyperlinks to Web pages [13]. Robust hyperlinks employ words that uniquely identify a page on the Web and use these to locate a Web page via a search engine. Phelps and Wilensky showed that about five words are often sufficient to accurately recall a page. We extend this work by using keyphrases rather than keywords. Our experience is that keyphrases are more effective than keywords in locating semantically similar pages in addition to the original pages focused on by Phelps and Wilensky.

5.1.1 Keyphrases

Following Turney’s lead, we define keyphrases to be one, two or three word phrases with no interceding stop words or punctuation marks [15]. As we are dealing with HTML documents, we further restrict these to be phrases that are not separated by intervening HTML tags. While Web page creators may often highlight important phrases by making them bold, underlined or italicized, it is unlikely that they will tag only a part of an important phrase. For example, it is unusual to render only a part of an important phrase in bold. However, we do make an exception for <A> and <LINK> tags that may occur within phrases. Sometimes, page creators may link words within a specific phrase to other pages that discuss a related or a more general concept. Similarly, parts of phrases may also be linked to their dictionary entries.

We also restrict the keyphrases to specific parts of speech. In this work we have focused upon keyphrases that contain an optional adjective followed by one or more nouns. Thus, these phrases are of the form *noun*, *noun-noun*, *adjective-noun*, *noun-noun-noun*, or *adjective-noun-noun*.

5.1.2 Keyphrase Extraction

The keyphrase extraction phase comprises of three operations:

- *Retrieve the page content.* The HTML source of all pages in the path is retrieved from the Web. The Path Manager analyzes this source to create a document signature [9]. In addition to page signature generation, this source is now used to extract keyphrases from the pages.
- *Tag the content with a part-of-speech tagger.* In this step, each of the words in document is appended with a tag identifying the part that it plays in the sentence. The HTML tags within the document are ignored.
- *Extract phrases that match the expected patterns.* We identify all phrases that match the criteria specified earlier from the tagged page contents and create a list of these phrases. The phrases in this list are used for further refinements and the rest of the document is discarded. While these phrases are indicative of the document contents we need to pick those that are representative of the document when searching for similar or exact page replacements. Due to the differences in strategies for locating similar and exact replacements for pages the phrases are stored in independent keyphrase lists that are ordered by their weight relative to their future use. The system employs Web search engines to calculate the rarity of phrases among Web documents, which in turn is critical to evaluating the effectiveness of the term when retrieving replacements for the page. While the path maintainers may use any search engine, it is recommended that they use the same search engine for generating keyphrase lists and for finding replacement pages.

5.1.3 Locating identical pages

Recalling a specific page from the Web involves choosing phrases that help discriminate this page from others on the Web. We use a TF-IDF-based measure to order the list of phrases identified in the earlier stage. The term frequency (TF) conveys the commonness of the phrase within the document and the inverse document frequency (IDF), the rarity of documents that contain this phrase within the corpus of documents. To calculate the IDF we divide the total number of documents that a given search engine indexes by the size of the result set returned when querying for the phrase.

While the number of indexed documents as well as the size of the result set may change over time, we have experimentally confirmed that typically these changes do not significantly affect the calculated IDF's. The keyphrase-list for exact page retrieval is ordered in the decreasing value based on the TF-IDF measure.

When the existing page is deemed unacceptable by the path maintainer the keyphrase list generated in the earlier phase is used to search for any copies of the page (in its original form) that may exist on the Web. The system allows its users to specify the number of initial keyphrases to search with and the maximum number of search results desired, then searches the Web via a search engine. If the set of search results returned is larger than the expected size this indicates that the query needs to be more specific. The algorithm adds the next keyphrase from the list and returns the result for this query. The result set is returned to the user when the results generated are fewer than the user had specified. While searching for a particular page, the algorithm gradually tightens the query to restrict the size of the result set returned. When a Web page has ceased to exist and no other copies are available the search may not return any results.

Location of similar pages may be used either as a backup measure in this case or as an optional feature to help path authors locate Web content that is somewhat alike the pages on their paths. This feature uses the keyphrase-list for similar pages.

5.1.4 Locating similar pages

The rarity of a document on the Web, while an asset when searching for the exact page, may actually hinder the search for similar pages. For example, a misspelled word in a phrase may render the document unique amongst others on the Web. While this misspelling is a boon when locating exact copies of a page it may yield no results when searching for similar pages containing this phrase. To account for conditions that may accidentally introduce such rarity, we weed out phrases that have occurred below a certain threshold within the document. The remaining phrases are then ordered in the descending TF-IDF-based value to create a keyphrase-list for similar page retrieval.

To find the closest possible set of matches for a given page the system uses a different strategy while searching for similar pages. It begins with the most restrictive set of phrases and generalizes it until the query returns at least a specified number of results. The user may configure the initial as well as final size of the query—in terms of the number of keyphrases to use—and the minimum size of the expected result set. The algorithm begins with the most keyphrases and generates a result set. If this result set is smaller than the expected set size it removes the least significant phrase (the phrase with the least TF-IDF weight) and reruns the query on the search engine to generate a broader result set.

