

# Perception of Content, Structure, and Presentation Changes in Web-based Hypertext

Luis Francisco-Revilla, Frank M. Shipman III, Richard Furuta, Unmil Karadkar, Avital Arora

Center for the Study of Digital Libraries and Department of Computer Science

Texas A&M, College Station, TX, 77843-3112, USA

+1 979 845 4924

{l0f0954, shipman, furuta, unmil, avital}@csdl.tamu.edu

## ABSTRACT

The Web provides access to a wide variety of information but much of this information is fluid; it changes, moves, and occasionally disappears. Bookmarks, paths over Web pages, and catalogs like Yahoo! are examples of page collections that can become out-of-date as changes are made to their components. Maintaining these collections requires that they be updated continuously. Tools to help in this maintenance require an understanding of what changes are important, such as when pages no longer exist, and what changes are not, such as when a visit counter changes. We performed a study to look at the effect of the type and quantity of change on people's perception of its importance. Subjects were presented pairs of Web pages with changes to either content (e.g., text), structure (e.g., links), or presentation (e.g., colors, layout). While changes in content were the most closely connected to subjects' perceptions of the overall change to a page, subjects indicated a strong desire to be notified of structural changes. Subjects only considered the simultaneous change of many presentation characteristics as important.

## Keywords

Walden's Paths, perception of change, fluid documents, managing fluid Web pages, meta-document.

## INTRODUCTION

The Web presents a rich and vast collection of information resources authored and published by a multitude of sources. In this collection it is possible to find information on virtually every topic. Even when referring to a single topic, it is often possible to find multiple sources of information varying in many aspects such as point of view, validity, semantics, rhetoric, and contextual situation. When exploiting this body of information people often refer to more than one of these sources. This has prompted the development of systems capable of aiding readers by providing an interpretation of Web pages.

Walden's Paths is an application that allows creation of trails or paths [3] using pages authored by others [14, 6]. The paths represent a higher order construction, or meta-document, that organizes and adds contextual information to the original pages. In order to provide the intended functionality, meta-documents need to address the issue that their fundamental building blocks (Web pages) are subject to unpredictable changes. This is particularly important given the high fluidity of Web pages [2], where authors are prone to change their pages just to maintain a "fresh" look and feel.

Meta-documents that are to have a lasting value need to adapt to their component changes. Hence, changes must be detected. One way to identify changes is to compare cached versions of the pages with the current version available on the Web. This comparison could provide a Boolean result of whether the page or sections of the page has changed. However, this simplistic detection scheme is not enough, since changes might be minor or irrelevant like a typographic correction or a minor cosmetic modification. Instead, an approach that measures the degree of change seems necessary. This means being able to locate changes and identify those that are meaningful. But what changes do people find meaningful? It is answering this question that is the focus of this paper.

The next section explains the problem and our motivation in greater detail. Then, we describe a study designed to determine what changes people consider relevant and how difficult these changes are to perceive. We then present the results of the study, emphasizing how the type and amount of change impact its perception. Subsequently, we discuss the results, their potential explanations, and their implications. We continue with a description of Walden's Paths Path Manager and how the observations obtained during the experiment informed its design.

## PROBLEM

Every document has some degree of fluidity and some degree of fixity [11], although, some documents tend by nature to be highly fluid. This is typically the case of digital documents such as Web pages. This fluidity rapidly became apparent in the Walden's Paths project. Paths would frequently become out of date within months of being authored due to pages that moved, changed or were no longer available [13]. As a result there was a constant necessity for project members to check for changes, revise paths that had changes, and retire paths that were no longer

viable. This maintenance of the collection of paths took more resources as the collection grew.

Fluidity by itself was not the problem. The fluidity of Web pages can be very beneficial. Authors and readers take advantage of the ease of generating and retrieving up to the minute information updates and error corrections. However, the frequent changes made to Web pages increase the complexity of managing digital document collections that include these resources. While we will discuss this problem in terms of Walden's Paths, this is also a problem, on a smaller scale, for individuals maintaining their bookmarks and, on a larger scale, for categorization-based Web portals like Yahoo!

Presently, when changes occur, people assess the relevance of the change in order to decide how to react. Hence, the relevance of the change is influenced by the human perception of the change. In fact, "relevance" is a highly subjective, abstract, and contextual concept. Different individuals can disagree in their relevance judgment about any given change. Nevertheless, as difficult as it might be, relevance assessment is still necessary.

To ease this situation, it is required to find approaches that aid in assessing the relevance of changes. In the case of Walden's Paths, one approach that addresses the fluidity issue and expedites delivery is caching the Web pages. This approach is similar to the caching functionality provided by CiteSeer [7] and AltaVista [1]. This approach alleviates the issue but does not solve it, since some changes are desirable. Some paths include pages that change by nature, such as the current weather forecast or the headlines from a local newspaper.

Therefore, selective caching and retrieval can improve this approach. The Walden's Paths Authoring Tool allows authors to specify the caching strategy for individual pages [10]. Nevertheless, caching Web pages created by third parties brings up issues of versioning, intellectual property, and information expiration.

Walden's Paths' ephemeral paths provide an alternative approach to this problem [5]. Ephemeral paths are created as the result of some computation and exist for only a short period of time. This mechanism can be applied to the problem of coping with changing pages by selecting only the pages that have not moved or changed as part of the ephemeral path. Unfortunately, this solution results in paths that are no longer very useful. The reason is that paths often have a rhetorical coherence based on the linearity imposed by the path author. Hence, when pages are removed from a path, the rhetorical structure can be broken, diminishing the path's effectiveness and undermining the author's original intent.

