Design and Evaluation of Maintenance Tools for Distributed Digital Libraries

Frank M. Shipman (shipman@csdl.tamu.edu)
Richard Furuta (furuta@csdl.tamu.edu)
NSF Grant # DUE-0121527

Handling the fluidity typical of Web pages is a general problem that affects everybody. This project intends to find ways to assist collection administrators in managing the continuous changes in their collection of Web pages. In order to due so, it will distinguish important changes from unimportant ones. The project's approach is based on observations of human perception of change and categorizing the desire for notification of such changes.

Motivation

The Web provides access to a wide variety of information but much of this information is fluid; it changes, moves, and occasionally disappears. This is a general problem. 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 are examples of page collections that can become out-of-date as changes are made to their components.

Increasingly, digital libraries are being defined that serve to collect pointers to World-Wide Web based resources, rather than to hold the resources themselves. In addition to the digital library that is provided by the "owner" of the documents there are a growing number of specialized libraries with distributed ownership of the documents. In these cases digital librarians often are limited to working with the pointers or links to the documents.

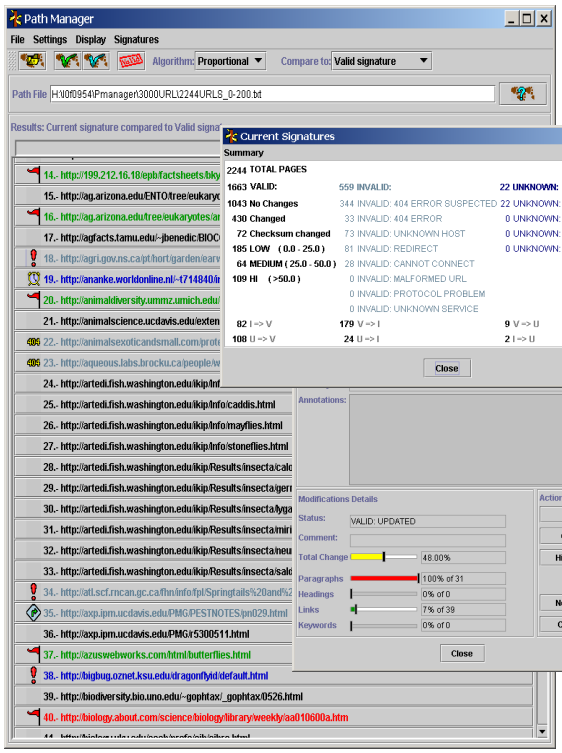
Maintaining these collections requires that they be updated continuously. This requires detecting and reacting to changes. However, while change detection is easily automated, assessing the relevance of the changes is not. People are typically recruited in order to evaluate the changes and provide guidelines for action. This is a difficult task highly dependent on the context and prone to ambiguities and disagreements. Two persons can and often do disagree about the importance of a change. Nevertheless, regardless of its difficult nature, this task needs to be performed continuously in order to manage the collection of pointers and Web pages. However, due to the fluid nature of the Web, this task is monumental for large collections. Hence it is necessary to create systems capable of detecting and reacting to changes in a proper way.

Approach

We are pursuing approaches to assist administrators in managing continuously occurring changes in their selected collection of Web pages. While our initial emphasis is on paths, our approach is equally applicable to Web-based domains that need to assess the relevance of changes to Web pages authored by third parties. We are extending our Path Manager to enable the maintenance of personal bookmark lists or other collections of resource links.

The challenge is to automatically determine the relevance of changes. The obvious approach would be to assess the magnitude or amount of change. Unfortunately, there is not an easily computed metric that provides a direct correlation between syntactic and semantic changes in a Web page. For example, there is no clear relationship between the number of bytes changed and the relevance of the change to the reader.

Determining the relevance of changes to a document can only be done in the context of how that document is to be used. Typically this knowledge is not available to the system that supports the



The Path Manager allows identifying the magnitude and relevance of changes to Web pages

assessment of relevance. Therefore our approach is to provide a variety of information to the person about changes of Web pages through a concise interface. To do so, we use classify the possible changes into four kinds of change: *content changes*, *presentation changes, structural changes*, and *behavioral changes*.

Content changes refer to modifications of the information contents of the page from the reader's point of view. Presentation changes are changes related to the document representation (such as fonts, colors, etc.) that do not reflect changes in its topic. Structural changes refer to the underlying connection of the document to other documents, or in other words the links of the page. Behavioral changes refer to modifications of the active components of a page such as scripts and applets.

The goal is to obtain an efficient measure of the semantic distance between two versions of a document. However, at the moment it is only possible to infer the semantic distance based on the syntactic characteristics of the documents. This is accomplished using heuristics and the Web page's markup. For example, a document can be partitioned based on heuristics such as "Paragraphs tend to encapsulate concepts".

The Study

As we mentioned before, people usually perform the evaluation of the relevance of changes. We have conducted a study to observe how humans perceive changes of Web pages in order to inform and validate the approach and design of systems that aid in the automatic assessment of change relevance.

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 System

Our approach is explored in the Path Manager. It supports the maintenance of collections of Web pages by recognizing, evaluating, and informing the user of changes. The evaluation of change is based on document signatures of the paragraphs, headings, links, and keywords. The system keeps track of original, last valid, and last collected signatures so users can determine both long-term and short-term change depending on their particular concern. The Path Manager has been designed to work in a distributed environment where connectivity to documents is unpredictable. Its architecture and instantiation provide users with control over system resources consumed. Also, the system provides feedback about documents that are not successfully evaluated.

For more information see:
http://www.csdl.tamu.edu/walden