5.2 Evaluation

We tested the performance of these algorithms over a randomly selected subset of pages from existing paths on the Walden's Paths server. We used the same set of pages for retrieving exact copies as well as similar pages. The results returned by the algorithms were reviewed manually to test for the effectiveness of these algorithms.

The system returned at least one exact copy for each page in the test set. Where multiple page mirrors were found the system returned more than one potential page replacements. We verified

that the returned results were exact copies via visual inspection of the content, layout as well as the links that emanate from these pages and also compared the HTML source for each of the potential replacements with the original page included in the path. Overall, the system performed very well when searching for identical pages to replace a page in a path.

When searching for similar pages, the algorithm generated between three and ten similar pages over the duration of the test. In a few cases, the system returned some of the identical pages—pages that were suggested as exact page matches—among the top few matches. To focus solely on the similar pages we discarded the exact copies of pages while analyzing the results. In general, human evaluators had more disagreements about the accuracy and acceptability of these results when compared to the results returned for exact pages.

The system demonstrated no clear preference between one, two or three word phrases. Manual examination of phrase ranking in both the keyphrase lists indicates that majority of the keyphrases represent the central concept of the page.

Length of the document was a factor in the quality of the keyphrases and hence the accuracy of suggested page replacements returned by the search engines. Word frequency distributions for short documents—documents that consisted of only a few sentences—caused the rarest of phrases to appear higher in the list irrespective of the actual context of the document. This tended to cause a deviation in the results from what was expected. Hence it is essential that the algorithm work around the problem of extremely concise documents.

5.3 Discussion

Sometimes path authors may link to non-textual content such as images or audio files embedded within Web pages leaving the algorithm with no text at all to attempt to retrieve the resource. The lack of significant textual content in Web resources included in a path presents the Path Manager with a significant challenge when extracting information to recall these resources. The scarcity of textual content can be countered by using other attributes of the resources as well as the paths that include them. The annotations added by the path creator offer a starting point. Keyphrases extracted from the annotations may be exploited to locate similar Web pages. In some cases, the structural layout of the page, or other meta-information tags included by page authors may help in searching for it more accurately. The filename of images or videos provide another attribute for locating these resources.

Efficiently retrieving or computing the TF scores for keyphrases is a challenge. While the term document frequencies for over 30 million individual terms (words) for about 51 million Web pages are readily available in the Berkeley Digital Library's Web Term Document Frequency and Rank page (<http://elib.cs.berkeley.edu/docfreq/index.html>) no such index exists for phrases or combinations of terms. The term document frequencies for phrases can be obtained by querying search engines and computing the result size but this method results in network load and is impractical when processing large documents.

Determining whether a page is similar to another is more complex than deciding whether two pages are identical due to the different metrics for gauging similarity. These similarity measures have to

do with the attributes of the pages themselves, attributes of their context and connections with respect to the wider hypertextual net that includes them, and finally, with respect to their use in the paths. Some of the similarity measures that jump to mind are the textual content of Web pages, embedded objects such as images, and their visual layout, that is the relative positioning of the page elements. The links between pages as well as the reliance of the page on these links—whether the page is text-heavy or link-heavy—also plays a role in determining the similarity of pages. Sugiyama et al., argue in favor of including contents as well as characteristics of linked pages to increase the accuracy of Web-based information retrieval [14].

From a human perspective the role that the page plays within the path, or its context of use, also affects the similarity of pages. Similarity, in this context, represents more of an acceptability measure than a statement of textual or semantic similarity between two pages. Finally, path maintainers may emphasize some of these attributes more than others, leading to a difference in opinion about whether any two pages are similar enough that one can replace the other in a path.

Our current approach returns pages based on the textual similarity alone. We do not yet determine the potential for acceptance as a replacement page when considering similarity of pages. This limitation is, in part, due to the human element that is inherent in the path authoring activity. We do not yet have a model of intentionality for path creation. In other words, we lack the means to programmatically represent or, at a later stage, determine the authors' intentions for including a page in the path. The annotations added by authors may, in some cases, provide an insight into their intentions, but the significance of using annotations for determining purpose is as yet unclear.

When searching for exact replacements, the system returned the desired pages when searching with five or six keyphrases. This behavior validates the observations and results presented in [12,13] when using keyword-based algorithms. Our experiments extend these observations to keyphrase-based algorithms.

6. CONCLUSIONS AND FUTURE WORK

Motivated by the work mentioned above, Path Manager is going through a redesign phase that includes three major considerations: integration of context-based algorithms; augmentation of the system with mechanisms that address the re-location and disappearance of integration of Web pages; and general improvements to cope with changes on the Web.

The integration of context-based algorithms is not just to add them as an alternative to the content-based algorithms, but also to combine them into hybrid algorithms that take into consideration both content and context. This approach requires modifications to the user interface and system implementation and additionally requires changes on the general perspective about the system as a whole. One philosophical change in the design is a shift from *paths* into a more general concept of *collections*. We foresee this as a necessary step that will facilitate defining context in accordance with the kind of collection. This is critical for the context-based algorithms to work properly.