A better solution is to create tools that facilitate an automatic reaction to change in an intelligent manner by taking into consideration the relevance of the changes. In order to inform and evaluate the approach and design of tools that aid in the automatic assessment of changes to Web pages there is a necessity to better understand what people consider important. With this goal in mind, the following experiment was conducted to determine what

changes people consider relevant and how people perceive changes to Web pages.

## **THE STUDY**

The study was designed to evaluate the ease of perception and assessed importance of changes as the type and quantity of the change was varied.

### **Kinds of Change**

One issue when studying the human perception of changes is that there are many types of change to Web pages. Thus it is important to classify the nature of changes in order to better understand the results. A possible taxonomy for changes is to classify them as content, presentation, structural, and behavioral changes.

Content changes refer to modifications of the page contents from the reader's point of view. For example, a Web page that originally contains information about a company's soccer tournament might be continuously updated as the tournament progresses. Later, after the tournament has ended, the page might change to a presentation about sports injuries or the upcoming company picnic.

Presentation changes are changes related to the visual encoding of material that do not change the material conveyed. Examples of presentation changes are different fonts, backgrounds, styles of bullets, or changing the layout of content.

Structural changes refer to the underlying connection of the Web page to other Web pages. As an example, consider changes in the link destinations of a "Weekly Hot Links" Web page. While this page might be conceptually the same each week, the fact that the destinations of the links change might be relevant, even if the anchor text of the links has not. Structural changes may not result in any visually perceptible change to a Web page except when the mouse lingers over the links.

Behavioral changes refer to the modifications in scripts, plug-ins and applets. The consequences of these changes are harder to predict, especially since many pages hide the script code in other files.

### **The Source Materials**

Due to the complexity of behavioral changes, this experiment focuses on the perception and evaluation of changes in content, structure, and presentation.

Web pages used in the study were selected from paths previously created by teachers and from personal bookmark lists. A second version of each page was created to reflect a single type of change, and was classified by the type of the change and a quantitative measure of its magnitude.

Content changes included modifications to the text presented. Structure changes referred to modifications of URLs and the anchor text of links. Presentation changes considered changes to fonts, spatial positioning of the information, background colors, navigational images and drastic changes (combinations of many presentation changes).

The quantitative measures of change were calculated based on the percentage of units modified in the document. For instance, a content change of 50% would imply that three

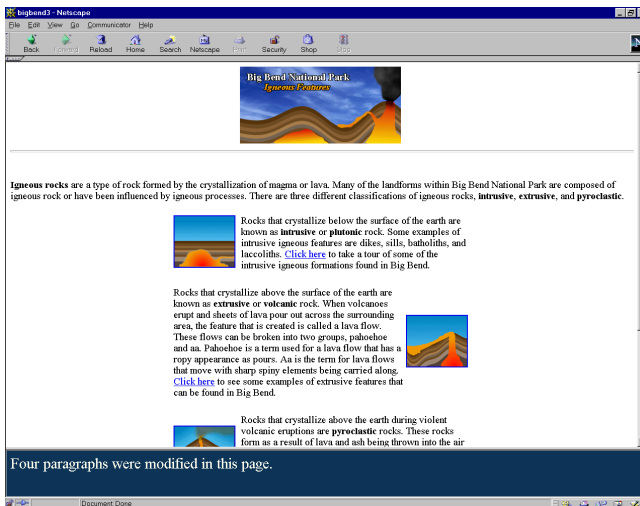


Figure 1: Example of a Web page pair in phase I with content changes to all four paragraphs.

out of six paragraphs in the document were modified. While these metrics are arbitrary they provide an objective measure we expect to be correlated to the human perception and assessment of change relevance.

### Participants

In order to conduct the study, we recruited 18 adults. Subjects were students at Texas A&M pursuing studies in fields ranging from sciences to humanities, and levels from undergraduate to post-doctoral. Subjects were randomly assigned to two groups, A and B.

### Methodology

The experiment consisted of simultaneously presenting the participant with two versions of a Web page, the original version and a possibly modified version (some were unmodified copies of the original). Figure 1 shows an example pair of pages from the study. The participant was then asked to evaluate the change magnitude, as will be described.

The goal of the study was to address the following three questions:

1. Do people view the same changes in a different way when given different amounts of time to analyze the pages?
2. What kinds of changes are easily perceived?
3. Of what kind of changes do users want to be notified?

Before the evaluation we familiarized the participants with the testing software and our definitions of the three types of change. Additionally, we provided a context for evaluating the changes in the Web pages by providing a scenario. The participant was asked to act the role of the Information Facilitator in a K-12 school (Kindergarten to High School). In this scenario, teachers had chosen pages from the Web for use by their classes and it was the participant's responsibility to notify the teacher of important changes.

The first task in the study was to fill in a pre-evaluation questionnaire collecting demographic data about the participants and their computer literacy. The study was then divided into three phases:

1. In phase I, the participant was given 60 seconds to view each Web page pair. This phase consisted of 8 pairs of pages, 3 with content changes, 3 with structural changes, and 2 with presentation changes. The system also provided a textual description of the change to the participant. This phase also provided training for the next phases where no textual description was provided. Figure 1 shows an example of two pages presented with the textual description of the change.
2. In phase II, the participant was only allowed to view the page pairs for 15 seconds before evaluating them. There were 31 pairs of pages in this phase.
3. In phase III, the participant was given 60 seconds to view the Web page pairs. There were 31 pairs in this phase.

Phase I aims to answer what types and quantities of change result in a person wanting to be notified of that change. Phases II and III address the questions of what changes are easily perceived, and how this perception changes given more time. They also inform about the importance of perceived changes.

To balance the difficulty of locating specific changes among the subjects, Web pages in phase II for group A were shown to group B during phase III. Similarly, Web pages in phase III for group A were shown to group B during phase II. This allowed comparing the results for the same page when given 15 or 60 seconds for observation.

For each of the 70 pairs of Web pages viewed, the subject was presented with five questions:

1. From the Content perspective, the degree of change is:

- From the Structure perspective, the degree of change is:
- From the Presentation perspective, the degree of change is:
- Overall, how significant are the changes?
- If this page were in my bookmark list, I would like to be notified when changes like these occur.

Questions 1-4 were answered in a 7-stop scale ranging from “none” to “moderate” to “drastic”. Question 5 was answered in a 7-stop scale ranging from “strongly disagree” to “strongly agree.”

The final task in the study was to fill a post-evaluation questionnaire that gathered the participant's general comments about how easy it was to assess the magnitude of changes in Web pages, and about the evaluation software.

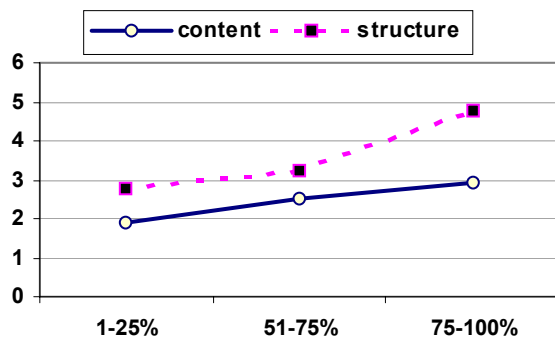
In all phases, the testing software controlled the presentation of the Web pages directly. The time required to record the participant’s answer was not controlled. The total time for the study varied from 1 to 2 hours per person, depending on the time the participants took in filling in the answers, and whether they decided to rest between phases.

## RESULTS

Results for the content and structure changes are presented as graphs. The X-axis represents a quantitative metric of change, such as percentage of paragraphs changed, while the Y-axis represents the human perception of change.

### Phase I Results

During this phase, subjects were provided a textual description of the changes and a minute to look at each pair of pages. As a result, they could evaluate all the changes rather than just those changes they noticed. This phase addresses the questions of what types and quantities of change are considered significant and which changes users would like to be notified of.

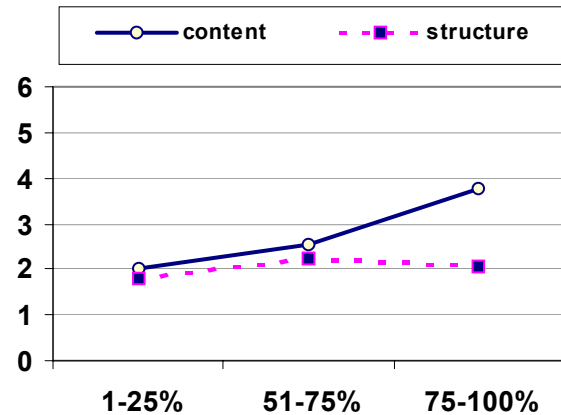


**Figure 2:** Notification desire in Phase I

Figures 2 and 3 show the results of phase I for low (1-25% change), medium (51-75%) and high (75-100%) change magnitudes (to keep the time needed to complete this phase manageable, pages with 0% change and 26-50% change were excluded from the phase’s test set). Each point is the average of 18 assessments of one Web page pair. The participants assessed the changes based on both the appearance of the two versions of the page and the textual

description of the change. Figure 2 shows the average of all participants to the question regarding notification. For both content and structure changes, there is a clear increment as the change magnitude increases, although the structural changes led to greater desire for notification.

Figure 3 shows the average evaluation of the overall perception of change. The subjects’ evaluation of the content change shows a clear increase as the magnitude of change increases. The perception of structure change shows an almost horizontal behavior.



**Figure 3:** Overall change perception in Phase I

### Content Changes

Content changes obtained from phase II and phase III are shown in Figure 4. Each point represents the average of the assessments for all the Web page pairs with that particular magnitude of change and with that particular type of change. There were three Web pages corresponding to each combination of magnitude and type of change. These were evenly split between the 15 second and 60 second conditions so each point represents the average of 27 assessments. This is also true for Figure 5.

Looking at Figure 4 it is possible to observe that as the magnitude of modifications increased, so did the perception of the change in content. The perception of overall change and desire to be notified appear to follow the perception of content change. Additionally, there is a clear gap between these and the perceived level of structure and presentation change.

The lines representing the answers for 15 seconds (solid lines) and 60 seconds (broken lines) show a similar behavior. The extra 45 seconds did not drastically change subjects' perceptions of the changes.

There is a plateau in the middle of the graph where the assessments for content change, overall change, and desire for notification can be seen to be the same or slightly higher for the page pairs in the 26-50% change range than those in the 51-75% range. At the same time, the presentation and structure change assessments moved slightly higher.

### Structure Changes

The results when structure was modified can be seen in Figure 5. The graph shows that the amount of time given to subjects was critical to their assessment of change—

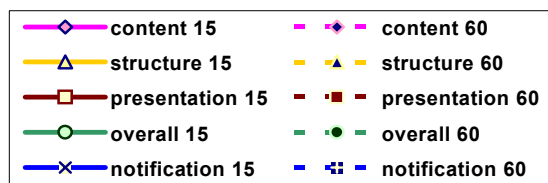


Figure 4: Content changes

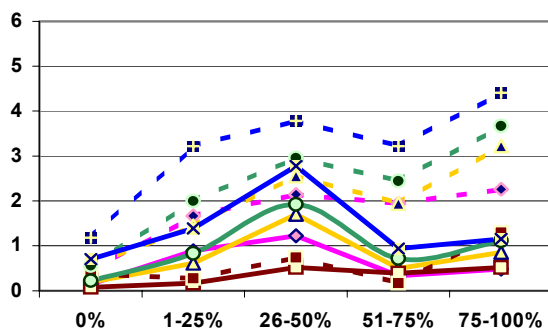


Figure 5: Structure changes

indicated by the dotted lines being higher than the solid lines. Only the line for presentation 60 remains relatively low.