The design of the new Path Manager considers the use of the mechanisms we devised for responding to the inevitable problems of page relocation and page disappearance. By solving the

problem of relocating the Web pages these mechanisms benefit collection managers by aiding in the preservation of collection integrity and consistency.

The algorithms for keyphrase generation can be extended along several axes to improve the efficiency of page retrieval. Broadening the characteristics of keyphrases, for example, keyphrases with multiple adjectives (adjective-adjective-noun form), verbs and stop words are possible candidates for focusing our efforts. Phrases that Web page creators visually distinguish via text formatting are also promising candidates. Use of other page and resource attributes may aid in improving the retrieval accuracy for short Web pages or Web page components included in the paths. Inclusion of page annotations for extracting keyphrases and for readjusting the weights of keyphrases from the pages may also present effective alternatives.

We are exploring mechanisms to generate document frequencies for phrases efficiently and accurately. While these are not directly available yet and are susceptible to change due to the fluidity of the Web a phrase frequencies list may serve as an important resource to various projects that include text-processing features.

Finally, the design of the next version of Path Manager addresses other kind of changes in the Web: changes in standards and authoring practices. Time passes very fast on the Web, affecting the underlying standards, technologies and people using it. Since Path Manager was developed, HTML (and now XHTML) has evolved. Web page authoring techniques have become more advanced, changing the way pages are written and designed, whether it is manually or using a WYSIWIG authoring environments. These new developments on the Web often arise as improvements over the previous state of affairs. Consequently Path Manager also needs to evolve in order to cope and take advantage of the new features of the Web.

7. ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation grants IIS-0219540 and DUE-0121527.

8. REFERENCES

- [1] Ashman, H. (2000). "Electronic Document Addressing – Dealing with Change," ACM Computing Surveys 32, pp. 201-212.
- [2] Brewington, B. & Cybenko, G. (2000). "How Dynamic is the Web," Proceedings of WWW9 –9th International World Wide Web Conference (IW3C2), pp. 264-296.
- [3] Broder, A., Glassman, S., Manasse, M., & Zweig, G. (1997). "Syntactic Clustering of the Web," Proceedings of WWW6, cited in Fetterly, D., Manasse, M, Najork, M., Wiener, J. (2003). "A Large-Scale Study of the Evolution of Web Pages," Proceedings of WWW03.
- [4] Bush, V. (1945). "As We May Think," The Atlantic Monthly, (August 1945), pp. 101-108.
- [5] Chakrabarti, S., Dom, B. E., Gibson, D., Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). "Mining the Link Structure of the World Wide Web," IEEE Computer, 32, 8, pp. 60-67.

- [6] Cho, J. & Garcia-Molina, H. (2000). "The Evolution of the Web and Implications for an Incremental Crawler," Proceedings of VLDB 2000, pp. 200-209.
- [7] Fetterly, D., Manasse, M., Najork, M., & Wiener, J. (2003). "A Large-Scale Study of the Evolution of Web Pages," Proceedings of WWW03, pp. 669-678.
- [8] Fogg, B., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P., & Treinen, M. (2001). "What Makes Web Sites Credible? A Report on a Large Quantitative Study". Proceedings of the SIGCHI conference on Human factors in computing systems, March 2001, Seattle, WA, ACM Press, New York, NY, 61-68.
- [9] Francisco-Revilla, L., Shipman, F., Furuta, R., Karadkar, U., & Arora, A. (2001). "Perception of Content, Structure, and Presentation Changes in Web-Based Hypertext," Proceedings of Hypertext, pp. 205-214.
- [10] Furuta, R., Shipman, F., Marshall, C., Brenner, D., & Hsieh, H. (1997). "Hypertext Paths and the World Wide Web: Experiences with Walden's Paths," Proceedings of Hypertext'97, ACM Press, pp. 167-176.
- [11] Johnson, D.B. & Tanimoto, S.L. (1999). "Reusing Web Documents in Tutorials with the Current-Documents Assumption: Automatic Validation of Updates," Proceedings of EDMEDIA 99, AACE, pp. 74-79.
- [12] Martin, J.D. & Holte, R. (1998). "Searching for Content-Based Addresses on the World-Wide Web". Proceedings of the third ACM conference on Digital Libraries, (Pittsburgh PA, June 1998), ACM Press, New York, NY, 299-300.
- [13] Phelps, T. & Wilensky, R. (2000). "Robust Hyperlinks: Cheap, Everywhere, Now," Proceedings of Digital Documents and Electronic Publishing 2000 (DDEP00), Munich, Germany, 13-15 September 2000.
- [14] Sugiyama, K., Hatano, K., Yoshikawa, M., & Uemura, S. (2003). "Refinement of TF-IDF schemes for Web pages using their hyperlinked neighboring pages", Proceedings of the Fourteenth ACM conference on Hypertext and Hypermedia (Nottingham UK, August 2003), ACM Press, pp. 198-207.
- [15] Turney, P. (2000). "Learning Algorithms for Keyphrase Extraction". In Information Retrieval 2(4), pp. 303-336.