As the quantitative measure of structure change increases (% of links changed), the perception of the change goes up rapidly at first and then levels out in the 60-second case. Note that the lines for perception of content change tend to follow the line of perception of structure change.

This graph shows a pronounced drop off in the perception of change between the 26-50% and the 51-75% pairs for the 15 second case. This could be due to an uncontrolled variable (document length, number of links, etc.) Longer pages in the 51-75% set would mean that subjects would spend more time comparing content and not have time to locate changes in links in 15 seconds. These issues are orthogonal to the conclusions that we draw from the experiment; discussed later in this paper. However, a follow on study should be performed to determine impact of length of page on determination of change.

#### Presentation Changes

Presentation changes are more difficult to measure on a single continuous dimension. Therefore, the study only covered the previously mentioned presentation change subcategories, namely fonts, spatial positioning,

background, navigational images, and drastic changes. The results of these subcategories are shown in Figure 6.

The results show that changes to fonts were not perceived as changes. There is little difference between the 15 and 60 second results.

Changes in the screen position (or layout) resulted in higher scores for notification desire than for perception of change. An example of this type of change is rearranging the cells in an HTML table used for alignment/layout purposes. Looking at the adjacent vertical bars in Figure 6 it is possible to note that there tended to be a drop in perceived change, particularly for content change, when subjects moved from 15 to 60 seconds to evaluate the change.

Background changes, such as color changes, were detected easily as noted by the two middle bars, which correspond to the perception of presentation change. However, participants did not care to be notified about them.

Similarly the score is low for the desire of notification when changes to navigational images occur. These changes included modifications of images for bullets, bars, etc. Again, there was a drop in the perception of change as subjects had more time.

Drastic changes in presentation were those having multiple types of presentation change between the pairs. This included Web sites that have a text only version, and a text and image version. The paragraphs and the links were the same. When faced with drastic changes, the participants perceived changes on every change category, even though the content and structure remain the same. Additionally, the perception of overall change and the desire for notification obtained high scores.

#### DISCUSSION

The results presented help answer some of our original questions while also creating new questions requiring further research. This discussion focuses on explanations for and implications of the unexpectedly high desire to be notified of structural changes, the differences between the perception of structure change and content change, the effects of increasing the amount of change and the time available to perceive the change, and the perception of presentation changes.

#### Notification of Structural Changes

One unexpected result from this study is the high desire to be notified of structural changes.

Phase 1 provided subjects with a description of the differences between the pages, meaning they knew the extent of the changes but not necessarily the specifics of the change. Figure 3 shows that the perception of overall change is more dependent on the amount of content change than on the amount of structural change. Intuitively, this is in contrast with the results of Figure 2, where the degree of structural change had the larger impact on whether the subjects wished to be notified of the change.

When not told how many links had changed, subjects wished to be notified of the structural changes they found. Figure 5 shows a high notification value relative to the other measures of change. One reason for this desire for

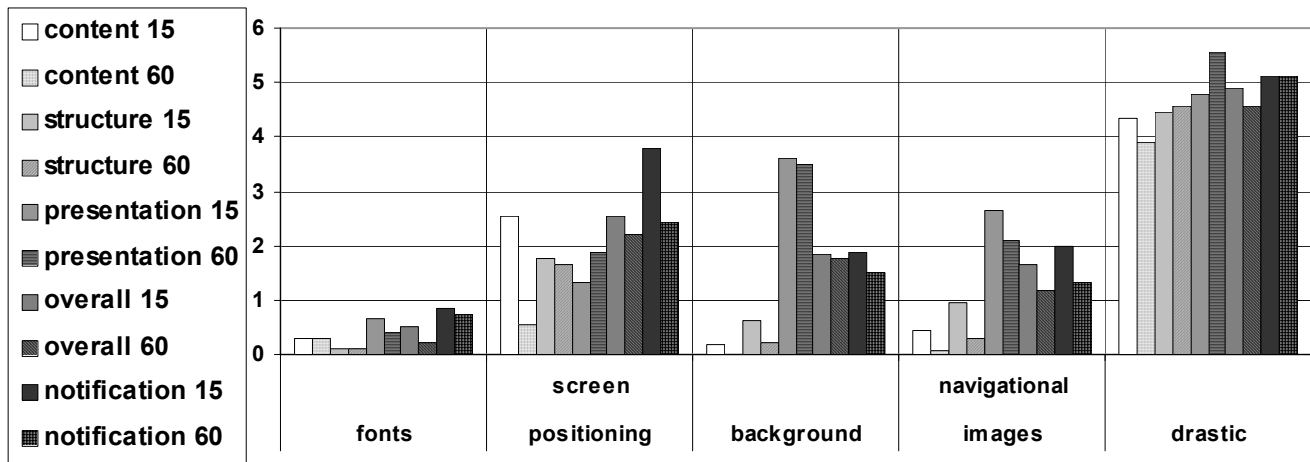


Figure 6: Presentation changes

notification is the difficulty of recognizing structural changes with only a casual glance at a page. Subjects found the effort required to identify link changes in 15 seconds “*very difficult, as there was not enough time!*” When given 60 seconds a subject commented:

*“(It is) easier but still the task of finding faulty links would require more time and it is something that would have a much more effect on whether the person wanted to be notified or not. ‘Cos nobody wants a page with different links.”*

#### Effect of Time

The effort required to recognize structural changes can be seen clearly in the results of giving subjects more time with each pair of pages. Figure 4 shows that the increase in time did not cause much of a change in the case of content changes. Presumably, these changes were more visually obvious and subjects developed methods for quickly locating these changes, such as noticing changes in word wrapping for paragraphs.

In contrast, time was critical for locating structural changes because they are not visually apparent and subjects have to move the mouse over links to search for the changes. When pages have lots of links and subjects found some changes, the subjects may have presumed there were potentially many more.

While not a large effect, providing subjects with more time reduced the perception of change for presentation changes, except in the case of drastic changes. This was especially true for screen positioning and navigational images. The extra time let subjects compare the pages and realize that the content and structure were the same.

#### Content and Structure

This study required subjects to both locate and evaluate the degree of content and structure change. The similarities and differences between subjects’ assessments of these types of change are interesting.

Perception of content change is highly correlated with perception of overall change and desire to be notified of the

change, as seen in Figure 4. Content changes were easily identified and were not perceived as structure or presentation changes (with the possible exception of large changes.) On the other hand, assessing the relevance of content changes seemed to require more effort than the other types of change. One participant commented that:

*“Content based changes required much more time to decide on the magnitude of change—for example deciding on the relevance of the deleted/altered text required time.”*

Figure 5 shows that subjects classified structural changes largely as structural changes and, to a lesser extent, as content changes. When asked about the difficulty of classifying the change type, some subjects found that “*it was fairly easy to determine what type of change it was*”, while other subjects expressed that “*it was relatively easy (easier) to identify content and presentation changes than structure changes*” and that there was “*ambiguity between content and structure*”.

This ambiguity could be due to text anchors for links changing, but also might be a result of the role of the page—when a page is primarily acting as a resource for browsing other pages, then its links may be its content.

#### Effect of Quantity of Change

As the quantity of change increased, subjects perceived the changes to be of all types. Figure 4 shows that large content changes were perceived as changes of structure and presentation as well. A similar trend appears for large presentation changes. As seen in Figure 6, subjects perceived drastic presentation changes as changes in content and structure.

Figures 4 and 5 also show that the perception of overall change and the desire of notification increase along with the quantitative metrics used in this study for the estimation of the magnitude of the changes. The perceived change increases rapidly at low levels of change and levels off at the high end.

## Presentation Changes

Besides cases where there were drastic presentation changes (cases of complete visual redesign), other presentation changes were perceived as small and not important to be notified of. Fonts were ignored—perhaps because browsers allow the user to change these so there is no expectation they are fixed. Changes in screen positioning were somewhat important, particularly in the 15 seconds case, when subjects might not have enough time to determine that all the content was the same. Similarly, changes in navigational images were perceived as more relevant in the 15 seconds case than in the 60 seconds case.

Presentation changes were easier to detect than other kinds of changes. One subject mentioned that:

*“Presentation changes were easy to find. The other changes required the user to play around with the links and read the articles.”*

This early detection of presentation changes provided the subjects with more time to evaluate their relevance. Therefore it is interesting that drastic changes still were perceived as changes in all other categories. It would be interesting to further investigate when combinations of small presentation changes start to be perceived as a drastic change.

## IMPLEMENTATION OF A PATH MANAGER

As mentioned in previous sections, the motivation for the study was to inform and evaluate the approach and design of tools that assist maintaining paths, which are collections of Web pages. Consequently, the observations made during the study were used to guide the design and development of Walden's Paths' Path Manager.

We give a brief overview of the Path Manager in this section. More complete details may be found in a companion paper [4].

### From the study to a design approach

The goal in developing the Path Manager is to assist a collection's maintainer in discovering when relevant changes occur to Web pages in paths. Although determining the relevance of changes to a document can only be done in the context of how that document is to be used, the study suggests how measures of content, structure and presentation change might be combined. Yet this still requires methods for determining the magnitude of the various categories of change.

The approach taken is to obtain concise document metrics that can be used heuristically to compare versions of an HTML page and quantify the magnitude of the changes. In conjunction with the URL, these metrics describe the document and are stored as the document's "signature." Notice that these document signatures are not the same as the "signature files" used in Information Retrieval (as in Witten, et al. [15]), which generally refer to strings of bits created by hashing all the terms in the document.

HTML documents have many candidate features that could be used for the assessment of content, structure and presentation changes. However, a goal of the Path Manager was to find a small set of features that could be computed,

stored and used efficiently. Therefore, in the current implementation, the system computes the signature over only five features for each document version, namely: paragraph checksums, headings, links, keywords, and a page checksum.

- *Paragraphs Checksums.* These are used primarily to measure content changes. The inclusion of individual paragraph checksums provides a finer granularity than page checksums.
- *Headings.* Since headings typically highlight important text and titles, changes to headings may indicate relevant content changes. In addition, headings also affect the presentation of the document, by making the selected text more prominent than the body of text.
- *Links.* Links provide measures of structural change that strongly impacted our subjects' desire to be notified in the study.
- *Keywords.* The presence or absence of keywords in a page is usually helpful in identifying relevant content changes. The current implementation of the Path Manager uses reader-provided keywords. Keywords provided by users have been found good at distinguishing relevant pages from non-relevant pages [12]. The question remains as to whether users will provide them. Therefore we are considering evaluating different algorithms for automatic keyword generation.
- *Page Checksum.* This is used to detect possible changes that do not affect any of the other metrics, helping to differentiate *no changes* from *unknown changes*.

The computation and use of these metrics for content, structure and presentation characterization of a Web page is heuristic by nature. There are many different ways to compute and combine these Web page features in order to infer the relevance of the changes between two versions of a document. Thus, we are investigating the two different algorithms in the implementation of the Path Manager. These algorithms are the subject of the next section.

### Algorithm descriptions

As was just mentioned, the Path Manager currently provides two different algorithms for estimating relevance. Both of these algorithms are based on the previously described document signatures. The first algorithm is based on David Johnson's work [8]. The second algorithm is simply called the Proportional Algorithm due to the nature of its inner workings, as will be explained in the following paragraphs.

#### *Johnson's Algorithm*

In his original approach, Johnson [8, 9] used paragraphs, headings and keywords in order to compute a "distance" between documents. However, in order to take account of structural changes, the Path Manager implements an extension version of Johnson's algorithm that includes links as an additional metric. In Johnson's algorithm, each individual feature of a Web page is compared to its corresponding feature in the second page in order to discover and count possible differences. For each individual feature, the algorithm identifies three types of differences: additions, deletions and modifications. Each

type of difference is weighted individually and then added to the others, yielding a weighted feature change measurement. The estimation of the total change is similarly computed by weighted summation of each feature change measurement.

Since Johnson's algorithm allows the weight of additions to differ from the weight of deletions, the distance from document A to document B (or version A to version B) is not necessarily equal to the distance from document B to document A. In addition, the distance from document A to document B is unbounded, capable of taking any positive real value. Thus using this algorithm it is difficult to normalize and define change values for concepts such as "little change" or "much change".

#### *Proportional Algorithm*

This algorithm provides a simpler computation of the distance between Web pages that provides a normalized and symmetric measurement. This algorithm simply computes the percentage of the number of feature instances that changed and the total number of features found. The result is a normalized and symmetric value.

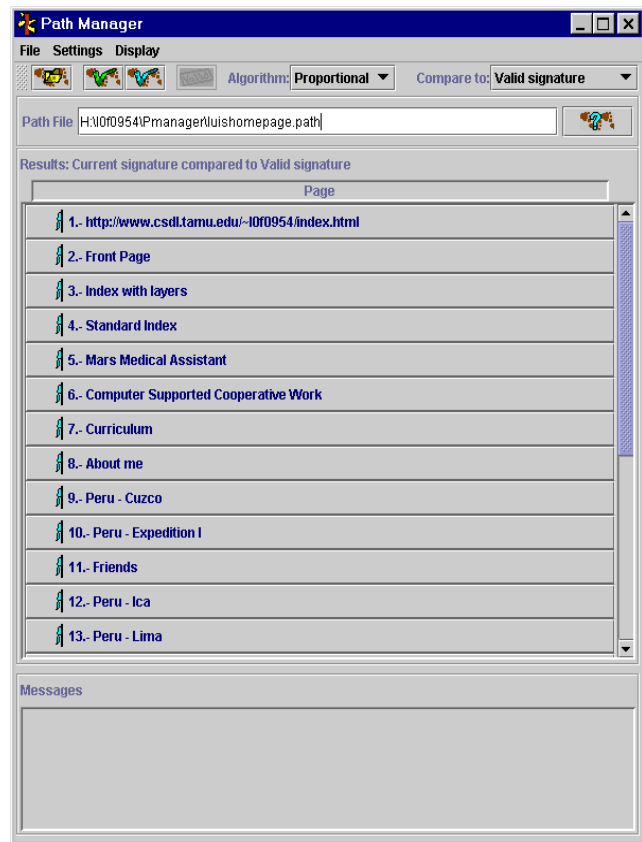
The normalized results makes it easier to compare different results. For example, if an algorithm adds some amount for each paragraph addition or deletion, the numeric value of change for long pages may be orders of magnitude higher than that for a short page in which every paragraph changed. By providing values within a range, users can more easily learn to interpret the results. The normalized results also allow for more effective visualizations. Values can be selected for identifying pages as "slightly changed" or "significantly changed".

#### **The System**

The Walden's Paths Path Manager is implemented in Java. Due to its Walden's Paths roots, the Path Manager is designed to take either Path files as input and then check all the pages specified in them. Alternatively, an HTML file can be specified as input, whether it is located locally (like a bookmark list) or remotely (specified by a URL). In the case of HTML files, every link considered as a page in the path to be checked.

The functionality of the Path Manager is straightforward. First, the system retrieves the page from the Web. It then parses the page contents and creates a new page signature. Next, the signature is stored in the path's signature file. Finally the new page signatures are compared with the previously stored signatures, and the system presents the users with an assessment of the relevance of the overall change of each Web page. At this point the user can review each page individually and validate the current state of the path.

Currently the Path Manager employs three versions of the signature to indicate the amount of recent and long-term change for a page. This signatures are the original, the last valid, and the latest time the path was checked. The original signature is the first signature obtained and is used to give a sense of the total amount of change since the path was first checked. As time passes and pages change, the user might update the signature, updating the valid signature. This



**Figure 7:** Path Manager initial interface

signature represents functional state of the page. Finally, the latest signature corresponds to the last time that the Path Manager retrieved and analyzed the pages, regardless of whether the user validated their state or not.

Initially the Path Manager shows a list of the pages in the path. This is shown in Figure 7. Pages are identified by either their title or, when the page title is not available, their URL. The system uses colors to represent the magnitude of the change. Pages are initially shown in blue, meaning that their current state is unknown as they have not been evaluated yet. The flag at the left is used to indicate if any change has occurred. Since this is initially unknown, the flag does not flutter.

The user can specify what algorithm, signature and other settings to use in order to check the pages in the list. Then the user starts the analysis, after which the system presents its relevance assessment of the changes in the pages. Figure 8 shows the results of using the proportional algorithm and comparing the signature with the last valid signature for our example path.

The system encodes the degree of relevance by the color of the page title or URL. In particular, black means that there is no relevant change. Green, yellow, and red text indicate low, medium, and high degrees of changes respectively. What is low, medium and high is determined by the user. In the future, these levels could be used for other purposes such as hiding a page that has changed more than a certain amount.

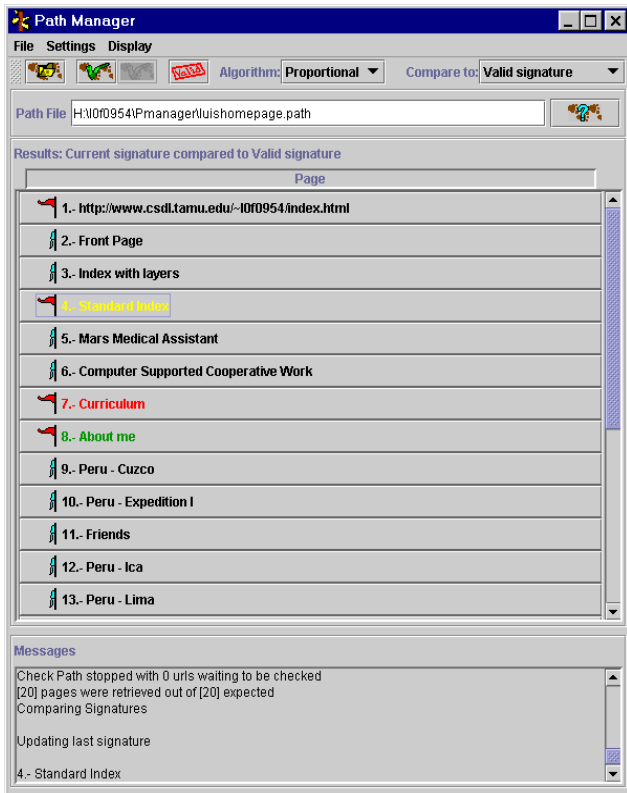


Figure 8: Path Manager relevance assessment

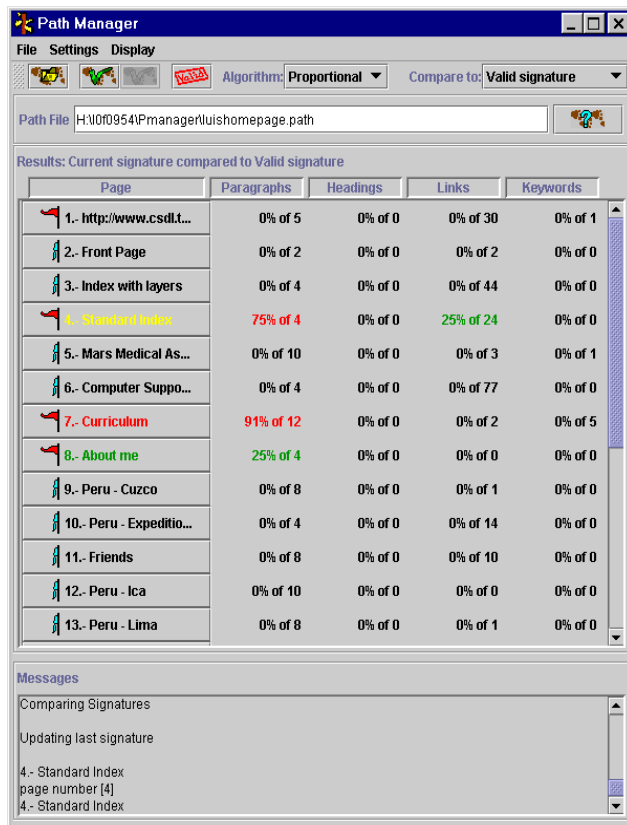


Figure 9: Detailed relevance assessment

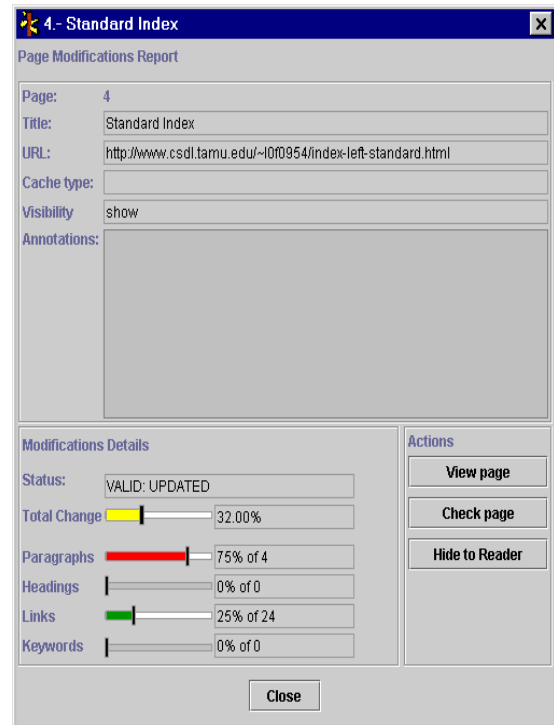


Figure 10: Individual page information

In cases where nothing has changed (the global checksum is the same) a blue hanging flag is shown next to the page title, but if there was a change a red flag is shown fluttering. These flags work as a boolean detection of even the slightest changes.

The user might wonder about what kinds of changes prompt the system to assess the overall relevance of the changes. In order to support the more inquisitive user, the Path Manager can display the amount of change to the particular document signatures used in the relevance assessment. Figure 9 shows the view of the different metrics used in the relevance assessment.

In addition, the user can obtain specific information about a particular Web page by clicking on the page title. This displays the dialog shown in Figure 10 (for page 4 in our example path.)

The top portion of the dialog shows information about the use of the page within the path, such as the cache strategy, annotations and visibility of the page. The bottom of the dialog presents the assessment of change to the page. This dialog also allows the user to check this page individually, or to display it in a Web browser.

At this point the user, who may not be the same person who authored the path, can judge the page inappropriate and choose to hide the page from the viewers. While this action prevents the material from being shown to readers of the path, it does not delete the page from the path, leaving that decision to the path author. In case there have been connection or retrieval problems, the manager can prompt the system to check this page again.

## Web Page Retrieval and Connectivity

When accessing the Web, connection times are never constant, as they depend on many variables out of the system or users' control. The time to retrieve two Web pages can vary dramatically. Even for a single page, the connection time often varies. When connecting to a Web page seems to take too long, sometimes the best strategy is to cancel and immediately restart the retrieval process, which may result in the Web page being loaded faster.

The Path Manager also takes advantage of the connection times by processing several pages in parallel. In this scheme, each page is retrieved and analyzed in an independent thread. Time is optimized by analyzing the already retrieved pages while waiting for the pages that take longer to be loaded. The user controls the maximum number of simultaneous threads.

Even using independent threads, is possible to encounter situations such as no response, slow connections, and very long pages. In order to deal with these situations the user can set individual timeouts for the system to complete the connection, retrieval, and analysis of the pages.

## CONCLUSIONS

In order to inform the design of tools that aid in the automatic assessment of changes to Web pages, a study was conducted, first to determine how people perceive changes to Web pages, and second to better understand what changes people consider relevant.

The study covered changes to the content, presentation, and structure of Web pages. The results show that the perception of content change is highly correlated with perception of overall change and desire to be notified of the change. Similarly, when faced with structural changes, people show a strong desire for notification. In contrast, changes of presentation are generally regarded as not important, with the exception of drastic changes.

The study also examined how time and the amount of change affect the perception of change. The amount of time given to assess the changes had little effect on the participants' perception of content changes, although time was critical for locating structural changes. The amount of change affected the perception of other types of change. As the quantity of change increased, subjects perceived the change to be of all types.

The results of the study influenced our design of algorithms to automatically assess the relevance of changes to Web pages within the context of the Walden's Paths' Path Manager. The Path Manager aids the user by providing a configurable evaluation and visualization of the changes to a path or other list of Web pages.

The next stage of this work is to test and evaluate the alternative metrics and algorithms within the context of Walden's Paths and personal bookmarks lists.

## ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Grant Nos. IIS-9812040 and DUE-0085798.

## REFERENCES

1. AltaVista (2000). Available at <http://www.altavista.com/>
2. Brewington, B. & Cybenko, G. (2000), How Dynamic is the Web, in Proc. of WWW9—9th International World Wide Web Conference, IW3C2, 264-296.
3. Bush, V. (1945) As We May Think. *The Atlantic Monthly*, 101-108.
4. Francisco-Revilla, L., Shipman, F., Furuta, R., Karadkar, U., and Arora, A. Managing change on the Web. Proceedings of the ACM & IEEE Joint Conference on Digital Libraries, 2001.
5. Furuta, R., Shipman, F., Francisco-Revilla, L., Karadkar, U., & Hu, S. (1999), Ephemeral Paths on the WWW: The Walden's Paths Lightweight Path Mechanism. *WebNet 1*, 409-414.
6. Furuta, R., Shipman, F., Marshall, C., Brenner, D., & Hsieh, H. (1997), Hypertext Paths and the World-Wide Web: Experiences with Walden's Paths, in Proc. of Hypertext'97, ACM Press, 167-176.
7. Giles, L., Bollacker, K., & Lawrence, L. (1998), CiteSeer: An Automatic Citation Indexing System, in Proc. of DL'98, ACM Press, 89-98.
8. Johnson, D. (1997), Enabling the Reuse of World Wide Web Documents in Tutorials. Ph.D. Dissertation, Dept. of computer Science and Engineering. University of Washington, Seattle, WA.
9. Johnson, D.B., & Tanimoto, S.L. (1999), Reusing Web Documents in Tutorials with the Current-Documents Assumption: Automatic Validation of Updates, in Proc. of EDMEDIA'99, AACE, 74-79.
10. Karadkar, U., Francisco-Revilla, L., Furuta, R., Hsieh, H., & Shipman, F. (2000), Evolution of the Walden's Paths Authoring Tools, in Proc. of WebNet 2000--World Conference on the WWW and Internet, AACE, 299-304.
11. Levy, D.M. (1994), Fixed or Fluid? Document Stability and new Media, in Proc. of the European Conference on Hypertext Technology '94, ACM Press, 24-41.
12. Pazzani, M., & Billsus, D. Learning and Revising Reader Profiles: The Identification of Interesting Web Sites. *Machine Learning 27* (1997), Kluwer Academic Publishers, 313-331.
13. Shipman, F., Furuta, R., Brenner, D., Chung, C., & Hsieh, H. (1998), Using Paths in the Classroom: Experiences and Adaptations, in Proc. Hypertext'98, ACM Press, 167-176.
14. Shipman, F., Marshall, C., Furuta, R., Brenner, D., Hsieh, H., & Kumar, V. (1996), Creating Educational Guided Paths over the World-Wide Web, in Proc. of ED-TELECOM'96, AACE, 326-331.
15. Witten, I.H., Moffat, A., and Bell, T.C., *Managing Gygabytes. Compressing and Indexing Documents and Images*, 2nd Edition, Morgan Kaufman, San Francisco, CA, 1999.